ESCUELA DE DOCTORADO
INTERNACIONAL DE LA USC

Fco. Javier
Temes Rodríguez

Tesis doctoral

CARACTERIZACIÓN DE LOS PATRONES DE METILACIÓN DEL RETROTRANSPOSON LINE1 EN GENOMAS TUMORALES MEDIANTE TECNOLOGÍAS DE SECUENCIACIÓN DE MOLÉCULA ÚNICA

Santiago de Compostela, 2023

Programa de doctorado en Medicina Molecular

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR DE LA TESIS

Caracterización de los patrones de metilación del retrotransposón LINE1 en genomas tumorales mediante tecnologías de secuenciación de molécula única

Don Fco. Javier Temes Rodríguez

Grupo Genomas y enfermedad

Universidade de Santiago de Compostela

Presento mi Tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1. La Tesis abarca los resultados de la elaboración de mi trabajo.
2. En su caso, en la Tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3. La Tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.
4. Confirmo que la Tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.

Fdo. Fco. Javier Temes Rodríguez

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DEL DIRECTOR Y TUTOR DE LA TESIS

Dr. Daniel García Souto en calidad de director

Dr. Xulio Maside Rodríguez en calidad de tutor

de la Tesis titulada:

Caracterización de los patrones de metilación del retrotransposón LINE1 en genomas tumorales mediante tecnologías de secuenciación de molécula única.

INFORMAN:

Que la presente Tesis, corresponde con el trabajo realizado por Don **Fco. Javier Temes Rodríguez** bajo nuestra dirección, y autorizamos su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como directores de ésta no incurre en las causas de abstención establecidas en Ley 40/2015.

En Santiago de Compostela 22 de agosto de 2023

Fdo.: Dr. Daniel García Souto

Fdo.: Dr. Xulio Maside Rodríguez

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Conflicto de interés

El autor y los directores de este trabajo acordaron presentar los resultados de esta Tesis y declaran no tener ningún conflicto de interés.

El autor declara que todas las figuras usadas en el manuscrito han sido creadas por él mismo, excepto aquellas en cuyo pie se detalla la autoría original.

Aprobación del comité de ética:

Todas las muestras conseguidas de biobancos nacionales para este proyecto cuentan con permiso del CEIC. (The impact of L1-mediated somatic rearrangements in the origin and development of human cancer" promovido por José Manuel Castro Tubío, con código de registro 2018/586, y evaluado por el Comité el 19/12/2018 (Anexo I). El resto de los datos genómicos que se utilizarán ya han sido publicados o están en vías de publicación, por lo que no se identifica ningún tipo de implicación ética o legal.

Esta tesis ha sido financiada con cargo a los siguientes proyectos

- LABAE20053TUBI, Impacto funcional de las interacciones de la cromatina mediadas por retrotransposición en el genoma 3D del cáncer. Spanish Association for Cancer Research (AECC). Tubio 1. (University of Santiago de Compostela). 01/12/2020- 30/11/2023. 300,000 €.
- PGC2018-102245-B-100, El impacto de los reordenamientos somáticos mediados por L1 en el origen y desarrollo del cáncer humano. Ministerio de Ciencia, Innovación y Universidades. Tubio 1. (University of Santiago de Compostela). 01/01/2019- 31/12/2021. 350,900 €.
- 716290 - SCUBA CANCERS, SCUBA CANCERS: Finding the genetic causes of contagious metastases under the sea. European Research Council (ERC). Tubio 1. (University of Santiago de Compostela & University of Vigo). 01/01/2017-31/12/2021. 1,480,000 €.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

0. AGRADECIMIENTOS

Esta Tesis no habría sido posible sin el constante apoyo, supervisión, consejos y buen hacer de Daniel García Souto, director de la Tesis, buen amigo y mejor compañero al cual le estoy infinitamente agradecido. También quisiera agradecer a Jose C. Tubio iniciador de la línea de investigación sobre la que versa este trabajo, por acogerme en su grupo de investigación y permitirme el uso de recursos y datos utilizados en la Tesis.

Grandes compañeros me han ayudado y animado en este viaje como Sonia Zumalave, Jorge Rodríguez, Martín Santamarina y Vanesa Hurtado los cuales gentilmente me han permitido usar parte de su trabajo como punto de partida para alguna sección de este. No quisiera olvidarme del resto de los compañeros del grupo de Genomas y Enfermedad con quien tuve la suerte de compartir estos años de la Tesis como Berni, Mónica, Paula, Nuria, Jorge, Juan, Iago, las Anas e Iria.

Para el desarrollo de este trabajo hicieron falta gran cantidad de recursos informáticos proporcionados por el CESGA, así como el conocimiento y apoyo del personal del mismo, especialmente Pablo Rey y Carlos Fernández, a ellos mi agradecimiento.

Mi especial gratitud a Xulio Maside por sus gestiones y a los miembros del tribunal por el esfuerzo de evaluar la presente Tesis.

Sobre todo, quisiera a gradecer a mi familia, mi madre Teresa, mi padre Javier, a mis hermanos Luis y Teresita, y a Chus el sol de mi vida, la paciencia que han tenido, y el cariño que me han proporcionado durante la confección de esta Tesis y siempre. A ellos va dedicado este trabajo de investigación.

Innumerables amigos, entre ellos los BoSCQuejeros me han animado durante estos años, también a vosotros muchas gracias.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1. RESUMEN

1.1 GALEGO

Título: Caracterización dos patróns de metilación do retrotransposón LINE1 en xenomas tumorais mediante tecnoloxías de secuenciación de molécula única

Resumo: O xenoma humano é moi repetitivo, en gran parte debido aos elementos móbiles que contén, dos que os retrotransposóns forman unha parte importante. Estes retrotransposóns están implicados en varios procesos mutacionais. O estado de metilación dos seus promotores é moi importante, xa que determina en gran medida a súa actividade mutacional dentro dun tumor. O único retrotransposón autónomo e activo no ser humano e o LINE1, contendo o xenoma humano mais de 500,000 copias. Debido á súa natureza repetitiva, así como á súa lonxitude, estudar estes patróns é un desafío. A aparición de tecnoloxías de secuenciación de lectura longa 'single molecule sequencing', como as de Oxford Nanopore (ONT), permite superar estas dificultades. Por un lado, as lecturas longas permiten cubrir de xeito inequívoco toda a lonxitude do mesmo elemento, evitando erros de cartografía como consecuencia da súa natureza repetitiva. Por outra banda, a tecnoloxía baseada en nanoporos permite secuenciar o ADN de forma nativa, sen necesidade de tratar as mostras con bisulfito e diferenciar as citosinas metiladas das non metiladas. Neste proxecto de Tese, facemos uso desta tecnoloxía para caracterizar os patróns de metilación dos retrotransposóns en tecidos tumorais en non tumorais, para responder a importantes preguntas biolóxicas sobre a súa actividade nos dous tipos de tecidos.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1.2 ESPAÑOL

Título: Caracterización de los patrones de metilación del retrotransposón LINE1 en genomas tumorales mediante tecnologías de secuenciación de molécula única.

Resumen: El genoma humano es altamente repetitivo, en gran parte debido a los elementos móviles que contiene, de los cuales los retrotransposones forman una parte importante. Estos retrotransposones están implicados en varios procesos mutacionales. El estado de metilación de sus promotores es muy importante, ya que determina en gran parte su actividad mutacional dentro de un tumor. El único retrotransposón autónomo y activo en los seres humanos es el LINE1, conteniendo el genoma humano más de 500,000 copias. Debido a la naturaleza repetitiva de éstos, y también a su longitud, el estudio de estos patrones representa todo un desafío. La aparición de tecnologías de secuenciación de lecturas largas 'single molecule sequencing', como la de Oxford Nanopore (ONT), permite solventar estas dificultades. Por un lado, las lecturas largas permiten secuenciar de manera inequívoca toda la longitud de un mismo elemento, evitando errores de mapeo como consecuencia de su naturaleza repetitiva. Por otro lado, la tecnología basada en nanoporos permite secuenciar el ADN de forma nativa, sin necesidad de tratar las muestras con bisulfito, y diferenciar las citosinas metiladas de las no metiladas. En este proyecto de Tesis, hacemos uso de esta tecnología para caracterizar los patrones de metilación de los retrotransposones en tumores y tejidos no tumorales, para responder a importantes cuestiones biológicas sobre la actividad de los mismos en ambos tipos de tejidos.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

1.3 ENGLISH

Title : Characterization of the LINE1 retrotransposon methylation patterns in tumour genomes using single molecule sequencing technologies

Summary: The human genome is highly repetitive, largely due to the mobile elements it contains, of which retrotransposons form an important part. These retrotransposons are involved in various mutational processes and the methylation status of their promoters is very important, as it largely determines their mutational activity within a tumour. LINE1 is the only autonomous and active retrotransposon in humans. The human genome contains more than 500,000 copies. Due to their repetitive nature, as well as their length, studying these patterns is challenging. The emergence of long reads 'single molecule sequencing' technologies, like the Oxford Nanopore (ONT), makes it possible to overcome these difficulties. On the one hand, long reads allow to cover unequivocally the entire length of the same element, avoiding mapping errors as a consequence of its repetitive nature. On the other hand, nanopore-based technology allows DNA to be sequenced natively, without the need to treat the samples with bisulfite, and to differentiate methylated from non-methylated cytosines. In this thesis project, we make use of this technology to characterize retrotransposon methylation patterns in tumoral and non tumoral tissues, to answer important biological questions about their activity in both types of tissues.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2. TABLAS DE CONTENIDOS Y LISTADOS

2.1 TABLA DE CONTENIDOS (RESUMIDA)

0. AGRADECIMIENTOS 7
1. RESUMEN 9
2. TABLAS DE CONTENIDOS Y LISTADOS 15
3. INTRODUCCIÓN GENERAL 29
4. OBJETIVOS 73
5. METODOLOGÍA 77
6. RESULTADOS y DISCUSIÓN I. Resultados metodológicos 107
7. RESULTADOS Y DISCUSIÓN II. Muestras y datos 177
8. RESULTADOS Y DISCUSIÓN III. Resultados biológicos 205
9. DISCUSIÓN GENERAL 259
10. CONCLUSIONES 277
11. GLOSARIO Y ABREVIATURAS 281
12. BIBLIOGRAFÍA 287
ANEXO I. Aprobación del comité de ética 299
ANEXO II. Permisos de imágenes y gráficas 303

2.2 TABLA DE CONTENIDOS (COMPLETA)

0. AGRADECIMIENTOS 7
1. RESUMEN 9
  1.1 Galego 9
  1.2 Español 11
  1.3 English 13
2. TABLAS DE CONTENIDOS Y LISTADOS 15
  2.1 Tabla de contenidos (resumida) 15
  2.2 Tabla de contenidos (completa) 15
  2.3 Listado de figuras 22
  2.4 Listado de tablas 27
  2.5 Listado de fragmentos de código 28

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. INTRODUCCIÓN GENERAL 29
3.1 Organización del presente manuscrito 31
3.1.1 Notas generales 31
3.1.2 Repositorio de datos, programas y resultados 32
3.2 Antecedentes 33
3.2.1 Retrotransposición y retrotransposones 33
3.2.2 Elementos LINE1 y L1HS 35
3.2.3 Retrotransposición de LINE1 36
3.2.3.1 Modelos de inserción de LINE1 38
3.2.3.2 Transducciones 39
3.2.3.3 Tipos de retrotransposición 40
3.3 Mutaciones, enfermedad y cáncer 42
3.4 LINE1, variación normal, enfermedad y cáncer 44
3.4.1 Regulación de la retrotransposición somática 47
3.4.2 Metilación del ADN 49
3.4.3 Islas CpG 50
3.4.4 Metilación de LINE1 51
3.5 Técnicas y herramientas para el estudio de los retrotransposones y sus patrones de metilación en cáncer 55
3.5.1 Tecnología de secuenciación de ácidos nucleicos 55
3.5.1.1 Primera generación: 55
3.5.1.2 Segunda generación 57
3.5.1.3 Tercera generación 58
3.5.1.3.1 Secuenciación en tiempo real de una sola molécula (SMRT) 59
3.5.1.3.2 Secuenciación de nanoporos 60
3.5.1.4 Aplicación de la secuenciación al estudio de los retrotransposones y cáncer 64
3.5.2 Técnicas y herramientas para el estudio de la metilación del ADN 66
3.5.2.1 Tratamiento con bisulfito y secuenciación 66
3.5.2.2 Detección directa de bases modificadas con secuenciadores de tercera generación 67
3.5.2.3 Flujo de trabajo para el análisis de la metilación 68
3.5.2.3.1 Flujo de trabajo para pretratamiento con bisulfito + secuenciación Illumina .. 68
3.5.2.3.2 Flujo de trabajo para datos de secuenciadores de 3ª generación 69
3.5.3 Herramientas para la búsqueda y análisis de retrotransposones 70
3.5.4 Técnicas y herramientas para el estudio de la metilación de los LINE1 71
3.6 Motivación de la Tesis 72
4. OBJETIVOS 73

4.1 Objetivos biológicos 75
4.1.1 Caracterizar los patrones de metilación de los LINE1 germinales (íntegros) y las regiones adyacentes en los genomas tumorales y su correspondiente tejido no tumoral. 75
4.1.2 Determinar la correlación entre el estado de metilación de los promotores de los LINE1 germinales (íntegros) y su actividad. 75
4.1.3 Caracterizar los patrones de metilación de las inserciones somáticas de los LINE1 y el tejido adyacente en el genoma tumoral y su correspondiente tejido no tumoral. 75
4.1.4 Determinar si las copias hijas de los LINE1 activos en el ser humano, heredan el patrón de metilación de los elementos fuente de los cuales provienen cuando se movilizan en los tumores. 75
4.2 Objetivos metodológicos 75
4.2.1 Adaptar e implementar en el laboratorio la tecnología de secuenciación de lecturas largas “single-molecule sequencing” MinION de Oxford Nanopore Technologies. 75
4.2.2 Desarrollar novedosos algoritmos informáticos para el estudio de patrones de metilación de elementos móviles. 75
5. METODOLOGÍA 77
5.1 Protocolos de laboratorio 79
5.1.1 Electroforesis en gel de agarosa 79
5.1.2 Reacción en cadena de polimerasa (PCR) 79
5.1.3 Extracción de ADN a partir de tejido 79
5.1.4 Extracción de ADN a partir de sangre y/o líneas celulares 80
5.1.5 Purificación del ADN 80
5.1.6 Limpieza del ADN 80
5.1.7 Cuantificación y control de calidad del ADN 81
5.1.7.1 Medición de la cantidad de ADN y contaminación por proteínas: 81
5.1.7.2 Medición de la cantidad de ADN de doble cadena 82
5.1.7.3 Medición de la calidad y longitud del ADN. 82
5.1.8 Conservación del ADN 83
5.1.9 Fragmentación del ADN 83
5.1.10 Selección de ADN de cadena larga (Circulomics SRE) 84
5.1.11 Secuenciación NGS 85
5.1.12 Secuenciación de nanoporos mediante la tecnología de Oxford Nanopore Technologies. 85
5.1.12.1 Preparación de la librería de ADN para secuenciación de nanoporos ONT (protocolos SQK-LSK108 y 109) 86

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.1.12.1.1 Preparación de una librería de fragmentos amplificados (amplicones) para secuenciación por nanoporos 90
5.1.12.1.2 Preparación de una librería con marcadores secuenciales (*barcoding)* para secuenciación por nanoporos 90
5.1.12.2 Preparación de la librería para secuenciación de nanoporos (protocolo Rapid Sequencing gDNA SQK-RAD004). 91
5.1.12.3 Ejecución del programa MinKnow para la secuenciación de la librería 92
5.1.12.4 Limpieza de las celdas de secuenciación 92
5.2 Pipelines y programas bioinformáticos 92
5.2.1 Alineación de datos de secuenciación de Illumina (lecturas cortas "pair end") 93
5.2.2 Basecalling de los datos sin procesar (raw) de lecturas largas ONT 94
5.2.3 Alineación de datos de secuenciación de lecturas largas MinION (ONT) 94
5.2.4 Alineación de datos de metilación. Pretratamiento con bisulfito + Illumina + Bismark 94
5.2.5 Alineación de datos de metilación. Lecturas largas obtenidas con ONT+Megalodon 97
5.2.6 Búsqueda de retrotransposones con Meiga 99
5.2.7 Búsqueda de retrotransposones con xTea, rMelt, PALMER 101
5.2.8 Inferencia de los haplotipos (phasing) con whatshap 101
5.2.9 Comprobación de la integridad de los ficheros fast5 102
5.2.10 Visualización de los datos de secuenciación con IGV 102
6. RESULTADOS y DISCUSIÓN I. Resultados metodológicos 107
6.1 Adaptación e implementación en nuestro laboratorio de las tecnologías de secuenciación de molécula única. 109
6.1.1 Trabajo de laboratorio 109
6.1.1.1 Secuenciación del fago λ 110
6.1.1.2 Secuenciación de un genoma humano. 115
6.1.1.3 Secuenciación de amplicones 118
6.1.1.4 Secuenciación con marcadores secuenciales (*barcoding*) 120
6.1.1.5 Pruebas de fragmentación del ADN 120
6.1.1.6 Pruebas de fragmentación Covaris G-Tube. 122
6.1.1.7 Selección de ADN de cadenas largas 123
6.1.1.7.1 Pruebas con membranas (o tripas) de diálisis. 123
6.1.1.7.2 Recorte de bandas 125
6.1.1.7.3 Gel de agarosa, migración en cubeta de electroforesis standard y extracción de ADN del gel 125
6.1.1.7.4 Gel de agarosa de grandes dimensiones y cubeta de electroforesis pequeña 126

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.1.1.7.5 Gel de agarosa de grandes dimensiones y cubeta de electroforesis de construcción propia 127
6.1.1.7.6 Conclusiones del recorte de bandas y la extracción del ADN de los geles de agarosa 128
6.1.1.7.7 Gel de agarosa de doble densidad y cubetas estándar con inversión de la corriente (GD-Piping) 128
6.1.1.7.8 BluePiping 130
6.1.1.7.9 Gel de agarosa con pocillos intermedios y cubeta de electroforesis standard 130
6.1.1.7.10 Invitrogen E-gel y cartuchos de selección de tamaño 131
6.1.1.7.11 Limpieza ADN con magnetic beads y buffer de limpieza de ONT 134
6.1.1.8 Problemas a la hora de cargar la librería 137
6.1.1.9 Recarga de ATP (refuelling) 137
6.1.1.10 Rellenado (refiling) 138
6.1.2 Diseño y construcción de un mezclador de baja velocidad 138
6.1.3 Diseño y construcción de cubeta de electroforesis de reducidas dimensiones 140
6.1.4 Diseño y construcción de una cámara refrigeradora 141
6.1.4.1 Características 142
6.1.4.2 Construcción de la cámara 142
6.1.4.3 Circuito térmico 143
6.1.4.4 Circuito eléctrico 144
6.1.4.5 Pruebas de la cámara refrigeradora 146
6.2 Desarrollo de programas informáticos. 148
6.2.1 Programas auxiliares 148
6.2.1.1 MGD_agente: el problema de los millones de ficheros 148
6.2.1.2 SVI. Integrador de variantes estructurales (Structural variants integrator) 149
6.2.2 Methyl-Meiga 149
6.2.3 Desarrollo de *scripts* de análisis estadístico y creación de gráficas (*notebooks* de Python). 156
6.3 Validación de la tecnología de secuenciación de datos de metilación 157
6.3.1 Datos de metilación de las líneas celulares obtenidos mediante pretratamiento de bisulfito + Illumina 158
6.3.2 Datos de metilación obtenidos con ONT + Megalodon 159
6.3.3 Superioridad de la tecnología de ONT para analizar y estudiar la metilación de los LINE1 159
6.3.4 Comparativa de los datos de metilación de Megalodon con los datos bisulfito + Illumina (“Gold standard”) 161
6.4 Comparativa de métodos gráficos y estadísticos de metilación general 164
6.4.1 Methylartist 164

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.4.2 Comparativa de *scripts* gráficos ad hoc creados por mí. 171
6.4.2.1 Análisis usando datos de probabilidad por cada CpG 172
6.4.2.2 Análisis usando datos binarizados por cada CpG 173
6.4.2.3 Descripción del método usado para generar gráficas y estadísticas 176
7. RESULTADOS Y DISCUSIÓN II. Muestras y datos 177
7.1 Muestras 179
7.1.1 Tumores primarios y su tejido adyacente 179
7.1.1.1 Pureza de los tumores 185
7.1.2 Líneas celulares 186
7.1.2.1 NCI-H2087 187
7.1.2.2 NCI-H2009 188
7.1.2.3 NCI-BL2087 y NCI-BL2009. 188
7.2 Conjunto de datos 189
7.2.1 Genoma de referencia 189
7.2.2 Secuencia consenso L1HS 189
7.2.3 LINE1 objeto de análisis 190
7.2.3.1 Análisis inicial con RepeatMasker obtenido de UCSC 191
7.2.3.2 Análisis inicial con datos obtenidos de la web de L1base y RepeatMasker 192
7.2.3.3 Base de datos definitiva de LINE1 germinales, íntegros 195
7.2.4 Patrones de dinucleótidos CpG e Islas CpG en los LINE1 195
7.2.4.1 Ta-1 196
7.2.4.2 Ta-0 196
7.2.4.3 Pre-Ta-0 197
7.2.4.4 L1PA1 (L1HS) 197
7.2.5 Datos de las muestras tumorales, tejido adyacente y/o sangre 198
7.2.6 Inserciones somáticas de LINE1 200
7.2.6.1 Inserciones somáticas de LINE1 obtenidos filtrando con datos de lecturas largas (ONT) 200
7.2.6.2 Inserciones somáticas de LINE1 obtenidos filtrando con datos de lecturas cortas (Illumina + xTea) 202
8. RESULTADOS Y DISCUSIÓN III. Resultados biológicos 205
8.1 Análisis general de las muestras 207
8.1.1 Metilación 207
8.1.1.1 Metilación de las líneas celulares 210
8.1.2 Hidroximetilación 212
8.2 Metilación de LINE1 germinales y regiones adyacentes en genomas tumorales y normales. 214

8.2.1 Búsqueda de regiones diferencialmente metiladas. 214
8.2.2 Estimación de la pureza basada en la metilación....215
8.2.3 Asignación de metil-alelos (epialelos)....218
8.2.3.1 Asignación de metil-alelos (epialelos) en líneas celulares. 222
8.2.4 Metilación de los LINE1 germinales íntegros por regiones. 224
8.2.5 Análisis individual de los LINE1 germinales íntegros 226
8.2.6 Metilación diferencial entre el promotor y el cuerpo de los LINE1 germinales íntegros.....231
8.3 Metilación y actividad de los LINE1 germinales 234
8.3.1 Análisis general....234
8.3.2 Discusión 239
8.4 Metilación de las inserciones somáticas de LINE1 en el genoma tumoral 240
8.4.1 Patrones de metilación de inserciones somáticas de LINE1 por lectura 243
8.4.2 Discusión 246
8.5 Las copias hijas de los LINE1 activos en el ser humano, no heredan el patrón de metilación de los elementos fuente de los cuales provienen 247
8.6 Análisis adicionales 253
8.6.1 Patrones de hidroximetilación de las inserciones somáticas 253
8.6.2 Epivariación proximal, falso hallazgo 255
9. DISCUSIÓN GENERAL....259
9.1 Puesta a punto de una unidad de secuenciación de lecturas largas 261
9.2 Methyl-Meiga es un software específico para la detección de la metilación de LINE1 somáticos y germinales.....266
9.3 Resultados biológicos....268
9.4 Futuros desarrollos, nuevas preguntas, posteriores vías de investigación.....274
10. CONCLUSIONES 277
11. GLOSARIO Y ABREVIATURAS....281
11.1 Glosario 281
11.2 Abreviaturas: 284
11.2.1 Castellano....284
11.2.2 Inglés....284
12. BIBLIOGRAFÍA 287
ANEXO I. Aprobación del comité de ética 299
ANEXO II. Permisos de reproducción de imágenes y gráficas....303

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2.3 LISTADO DE FIGURAS

Figura 1: Tipos de transposones y mecanismos de transposición. 33
Figura 2: Clases de transposones. 34
Figura 3: Transposones en el genoma humano. 35
Figura 4: Esquema de un retrotransposón LINE1 de humanos. 36
Figura 5: Proceso de retrotransposición de un LINE1. 37
Figura 6: Modelos de inserción de un LINE1. 39
Figura 7: Tipos de retrotransposición. 41
Figura 8: Evolución de las células cancerosas. 43
Figura 9: "Hallmarks of cancer". 43
Figura 10: Posibles mecanismos de alteración de la actividad génica. 46
Figura 11: Mecanismos de silenciamiento del LINE1 y su mantenimiento. 48
Figura 12: Fórmula estructural de la citosina, la metil-citosina y derivados. 50
Figura 13: Nombres de agrupaciones de CpG. 50
Figura 14: Actividad del LINE1. 52
Figura 15: Determinación de una secuencia de ADN/ARN por el método de didesoxi de Sanger. 56
Figura 16: Ejemplo de la salida de un secuenciador capilar. 56
Figura 17: Secuenciación NGS. 57
Figura 18: Términos relacionados con las lecturas NGS. 58
Figura 19: Preparación de la librería para la secuenciación en tiempo real de una sola molécula (SMRT). 59
Figura 20: Esquema de secuenciación SMRT de PacBio. 60
Figura 21: Contador Coulter. 61
Figura 22: Diversas imágenes del dispositivo MinION. 62
Figura 23: Relación N.º de moléculas ADN masa de la librería. 62
Figura 24: Secuenciación por nanoporos. 63
Figura 25: Dispositivos de secuenciación de Oxford Nanopore Technologies. 63
Figura 26: "The Cost of Sequencing a Human Genome". 64
Figura 27: Cánceres con alta tasa de inserciones somáticas de retrotransposones. 65
Figura 28: Escalera de pesos moleculares del ADN. 79
Figura 29: Espectrofotómetro One 81
Figura 30: Fluorímetro Invitrogen Qubit. 82
Figura 31: TapeStation 4200 (Agilent Technologies) y salida de la medición de la calidad del ADN de varias muestras. 83
Figura 32: Tubo de fragmentación Covaris. 83
Figura 33: Esquema de selección de cadenas de ADN largas usando el protocolo de Circulomics SRE. 84
Figura 34: Secuenciador NGS miseq de Illumina 85
Figura 35: Carga de una librería en un MinION de ONT 86
Figura 36: Esquema de una librería de secuenciación SQK-SLK109 ONT 87
Figura 37: Esquema de una librería de secuenciación rápida SQK-RAD004 ONT 91
Figura 38: Comparativa de MEIGA-LR con otros programas de descubrimiento de retrotransposición somática. 100
Figura 39: Matriz de confusión 100
Figura 40: Ficheros .fast5 estropeados, detectados por Megalodon. 102

Figura 41 UCSC Genome Browser.....103
Figura 42: Visualización de la metilación de un LINE1 con IGV.....105
Figura 43: Comprobación de calidad del ADN del fago λ mediante un espectrofotómetro (NanoDrop).....110
Figura 44: Salida del MinKnow de la primera secuenciación en nuestro laboratorio, un fago λ. 111
Figura 45: Salida del programa NanoPlot de la primera secuenciación en nuestro laboratorio, un fago λ.....111
Figura 46: Aumento de cobertura debido a la cadena de control. 112
Figura 47: Salida del IGV mostrando un alineamiento de lecturas largas (MinION de ONT).....113
Figura 48: Salida de NanoPlot mostrando el número de lecturas por longitud de la secuencia leídas (Fago λ).....114
Figura 49: Salida de NanoPlot mostrando la calidad de las lecturas frente a su longitud (Fago λ).....114
Figura 50: Salida de NanoPlot mostrando el porcentaje de identidad de las lecturas (Fago λ).....115
Figura 51: Autorización del "sujeto 1" para la secuenciación de su ADN. 115
Figura 52: Comprobación de calidad del ADN del “sujeto 1”. 115
Figura 53: Salida de NanoPlot mostrando el número de lecturas por longitud de la secuencia leídas ("sujeto 1").....117
Figura 54: Salida de NanoPlot mostrando la calidad de las lecturas frente a su longitud ("sujeto 1").....117
Figura 55: Salida de NanoPlot mostrando el porcentaje de identidad de las lecturas ("sujeto 1").....118
Figura 56: Fragmento de un gel de agarosa de amplicones de la muestra NCI-H2087. 118
Figura 57: Gráfica producida por el programa MinKnow mostrando el número de lecturas por tamaño.....119
Figura 58: Punta de jeringuilla 30G.....120
Figura 59: Fragmentación del ADN con el Covaris G-Tube.....123
Figura 60: Salida de la Tape Station de las pruebas de fragmentación con el Covaris G-Tube. 123
Figura 61: Fragmentos de ADN de pequeño tamaño que no interesa secuenciar. 124
Figura 62: Pruebas de eliminación de fragmentos de ADN de reducida longitud mediante membranas de diálisis.....124
Figura 63: Recorte de bandas I. 125
Figura 64: Recorte de bandas II.....126
Figura 65: Creación de un gel de agarosa de grandes dimensiones.....126
Figura 66: Gel de agarosa de grandes dimensiones en el transiluminador, mostrando la ausencia de ADN. 127
Figura 67: Gel de agarosa de grandes dimensiones en el transiluminador, con ADN migrado en la cubeta de lectroforesis de construcción propia. 127
Figura 68: Cubeta de electroforesis de construcción propia.....127
Figura 69: Esquema de funcionamiento del GD-Piping.....128
Figura 70: Gel de agarosa del sistema GD-Piping.....129
Figura 71: Gel de agarosa al 5% con ADN.....129

Figura 72: BluePiping. 130
Figura 73: Gel de agarosa con dos grupos pocillos. 130
Figura 74: Invitrogen e-gel. 131
Figura 75: Migración del ADN en el dispositivo invitrogen e-gel, junto con la escalera de ADN para referencia. 132
Figura 76: Segunda prueba de migración de ADN en el invitrogen e-gel. 133
Figura 77: Gel de agarosa con el ADN recuperado del invitrogen e-gel. 134
Figura 78: Gel de agarosa con el ADN tras limpiarse con distintos procedimientos con bolitas magnéticas.. 136
Figura 79: Inclinando la flow cell de un MinION para ayudar a la introducción de la librería. 137
Figura 80: Salida del programa MinKnow mostrando la velocidad de secuenciación de una librería tras el refuelling. 138
Figura 81: Ejemplo de un mezclador comercial. 138
Figura 82: Diseño de un mezclador de baja velocidad con Lego 139
Figura 83: Mezclador de baja velocidad de construcción propia. 139
Figura 84: Diseño de cubeta de electroforesis para recorte de gel. 140
Figura 85: Detalle de unión de las piezas de la cubeta de electroforesis. 140
Figura 86: Cartuchos de los cuales se extrajeron los electrodos (izquierda) y prueba de funcionamiento con uno de los electrodos deshaciéndose. 141
Figura 87: Cubeta de electroforesis de construcción propia (izquierda y centro) y detalle de la cubeta dentro del transiluminador con un gel de grandes dimensiones (derecha). . 141
Figura 88: Salida del MinKnow mostrando las desviaciones de temperatura durante la secuenciación de una librería. 142
Figura 89: Construcción de la cámara refrigeradora. 143
Figura 90: Esquema de los componentes electrónicos y su interconexión (realizada con el programa Fritzing). 144
Figura 91: Esquema eléctrico del circuito eléctrico de la cámara refrigeradora. 145
Figura 92: Esquema del PCB (printed circuit board) 145
Figura 93: Pruebas de funcionamiento de la cámara refrigeradora. 147
Figura 94: Gráficas de temperatura. (prueba de la cámara refrigeradora) 147
Figura 95: Esquema del sistema MGD-Agente. 148
Figura 96: Imagen del IGV previo a marzo 2023, donde las citosinas no metiladas no se distinguen. 150
Figura 97: Esquema del filtrado de alineamientos realizado por Methyl-Meiga 151
Figura 98: Visualización en el IGV del filtrado de alineamientos espurios. 152
Figura 99: Diagrama de bloques del funcionamiento de Methyl-Meiga. 152
Figura 100: Salida de Methyl-Meiga mostrando un LINE1 germinal. 154
Figura 101: Composición de figuras producidas por Methyl-Meiga con distintos niveles de zoom. 155
Figura 102: Diferencias de mapabilidad sobre un LINE1 entre lecturas cortas de Illumina y largas de ONT 160
Figura 103: Comparación de la metilación de las líneas celulares (NCI-H2009, NCI-H2087) con el “Gold estándar”. 161
Figura 104: Correlación de la metilación de las líneas celulares. 163
Figura 105: Ejemplo de compactación de segmentos genómicos. 166

Figura 106: Gráfica de violín producida por Methylartist de la Metilación en regiones funcionales. 168
Figura 107: Gráfica de violín producida por Methylartist de la metilación en diversos LINE1 I. 169
Figura 108: Gráfica de violín producida por Methylartist de la metilación en diversos LINE1 II. 170
Figura 109: Pruebas de diferentes formas de representar la metilación con gráficas de violín. 172
Figura 110: Media de las medias 174
Figura 111: Pruebas de suavizado de las gráficas de violín. Datos de metilación binarizados. 175
Figura 112: Pruebas de suavizado de las gráficas de violín. Datos de metilación binarizados, media en ventanas de 10K pb. 175
Figura 113: Estadísticas de la secuenciación por ONT de las muestras de la cohorte. ..183
Figura 114: Gráficas de bigote de la cobertura (panel superior) y longitud de las lecturas (panel inferior) agrupados por tipo de tumor. 184
Figura 115: Leyenda de las figuras 116 y 1117 (circos plot de las líneas celulares). 186
Figura 116: Circos plot con información de la línea celular NCI-H2087. 187
Figura 117: Circos plot con información de la línea celular NCI-H2009. 188
Figura 118: Descarga de los datos del RepeatMasker del UCSC. 191
Figura 119: Disposición de las CpG e isla CpG del elemento Ta-1. 196
Figura 120: Disposición de las CpG e isla CpG del elemento Ta-0. 196
Figura 121: Disposición de las CpG e isla CpG del elemento pre-Ta. 197
Figura 122: Disposición de las CpG e isla CpG del elemento L1HS. 197
Figura 123: Gráfica de violín de la metilación global de tumores y tejido adyacente de nuestra cohorte (media por ventana de 10 Kpb). 208
Figura 124: Diferencias de la media y mediana entre la metilación de tejido adyacente - metilación del tumor. 210
Figura 125: Comprobación de la metilación de las líneas celulares. 211
Figura 126: Mapa de metilación - hidroximetilación de la muestra PD0277c 212
Figura 127: Gráfica de la hidroximetilación global de tumores y tejido adyacente de nuestra cohorte. 213
Figura 128: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1 I. 216
Figura 129: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1 II. 217
Figura 130: Correlación entre la clonalidad (estimada por Battemberg) y la pureza estimada usando el estado de metilación de los LINE1. 217
Figura 131: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1, por muestra. 218
Figura 132: Ejemplo de grafica de la media de metilación por LINE1 germinal, separando las lecturas por k-means. 219
Figura 133: Metil-alelos (epialelos) del LINE1 chr2 155671310 para las muestras PD0270 y PD0277. 220
Figura 134: Metil-alelos (epialelos) del LINE1chr6 29952435 para las muestras PD0266 y PD0331. 221

Figura 135: Ejemplo de los epialelos de una de las líneas celulares, mostrando el patrón caótico de metilación. 223
Figura 136: Comparativa tumor - normal (gráfica de bigotes) de la metilación de LINE1 germinales y regiones adyacentes. 225
Figura 137: Comparativa tumor - normal (gráfica de violín) de la metilación de LINE1 germinales y regiones adyacentes. 225
Figura 138: Mapa de calor (heatmap) de la metilación de los promotores de LINE1 germinales íntegros. 227
Figura 139: Mapa de calor (heatmap) de la metilación de los cuerpos de LINE1 germinales íntegros. 228
Figura 140: Mapa de calor (heatmap) de la metilación de las regiones adyacentes a los LINE1 germinales íntegros. 229
Figura 141: Gráfica suavizada de la metilación de todos los LINE1 germinales en los tejidos adyacentes 230
Figura 142: Gráfica suavizada de la metilación de todos los LINE1 germinales en las muestras tumorales 230
Figura 143: Correlación entre regiones de los LINE1 germinales (promotor, cuerpo, regiones adyacentes) 232
Figura 144: Correlación actividad – metilación de LINE1 germinales I 236
Figura 145: Correlación actividad – metilación de LINE1 germinales II. 236
Figura 146: Correlación actividad – metilación de LINE1 germinales III. 237
Figura 147: Correlación actividad – metilación de LINE1 germinales IV 237
Figura 148: Correlación actividad – metilación de LINE1 germinales V 239
Figura 149: Grafica resultante de ejecutar Methyl-Meiga sobre inserciones somáticas de LINE1 241
Figura 150: Metilación de inserciones somáticas. 242
Figura 151 Gráfica suavizada de la metilación de las inserciones somáticas de LINE1 menores de 5500 pares de bases en la cohorte de tumores. 243
Figura 152: Gráfica suavizada de la metilación de las inserciones somáticas de LINE1 mayores de 5500 pares de bases en la cohorte de tumores. 243
Figura 153: Gráfica de densidad (KDE) de la metilación por lectura de las inserciones somáticas desglosadas por tamaño de la inserción. 244
Figura 154: Grafica de densidad (KDE) de la metilación por lectura de las inserciones somáticas. 245
Figura 155: Gráfica de la media de metilación por lectura de las regiones promotoras y el cuerpo de las inserciones mayores de 5500 pb 245
Figura 156: Diagrama de cómo se realizó el análisis de los patrones de metilación de las copias hijas íntegras de LINE1. 247
Figura 157: Visualización en el IGV de un ejemplo de una inserción somática integra usando la secuencia proporcionada por MEIGA-LR (muestra NCI-H2087). 248
Figura 158: Ejemplo de gráfica comparativa con patrones similares de una inserción somática y el elemento fuente del que deriva (espacio genómico) 249
Figura 159: Ejemplo de gráfica comparativa con patrones similares de una inserción somática y el elemento fuente del que deriva (espacio CpG) 249
Figura 160: Ejemplo de gráfica comparativa con patrones distintos de una inserción somática y el elemento fuente del que deriva (espacio CpG) 249

Figura 161: Gráfica suavizada, de la hidroximetilación de las inserciones somáticas de LINE1 menores de 5500 pares de bases en la cohorte de tumores. 254
Figura 162: Gráfica suavizada, de la hidroximetilación de las inserciones somáticas de LINE1 mayores de 5500 pares de bases en la cohorte de tumores. 254
Figura 163: Comparación de epialelos informativos y no informativos de LINE1. 255
Figura 164: Gráfica de volcán, del log2 fold change de la metilación del alelo lleno / alelo vacío de inserciones somáticas (izquierda) y de inserciones germinales (derecha) 256
Figura 165: Gráfica de volcán del log2 fold change de la metilación del alelo lleno / alelo vacío de inserciones somáticas en las dos líneas celulares (NCI-H2087 y NCI-H2009) 257

2.4 LISTADO DE TABLAS

Tabla 1: Contaminación de las líneas celulares (NanoDrop) 82
Tabla 2: Características del computador de altas prestaciones del CESGA. 93
Tabla 3: Ficheros de salida de Megalodon. 99
Tabla 4: Estadísticas de secuenciación fago λ 112
Tabla 5: Estadísticas de secuenciación del “sujeto 1”. 116
Tabla 6 Pruebas de fragmentación/degradación del ADN 121
Tabla 7: Pruebas de fragmentación con el Covaris G-Tube 122
Tabla 8: Pruebas de selección de ADN con Invitrogen e-gel 131
Tabla 9: ADN recuperado con el dispositivo Invitrogen e-gel 132
Tabla 10: Recuperación de ADN en el e-gel 132
Tabla 11: Segunda prueba de recuperación de ADN con el Invitrogen e-gel 133
Tabla 12: Muestras del experimento de limpieza de ADN con bolitas magnéticas 136
Tabla 13: Lista de materiales de la cámara refrigeradora 146
Tabla 14: Lista de los ficheros producidos por Methyl-Meiga 156
Tabla 15: Ficheros de WGBS de las líneas celulares. 158
Tabla 16: Comparación de la metilación de las líneas celulares. 161
Tabla 17: Muestras usadas en el presente estudio. 181
Tabla 18: Estadísticas de los resultados de la secuenciación por ONT de las muestras. . 182
Tabla 19: Pureza de las muestras tumorales (obtenida con Battemberg). 185
Tabla 20: Secuencia consenso L1HS. 190
Tabla 21: Esquema de los datos del RepeatMasker descargados de la UCSC. 192
Tabla 22: Campos de la base de datos L1base. 193
Tabla 23: Cabecera del fichero de transposones RepeatMasker. 193
Tabla 24: Muestra del fichero de LINE1 con coordenadas de los LINE1 precisas. 194
Tabla 25: Inserciones somáticas de LINE1 en nuestra cohorte de tumores. 200
Tabla 26: Inserciones somáticas por muestra (muestras con tejido adyacente). 201
Tabla 27: Inserciones somáticas por muestra mayores de 6000 pares de bases (muestras con tejido adyacente). 201
Tabla 28: Inserciones somáticas por muestra (muestras filtradas con xTea). 204
Tabla 29:Inserciones somáticas por muestra mayores de 6000 pares de bases (muestras filtradas con xTea) 204
Tabla 30: Metilación global de las muestras de nuestra cohorte. 209
Tabla 31: Comprobación de la metilación. 211
Tabla 32: Comparativa de los epialelos, entre el tejido sano y los tumores (en número de elementos y porcentaje). 222

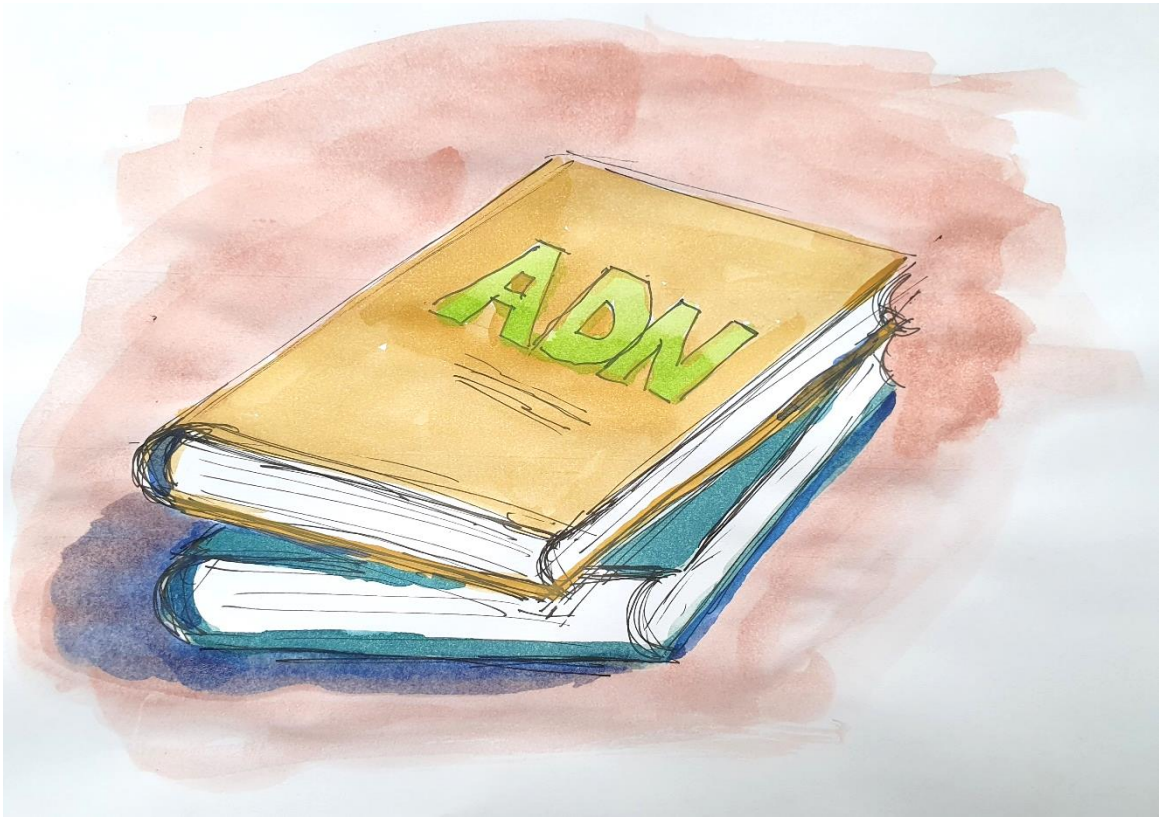
Tabla 33: Comparativa de los epialelos (líneas celulares), entre el tejido sano y los tumores (en número de elementos y porcentaje). 223
Tabla 34: Estadísticas de la metilación de LINE1 germinales (tumor y tejido adyacente), en regiones adyacentes, promotor y cuerpo de LINE1 224
Tabla 35: Datos de correlación entre regiones de los LINE1 germinales (promotor, cuerpo, regiones adyacentes) 231
Tabla 36: Regresión lineal entre los promotores y el cuerpo y regiones adyacentes de los LINE1 germinales. 232
Tabla 37: Abreviaturas de las regiones de la cromatina usadas 238
Tabla 38: Estadísticas de metilación de las inserciones somáticas 241
Tabla 39: Estadísticas de la metilación de las copias hijas y los elementos fuentes de los cuales derivan 252

2.5 LISTADO DE FRAGMENTOS DE CÓDIGO

Código 1: Preparación del genoma de referencia para Bismark. 95
Código 2: Script SLURM para la preparación de datos para Bismark. 96
Código 3: Ejemplo de lanzamiento del script de preparación de datos para Bismark. 97
Código 4: Lanzamiento de extracción de datos de metilación de Bismark. 97
Código 5: Lanzamiento de Megalodon. 98
Código 6: Lanzamiento de fast5_integrity_check. 102
Código 7: Ejemplo de código para la obtención de regiones genómicas (regions.sh) 165
Código 8: Script para el lanzamiento de Methylartist. 167
Código 9: Fusión de los datos de RepeatMasker con los de L1Base. 194
Código 10: Ejemplo de lanzamiento de MEIGA-LR en modo pareado. 199
Código 11: Eliminación de inserciones germinales usando datos de xTea. 203
Código 12: Ejecución de RADMeth. 215

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3. INTRODUCCIÓN GENERAL



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Conocimiento”. Acuarela y rotulador sobre papel (Ago. 2023)

3.1 ORGANIZACIÓN DEL PRESENTE MANUSCRITO

Este trabajo pretende arrojar luz sobre los patrones de metilación del retrotransposón LINE1 (L1) en cáncer, muy importantes en la regulación de su actividad. Dada la complejidad teórica y técnica de esta temática y con el cometido de facilitar la comprensión del lector, en el presente manuscrito se intentará dar una visión global de esta materia antes de entrar en cuestiones más específicas. Con tal fin, en la introducción general, expondré una panorámica de la investigación actual en el campo de la retrotransposición de LINE1, su implicación en los procesos cancerosos y en la evolución del tumor y el control de su actividad a partir de la metilación del ADN. De esta forma, describiré primero generalidades sobre los LINE1 y la metilación del ADN, para a continuación repasar la implicación de los LINE1 en el cáncer. Ya con más profundidad, describiré el estado del arte presente acerca de la metilación y los patrones de metilación de los LINE1 en cáncer. En los dos siguientes subapartados de la introducción se detallarán los recursos técnicos e informáticos que en la actualidad se emplean para realizar estudios genómicos y epigenómicos, así como su aplicación en el caso de los transposones. Se incluye un detallado recorrido por las tecnologías de secuenciación masiva existentes, punto de partida fundamental para entender los avances en las nuevas tecnologías de secuenciación de molécula única. Parte de esta Tesis está constituida por la puesta a punto en nuestro laboratorio de una de ellas, la tecnología de secuenciación por nanoporos. Tras describir brevemente la motivación, el enunciado de los objetivos de la presente Tesis doctoral, se explicará la metodología empleada para llevarla a cabo. Posteriormente se procederá a describir los resultados generados y a su discusión para, finalmente, establecer las conclusiones derivadas de mi trabajo.

3.1.1 Notas generales

Nota 1: Debido a que en este trabajo se hace uso extensivo de programas cuya salida presenta los números en formato anglosajón, y para evitar confusión, se va a usar ese criterio a lo largo de todo este manuscrito. En los números el separador decimal será un punto (.) y el separador de miles, millones será la coma (,) al contrario que en el sistema español.

Nota 2: Somos conscientes de que en algunas gráficas y tablas el texto tiene una fuente bastante pequeña, que no facilita su lectura. Es una decisión consciente y meditada consecuencia de tratar de mostrar la máxima información posible sin tener que trocearla en innumerables tablas y gráficas, haciendo complicado su manejo y comprensión global. El problema se minimiza en el manuscrito en formato PDF, disponible en un repositorio adjunto mencionado a continuación, donde también se encuentran las imágenes en alta resolución.

Nota 3: En el presente trabajo, aunque un tanto arbitrarias, se usan las siguientes "etiquetas" para referirnos al grado de correlación de Pearson en función de los valores absolutos de r: 0-0.19 se considera muy débil o inexistente, 0.2-0.39 débil, 0.40-0.59 moderada, 0.6-0.79 fuerte y 0.8-1 muy fuerte correlación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.1.2 Repositorio de datos, programas y resultados

Muchos trabajos de investigación en genómica del cáncer parten de una cohorte de datos de secuenciación masiva y requiere del desarrollo y optimización constante de programas informáticos específicos para llevar a cabo su análisis. Estos análisis generan una ingente cantidad de información en forma de código, imágenes y tablas de datos. La presente Tesis doctoral no es una excepción en este sentido y, consecuentemente, se hace imposible su inclusión en el presente manuscrito. Es por ello por lo que he optado por incluirlas en un repositorio público. Ahora bien, debido a que parte de esos resultados se deben a un trabajo de investigación del grupo Genomas y Enfermedad todavía pendiente de plasmarse en un artículo, el acceso es, a fecha de conclusión de este texto, todavía restringido. En cuanto el mencionado artículo sea publicado (a finales de 2023) el repositorio se hará de acceso público, siendo su dirección de acceso la siguiente:

<https://gitlab.com/fjt/javier-temes-repositorio-tesis.git>

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.2 ANTECEDENTES

3.2.1 Retrotransposición y retrotransposones

Una fracción relevante de los genomas eucariotas se compone de secuencias repetitivas, ya sea con distribución en tándem o dispersa a lo largo de su secuencia. Dentro de esta segunda categoría se encuentran los elementos transponibles o transposones. Descubiertos inicialmente por Bárbara McClintock en la planta del maíz (*Zea mais*) en 1944 (McClintock 1950) los transposones son secuencias de ADN que poseen o que en algún momento del desarrollo biológico o evolutivo de la especie hospedadora poseyeron capacidad de desplazarse por el genoma, o generar nuevas copias de sí mismos en posiciones distantes del genoma (Feschotte y Pritham 2007). Este proceso recibe el nombre de transposición. En un primer intento de clasificación, Finnegan (Finnegan 1989) discrimina dos clases de elementos transponibles en base a su mecanismo replicativo: Transposones de **Clase I**, o aquellos que requieren de un intermediario de ARN para llevar a cabo la transposición gracias a la acción coordinada de enzimas con actividad reverso transcriptasa e integrasa codificadas por el propio transposón, y transposones de **Clase II**, o aquellos que se copian directa-mente por acción de una trans-posasa sin intermediario de ARN. El primer mecanismo se denomina retrotransposición o también de "*copia y pega*" puesto que el elemento original no se moviliza. En contraposición, en los elementos de la clase II el mecanismo es de "*corta y pega*", donde la secuencia a transponer se escinde de la cadena de ADN y se inserta en otra posición del genoma (E. M. Ostertag y Kazazian 2001). Comúnmente los primeros elementos reciben el nombre de retrotransposones, en tanto que los de Clase II se suelen llamar transposones de ADN.

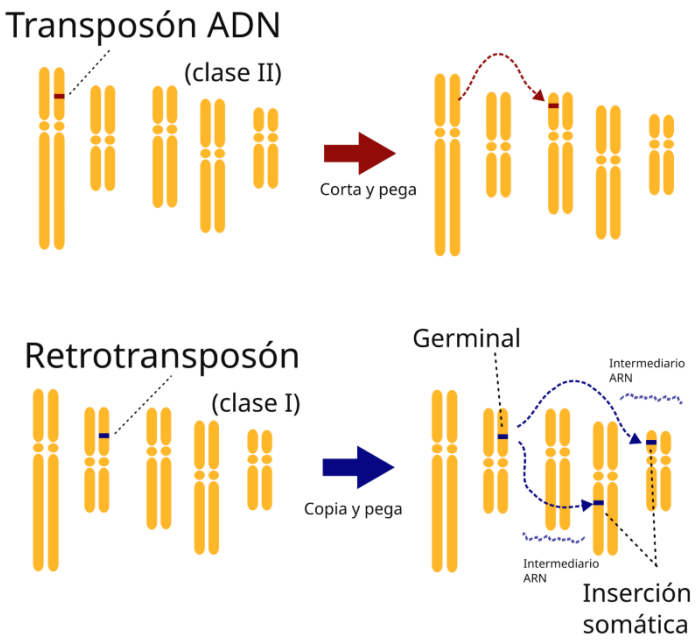


Figura 1: Tipos de transposones y mecanismos de transposición.

Esta clasificación, aunque correcta y efectiva, engloba transposones de muy diversa índole, tanto en composición génica, como en secuencia y origen en una misma clase. Debido a ello, en el año 2007 Wicker y colaboradores propusieron una clasificación jerárquica atendiendo a las características estructurales y el modo de replicación de los elementos móviles (Wicker et al. 2007). Si bien esta es la más ampliamente aceptada por la comunidad científica, algunas voces han reclamado su revisionismo, ya que no consideran el origen evolutivo de los transposones (Piégu et al. 2015; Arkhipova 2017). *Grosso modo*, Wicker *et al* definen dos grandes grupos de transposones para cada una de las clases inicialmente descritas por Finnegan (figura 2). Dentro de los retrotransposones (Clase I), que constituyen al menos el 42% del genoma humano (E. M. Ostertag y

Kazazian 2001), se encuentran aquellos que poseen repeticiones en tándem largas flanqueantes a sus puntos de inserción, los transposones LTR (Acrónimo en inglés de *Long Terminal Repeats*) y aquellos que no constan de las mismas, los retrotransposones no LTR. Los retrotransposones LTR, están constituidos por tres órdenes, LTR, DIRS y Penélope, son estructuralmente similares a retrovirus que carecen de los genes codificantes para las proteínas de la envuelta vírica y, de hecho, se cree que la mayoría son atavismos de provirus que fueron retro-transcritos e integrados en nuestro genoma en algún momento de nuestra evolución como especie. Al segundo grupo pertenecen los LINE (*Long Interspersed Nuclear Elements*) y los SINE (*Short Interspersed Nuclear Elements*). De los cinco órdenes posibles de retrotransposones (LTR, DIRS, Penélopes, LINEs y SINEs) en humanos sólo están activos los LINE y SINE, en concreto los pertenecientes a las familias de LINE1, Alu y SVA (H. H. Jr. Kazazian y Moran 2017; Beck et al. 2010; Deininger y Batzer 2002; Brouha et al. 2003; Belancio, Hedges, y Deininger 2008). En concreto, los LINE1 son uno de los retrotransposones más prevalentes en el genoma humano, debido a que son los únicos retrotransposones completamente autónomos y activos en el genoma humano actual. Otros elementos móviles, como por ejemplo los Alus y SVAs, englobados dentro de los SINE, son mucho más cortos y carecen del gen de la retrotranscriptasa, requieren de las enzimas codificadas por los LINE1 para propagarse y por lo que se consideran no autónomos.

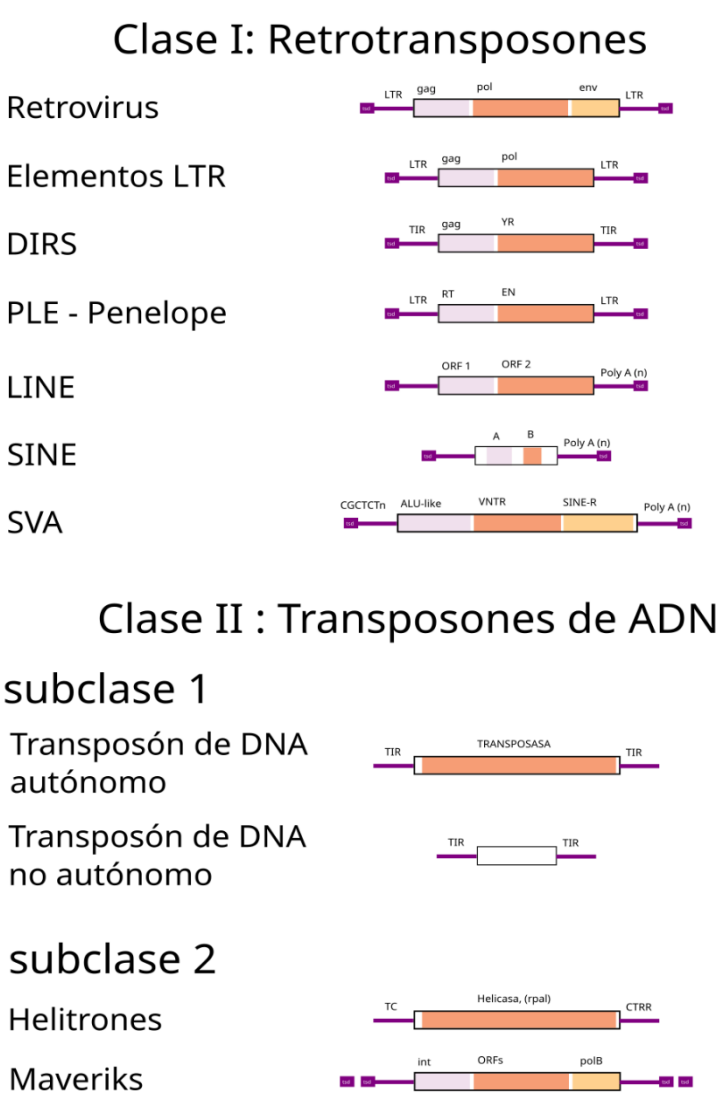


Figura 2: Clases de transposones. Adaptada de (Makałowski et al. 2019)

Las distintas clases de transposones varían en abundancia relativa en el genoma humano. Empleando herramientas informáticas específicas para el análisis de la fracción repetitiva a partir de datos genómicos tales como los almacenados en RepBase (Bao, Kojima, y Kohany 2015) (**http://www.repeatmasker.org**) se observa que alrededor del 52.5% del genoma humano son transposones, cuya mayoría pueden ser debidamente asignados a

alguna de las clases de transposones descritas (Figura 3). Aquellos que no han podido adscribirse a alguna de las clases son copias que han adquirido un número considerable de mutaciones a lo largo del tiempo, lo que hace que su identificación sea más complicada. Haciendo uso de herramientas más sofisticadas, Koning *y su grupo* (Koning et al. 2011) han encontrado que los elementos repetitivos pueden constituir entre el 66% y el 69% del genoma. De ellos los Alus, perteneciente a la familia de los SINEs constituyen el 39%, mientras que la familia de los LINE constituye el 21% del genoma. De estos últimos el LINE1 es el más abundante (entre 500,000 y 850,000 copias según los distintos autores) y el mejor estudiado, así como el objeto de estudio de esta Tesis doctoral.

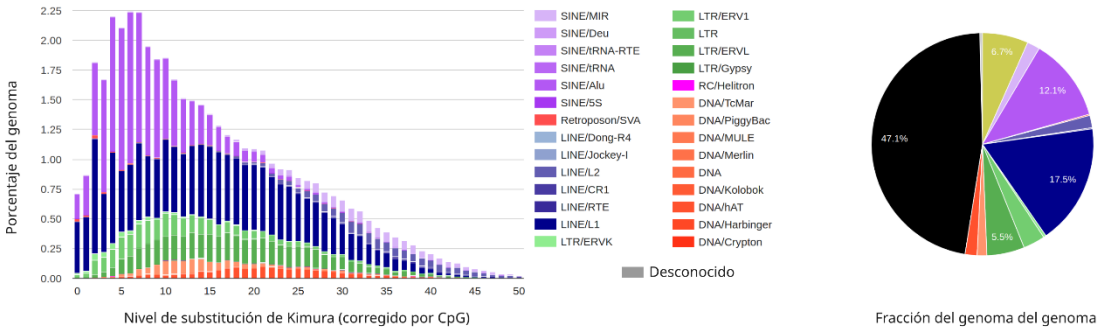


Figura 3: Transposones en el genoma humano.

3.2.2 Elementos LINE1 y L1HS

Como ya se ha mencionado, LINE es el acrónimo del inglés "*long interspersed nuclear element*", es decir, son elementos nucleares largos intercalados. Su origen aproximado se remonta a hace 160 millones de años, durante la radiación de los mamíferos, y desde entonces han evolucionado divergentemente en los distintos linajes de esta clase animal. Originariamente en primates coexistieron múltiples linajes de LINE, pero desde hace 40 millones de años solo ha prevalecido de forma ininterrumpida un único linaje de subfamilias (Goodier y Kazazian 2008). En humanos los únicos activos actualmente son los LINE1, en concreto de la familia L1HS (Smit et al. 1995; Beck et al. 2010), si bien coexisten en el genoma con copias de otras familias, como por ejemplo LINE-2 (Petri et al. 2019) y otros LINEs que estuvieron activos en los primates precursores del Homo sapiens moderno.

En mamíferos las copias completas de LINE1 rondan los seis mil pares de bases de longitud, al igual que la familia específica de humanos, los L1HS (Mir, Philippe, y Cristofari 2015). Estructuralmente los L1HS constan de dos marcos de lectura abiertos (ORF1 y ORF2) separados por un inter-espaciador de 63 pb en lo que convendremos en llamar cuerpo del transposón. El cuerpo del transposón se encuentra flanqueado por sendas regiones no transcribibles (UTR, "UnTranslated Region") y dos duplicaciones de sitio en tándem (TSD: "tandem site duplication"). La región UTR en 5', de 900 pares de bases (pb), contiene una zona de inicio de transcripción para la polimerasa II ("trans*cript*ion start site" - TSS), así como diversos motivos ricos en CG para unión de factores de transcripción y con actividad promotora. La región terminal 3' UTR que incluye una cola de poli adeninas (poliA) y otra zona de duplicación de sitio en tándem

(TSD). El ORF1 codifica para una proteína de 40kDa con funciones de unión al ARN (no específica de secuencia) con motivos de dedos de zinc, cremallera de leucina y hélices superenrolladas. Esta proteína actúa a modo de chaperona para el ARN del LINE1. El marco de lectura abierto ORF2 codifica para otra proteína, de 150kDa con actividad endonucleasa y retrotranscriptasa, requerida para la propagación de LINE1 (J. V. Moran, Kazazian, y Boeke 1996; Hancks y Kazazian 2016). Recientemente se ha encontrado otro ORF denominado ORF0 en la cadena complementaria (Hancks y Kazazian 2016). En la Figura 4 se muestra una versión esquemática de un LINE1.

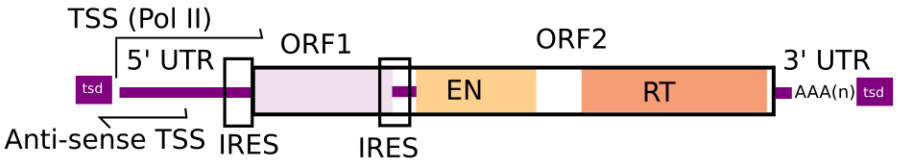


Figura 4: Esquema de un retrotransposón LINE1 de humanos.
El promotor del LINE1 coincide con la región 5' UTR (900 pb), el cuerpo está formado por los ORF1 y ORF2. TSS: Sitio de inicio de la transcripción, Pol II: polimerasa II, ORF: marcos de lectura a abiertos. IRES: Sitio interno de entrada al ribosoma, UTR: regiones sin transcribir, tsd: sitios de duplicación en tándem.

De las más de 500,000-850,000 copias de LINE1 presentes en nuestro genoma, se estima que al menos unas 100-200 son funcionales, ya que la mayoría han perdido su actividad debido a mutaciones puntuales, inversiones o truncamientos, por lo general hacia la de la región UTR 5' (Beck et al. 2010; Tubio et al. 2014; Rodriguez-Martin et al. 2020). Su movilización es una fuente importante de variación estructural, tanto en cáncer (Rodriguez-Martin et al. 2020) como a nivel germinal. De hecho, se estima que bajo condiciones normales hay una media de 285 diferentes inserciones de estos elementos entre dos individuos no emparentados (Hancks y Kazazian 2016; Sultana et al. 2019).

3.2.3 Retrotransposición de LINE1

El mecanismo clásico de retrotransposición, también llamado de "copia y pega", requiere de la transcripción de un fragmento de ADN a una cadena de RNA, su posterior retrotranscripción a ADN complementario por medio de una reverso-tran*script*asa codificada por los propios retrotransposones y su ulterior integración de nuevo en el ADN hospedador. En caso de que tal reverso-tran*script*asa fuese codificada por el propio elemento, nos encontramos con retrotransposones autónomos, por ejemplo, en humanos los LINE1. La relajación de los mecanismos celulares de represión de los LINE1, principalmente la pérdida de la metilación de las islas CpG dentro de su promotor interno (Howard et al. 2008; Muotri et al. 2010; Castro-Diaz et al. 2014; Barchitta et al. 2014; Goodier 2016; F. Yang y Wang 2016), y probablemente la relajación de la cromatina (Garcia-Perez et al. 2010), posibilita la transcripción de los ORF contenidos en el cuerpo de los LINE1, por acción de la ARN polimerasa II (Athanikar, Badge, y Moran 2004; Kroutter et al. 2009). Este promotor se ubica en el UTR 5' y consta de una serie de peculiaridades que lo distinguen de la mayoría de los genes endógenos humanos, como que carece de una caja TATA y posee un sitio de inicio no convencional (5'-GGGGG-3'). El inicio de la transcripción requiere del reclutamiento de una serie de factores de transcripción, entre otros, YY1, TCF/LEF y RUNX3. Además, los LINE1 tienen un

promotor anti sentido (ORF0) probablemente relacionado con el control de la abundancia de los transcritos de LINE1 (Denli et al. 2015; X. Zhang, Zhang, y Yu 2020). Normalmente, la terminación de la transcripción de LINE1 ocurre a la altura de la cola de poli adeninas (poliA) (AATAAA) localizada en la región UTR 3' del transposón.

El transcrito primario se traslada al citoplasma, donde los dos ORFs (ORF1 y ORF2) se traducen a sus correspondientes proteínas. A continuación, se forma un complejo ribonucleoproteico (RNP) entre las proteínas de nueva traducción y su ARN. Esta ribonucleoproteína es entonces importada al núcleo por un mecanismo aún por dilucidar, ya que su tamaño es notablemente superior al diámetro del poro nuclear. Se ha propuesto que su acceso a la cromatina genómica ocurriría durante la división celular, cuando se disgrega la membrana nuclear. Una vez en el núcleo se produce la retrotranscripción del ARN en ADN complementario, y su inserción en el genoma (E. M. Ostertag y Kazazian 2001). Errores en la actividad reverso trans*cript*asa de LINE1 pueden resultar en la integración aberrante de transcritos celulares endógenos, dando lugar a pseudogenes.

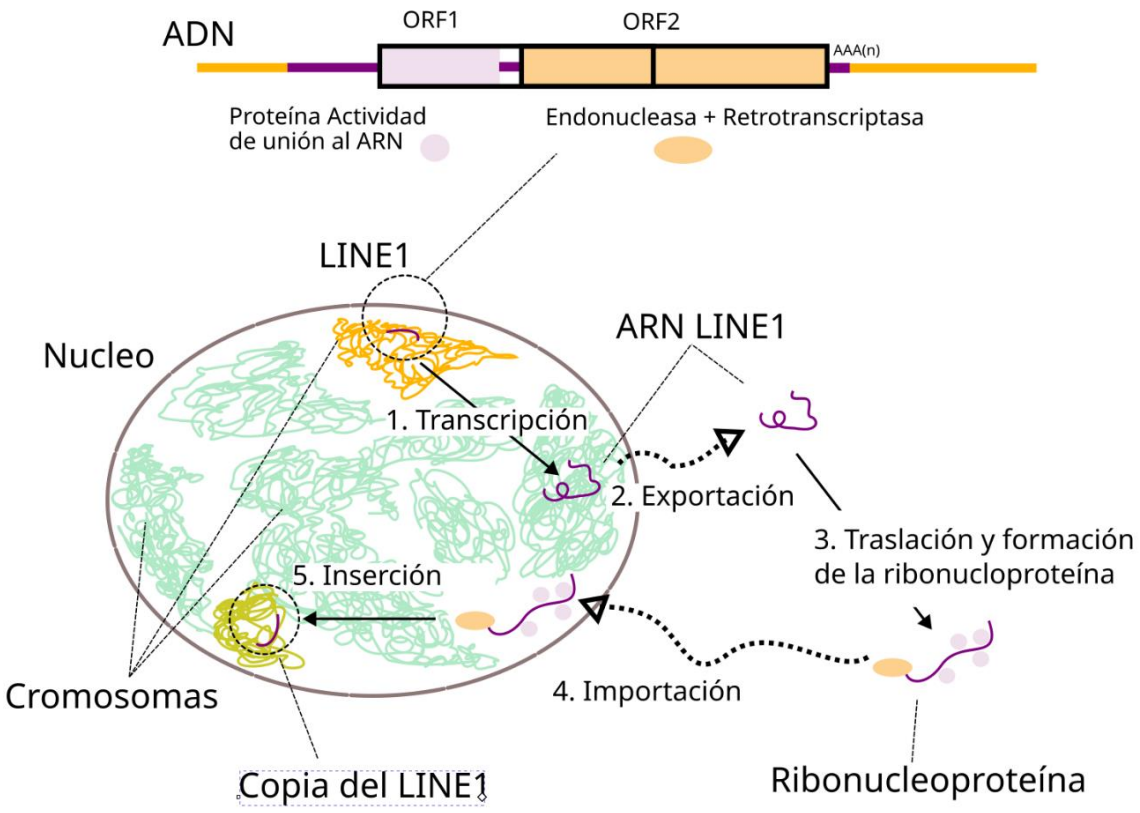


Figura 5: Proceso de retrotransposición de un LINE1.
Adaptada de (Bodak, Yu, y Ciaudo 2014)

Inicialmente no estaban muy claros qué factores influyen en el locus de inserción. Los estudios pioneros propusieron una relación con el estado de la cromatina (nivel de accesibilidad) e interacciones proteína-proteína con distintos factores nucleares (Hancks y Kazazian 2016). Evidencias posteriores (Sultana et al. 2019) sugieren, no obstante, que podría existir cierta especificidad en los sitios de inserción en regiones de baja densidad de nuleosomas. La inserción suele producirse en un sitio consenso 5′-TTTT/AA-3′ de la

cadena complementaria, tal y como evidencian estudios DNA aislado de la cristalización de la endonucleasa del LINE1 (J. V. Moran, Kazazian, y Boeke 1996; Miller et al. 2021). Son también comunes las variantes del tipo 5′-YYYY/RR-3′ (Y = pirimidina, R = purina) (J. V. Moran, Kazazian, y Boeke 1996; Cost et al. 2002; Monot et al. 2013; Hancks y Kazazian 2016). Estos motivos están mayoritariamente en fragmentos de ADN linker fuera de los nucleosomas, y de ahí se cree la preferencia de los LINE1 por insertarse en las regiones inter-nucleosomales (Sultana et al. 2019). Aunque técnicamente la inserción puede producirse en cualquier punto de la cromatina, existe cierto sesgo hacia las regiones de replicación temprana. Asimismo, y dado que es de esperar que la fitness celular se vea afectada por los loci de inserción de las nuevas copias de LINE1, linajes celulares con nuevas inserciones somáticas de LINE1 estarán sujetos a procesos de expansión clonal diferencial controlados por selección natural, perfilando la distribución global de copias LINE1 en el genoma hospedador (Tubio et al. 2014). Estos procesos de selección natural y purgado de clones con inserciones deletéreas operan tanto en el desarrollo somático normal como durante la evolución tumoral, si bien la literatura disponible se ha centrado mayoritariamente en comprobar la implicación de LINE1 en los procesos tumorales. De hecho, estudios previos de nuestro grupo de investigación han constatado una sobrerrepresentación de inserciones somáticas en la heterocromatina y las zonas intergénicas (Tubio et al. 2014). Si las nuevas copias de LINE1 se han transcrito e insertado íntegramente y los mecanismos celulares de represión así lo permiten, dispondrán de capacidad plena de producir a su vez nuevas copias (Tubio et al. 2014).

3.2.3.1 Modelos de inserción de LINE1

Son varios los modelos por los cuales LINE1 puede insertarse en el genoma hospedador en función de qué parte de la molécula de DNA emplea la retrotranscriptasa para iniciar (en adelante *primar*) la retrotranscripción del transposón: La denominada transcripción reversa primada por diana (en adelante TPRT, acrónimo de su denominación inglesa *target-primed reverse-transcription*) (John V. Moran y Gilbert 2007), considerado el más prevalente y "canónico", el primado gemelo (TP, *Twin priming*), variantes del primado gemelo y el primado dual (*Dual Priming*) y mecanismos de inserción independientes de la endonucleasa (Morrish et al. 2002). Describiremos con detalle los dos primeros, puesto que son los más frecuentes y mejor estudiados.

La TPRT (Cost et al. 2002) (Figura 6, panel a) se inicia con el mellado (SSB, *Single Strand Break)* de la cadena negativa (-) en el loci de inserción, generalmente en un motivo rico en timidina, por ejemplo, 5'-TTTT/A-3'. Esto libera una región de poliT a la cual se asocia, por complementariedad, el fragmento de poliA terminal de un mRNA de LINE1. La retrotranscripción se inicia entonces primada por el extremo 3' OH libre en la mella, generando la cadena de ADN complementaria al transcrito de LINE1 (ADNc). A continuación, se induce una segunda hidrólisis en la cadena positiva (+) del ADN diana, liberando un nuevo radical OH libre en 3' que prima la síntesis de la cadena opuesta al ADNc de LINE1 en formación, en un proceso menos estudiado y conocido. El intermediario resultante de este proceso se resuelve con la escisión e inserción del nuevo elemento en el loci diana, incluyendo la generación de pequeñas duplicaciones en tándem en el punto de inserción (TSD, *Tandem Site Duplications*) y, menos frecuentemente,

deleciones (Goodier y Kazazian 2008). Frecuentemente la retrotranscripción no recrea completamente y, como resultado, las copias derivadas suelen estar truncadas hacia 3', con una media de 900 pares de bases de longitud.

Alternativamente, el mecanismo de TP (figura 6, Panel b) requiere de la acción de dos reversotranscriptasas que actúan simultáneamente sobre un mismo loci que ha sufrido una rotura de cadena doble (DSB, *Double Strand Break*). La cadena negativa (-) sigue un proceso de reverso-transcripción similar al descrito anteriormente, en tanto que el extremo 3'-OH libre de la segunda rotura en la cadena positiva (+) prima otra reverso transcripción de LINE1 en sentido inverso. Se produce el apareamiento en las zonas complementarias de las dos hebras de LINE1 en síntesis, se eliminan los ARN en ambos heteroduplex y, por último, se sintetiza el ADN complementario de aquellas regiones de cadena simple. Este modelo es propenso a generar inversiones en el cuerpo de LINE1 (Eric M. Ostertag y Kazazian 2001).

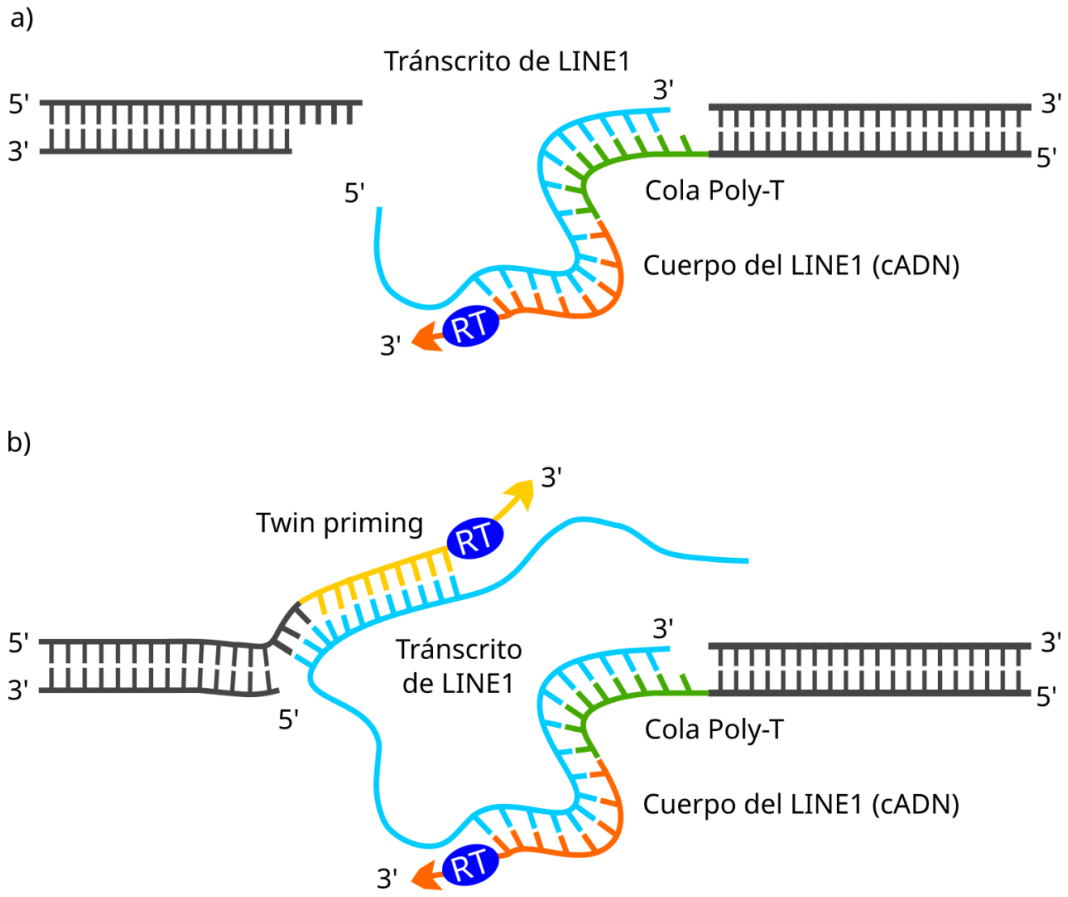


Figura 6: Modelos de inserción de un LINE1.
(a) target-primed reverse-transcription y (b) 'Twin priming'. Adaptado de (Nam et al., 2022)

3.2.3.2 Transducciones

En el supuesto de que la transcripción de LINE1 se prolongue más allá de la señal de terminación de poliadeninas, se generará un transcrito quimérico entre LINE1 y la región del genoma hospedador inmediatamente a 3' que la reverso transcriptasa podrá insertar en una nueva localización genómica (Hancks y Kazazian 2016; Tubio et al.

2014). Este proceso, por el cual parte de la secuencia corriente abajo de un transposón es copiada en otra parte del genoma debido a la retrotransposición de un LINE1 (que puede incluir o no el propio LINE1 o parte de él) se denomina transducción, en este caso hacia 3' (Moran et al., 1999). Del mismo modo, y aunque mucho más infrecuentes que las transducciones 3', si la retrotransposición se inicia *upstream* (aguas arriba) del punto teórico de inicio de transcripción de LINE1, tal región genómica podrá insertarse en una nueva posición genómica en lo que se conoce como transducción 5' (Beck et al. 2011). Las transducciones tienen una importancia crucial en la determinación de la actividad de los LINE1 y para la determinación inequívoca de la procedencia de las copias derivadas de un elemento fuente dado. Nuestro grupo de investigación ha desarrollado un algoritmo (denominado ***Source Inference,*** descrito en el apartado 8.3) para inferir la procedencia de las inserciones somáticas de LINE1 basado en nucleótidos diagnóstico, aunque carezcan de transducción, pero no es tan preciso, al ser un algoritmo probabilístico.

3.2.3.3 Tipos de retrotransposición

Se pueden definir distintos tipos de inserciones de LINE1 en función de si éste ha sido retrotranscrito de forma completa o parcial y en compañía o no de un fragmento del ADN genómico adyacente a su elemento fuente (Tubio et al. 2014). Aquellas conocidas como inserciones '*solo*' o '*solo-L1*' (TD0) incluyen exclusivamente la secuencia de LINE1, ya sea ésta una copia íntegra de su elemento fuente o una versión parcial o reorganizada del mismo. Las transducciones 3' acompañadas de una copia de LINE1, ya sea ésta completa o parcial, se identifican como '*partnered*' (TD1), en tanto que aquellas transducciones que han sufrido el truncamiento completo de la secuencia de LINE1 apreciándose únicamente la región transducida huérfana se llaman '*orphan*' u '*orphan-L1*' (Figura 7).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

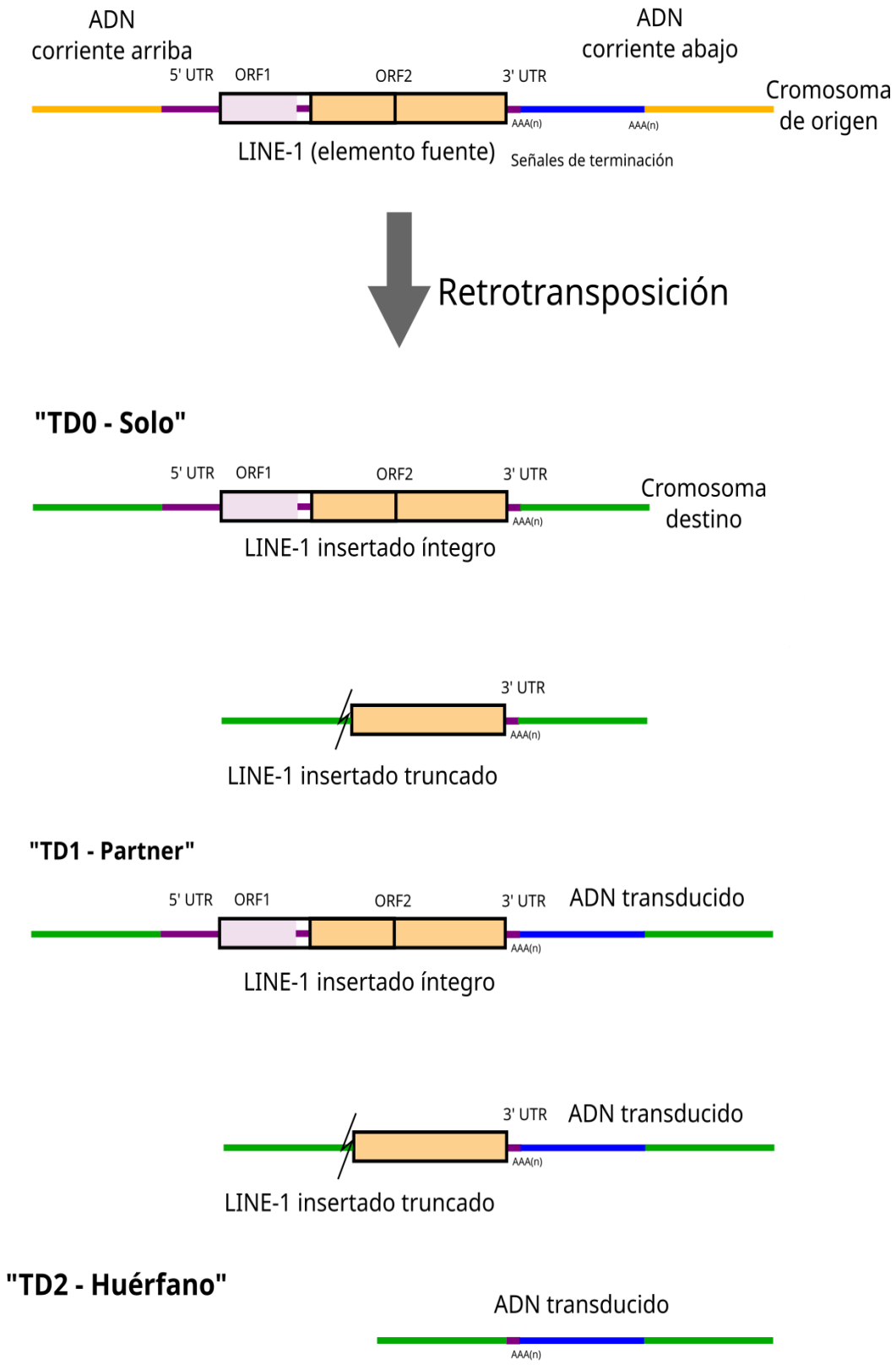


Figura 7: Tipos de retrotransposición.

3.3 MUTACIONES, ENFERMEDAD Y CÁNCER

Bajo el término cáncer se engloba una plétora de enfermedades caracterizadas por la proliferación anormal y descontrolada de un conjunto de células transformadas. En condiciones normales nuestro cuerpo actúa como un ecosistema multicelular en el que sus integrantes, las células, cumplen distintas funciones especializadas en sus correspondientes tejidos y órganos, actuando de forma colaborativa y coordinada. Esto se consigue a través de un intrincado sistema de señales extracelulares que todas las células emiten, reciben e interpretan, adaptando su comportamiento en consecuencia. Así las células actúan de forma socialmente responsable, manteniéndose en un estado quiescente, creciendo, diferenciándose, dividiéndose o muriendo, según las necesidades puntuales del organismo. Esto se consigue a través de un estricto sistema de rutas reguladoras intracelulares que controlan el metabolismo y el ciclo celular, la diferenciación y la apoptosis, lo que permite mantener la correcta homeostasis somática. Alteraciones en estos mecanismos de control pueden desembocar en el crecimiento descontrolado de algunas células, lo que puede, potencialmente, desembocar en un proceso tumoral. Tales alteraciones, habitualmente mutaciones genómicas, aunque no exclusivamente, se acumulan progresivamente en pasos incrementales durante la progresión tumoral, requiriéndose del orden de entre cuatro y siete de estos eventos para desencadenar el cáncer (Hanahan y Weinberg 2000; 2011). Estas alteraciones suelen conferir ventajas a las células que las poseen, estableciéndose una evolución Darwiniana, que puede llevar paulatinamente a un conjunto de células sanas a convertirse en una masa tumoral y posteriormente cancerígena. Tal proceso puede considerarse análogo al que da lugar a nuevas especies y, como tal, se basa en la preexistencia de mutaciones más o menos aleatorias con efecto fenotípico en la *fitness* celular y sometidas a un filtrado por selección natural (Stratton, Campbell, y Futreal 2009).

Todas las células del cáncer provienen del óvulo fertilizado a través de una serie de divisiones mitóticas, al igual que las del resto del organismo. A lo largo de la vida del individuo, sus células adquieren una serie de mutaciones no presentes en el conjunto de variantes que presentaba el zigoto. Estas variaciones se denominan *mutaciones somáticas*, en contraposición a las *mutaciones germinales*, o aquellas que se heredan de los progenitores. Las mutaciones somáticas a su vez se dividen en conductoras (o "*driver*") y pasajeras. Las primeras les confieren a las células que las contienen alguna ventaja evolutiva que permite o facilita su supervivencia y proliferación, mientras que las segundas no alteran la fitness celular, pero aparecen en una célula que coincidente o subsecuentemente adquieren una mutación driver. La aparición de mutaciones driver funda un clon de células que, si las circunstancias son propicias, puede expandirse clonalmente sobre el resto de las células de su entorno.

La sucesiva acumulación de mutaciones driver en un mismo clon celular favorece su metaplasia y anaplasia, pudiendo concurrir en una masa tumoral benigna y, en última instancia, un tumor (figura 8).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

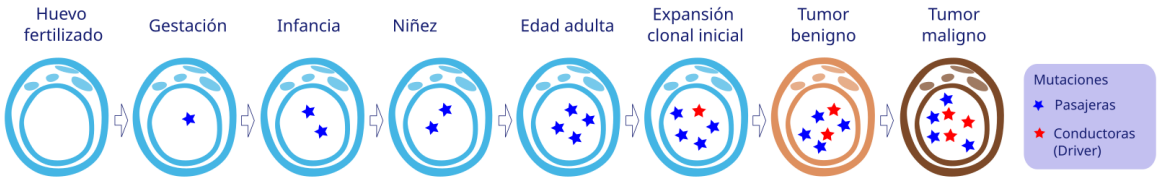


Figura 8: Evolución de las células cancerosas.
(Adaptada de Stratton et al., 2009)

Evidentemente, de entre todo el elenco de mutaciones algunas pueden participar de la activación o inactivación efectiva de genes, incluyendo SNVs (*Single nucleotide variants*), es decir la sustitución de un nucleótido por otro, inserciones o deleciones de pequeños fragmentos de ADN (*indels*), variación estructural, donde se producen inserciones o deleciones de grandes fragmentos del ADN, reestructuraciones consistentes en la fragmentación del ADN y su posterior reensamblaje (*i.e. cromotripsis*), alteraciones en el número de copias de fragmentos del genoma o inserciones de ADN exógeno, por ejemplo el de ciertos virus, algunos de los cuales son conocidos causantes de cánceres (Stratton, Campbell, y Futreal 2009). Las células tumorales sufren, además, una profunda reprogramación en todos los niveles de regulación epigenética, ya sean pre-transcripcionales, entre los que se cuentan alteraciones en el grado de condensación y accesibilidad de la cromatina, el estado de metilación del ADN e histonas, o postranscripcionales, tales como la abundancia y expresión de ARN de interferencia y microARN, contribuyendo a exacerbar el proceso canceroso. Estos cambios genéticos y, en principio, la mayoría de epigenéticos son heredables y su prevalencia en la comunidad de células tumorales está sujeta a una evolución darwiniana a pequeña escala (Stratton, Campbell, y Futreal 2009).

Aunque la naturaleza de las mutaciones y las dinámicas de cambios adquiridos por un tumor son muy variopintas, existe un conjunto de alteraciones recurrentes por las cuales una célula normal adquiere las ventajas necesarias para constituir un tumor. Algunas de ellas serían la autosuficiencia respecto a las señales proliferativas y de diferenciación celular,

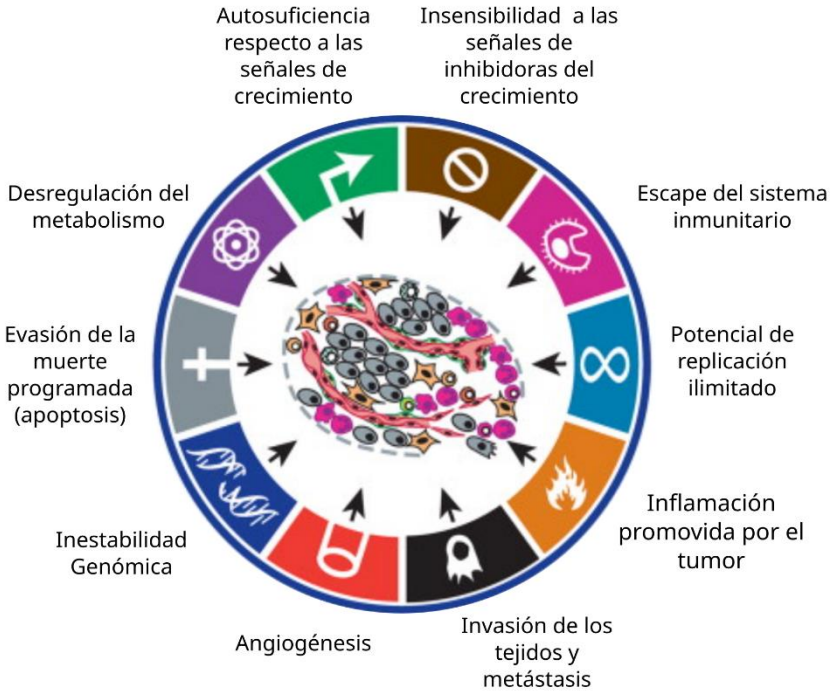


Figura 9: "Hallmarks of cancer".
(Adaptada de Hanahan and Weinberg, 2011)

insensibilidad a las señales de inhibidoras del crecimiento, evasión de la muerte programada por apoptosis, potencial replicativo ilimitado, inducción de la angiogénesis, capacidad invasiva de tejidos y metástasis, desregulación del metabolismo, escape del sistema inmunitario e inestabilidad genómica (Figura 9). Además, y aunque no son estrictamente características de las células tumorales, su capacidad de interacción con el entorno puede favorecer o posibilitar la aparición del cáncer, ya sea promoviendo inflamación local o alternando el microentorno del tumor para beneficiar su desarrollo (Hanahan y Weinberg 2000; 2011).

3.4 LINE1, VARIACIÓN NORMAL, ENFERMEDAD Y CÁNCER

La visión clásica de los transposones como parásitos en el genoma o material inerte no funcional como parte del mal llamado ADN basura, entra en conflicto con el hecho de que sean tan abundantes y que no se hayan eliminado en el curso de la evolución de los genomas. A día de hoy creemos que los transposones desempeñaron un papel muy importante en la evolución y radiación genómica, tanto de nuestra especie como de los eucariotas en general, ya desde sus orígenes más tempranos hace más de 600 millones de años en la era Precámbrica (Goodier y Kazazian 2008). Varios serían los mecanismos por los cuales pueden, potencialmente, modular la evolución de los genomas (ver figura 10): afectando a la función génica con sus inserciones *de novo*, lo que puede alterar el fitness a nivel individual o celular y generar linajes evolutivos potencialmente sujetos a selección. Induciendo reorganizaciones cromosómicas que actúen como barreras reproductivas por incompatibilidades meióticas. Como una nueva fuente de material genético, tales como nuevos genes o su combinación con otras secuencias reguladoras en cis y en trans. En este sentido (X. H.-F. Zhang y Chasin 2006) comprobaron que las regiones altamente repetitivas (sobre todo Alus), son la principal fuente de nuevos genes en los seres humanos y ratones. Finalmente, la movilización de transposones puede provocar cambios epigenéticos en los puntos de inserción, tales como hipermetilación, heterocromatinización o alteración de los dominios topológicos de regulación del ADN, afectando, potencialmente, a la expresión génica (Feschotte y Pritham 2007). Además, dado que la presencia de retrotransposones en mamíferos se remonta a varios cientos de millones de años atrás, es presumible que una relación huésped-parásito tan longeva haya propiciado que nuestras células desarrollasen una variedad de estrategias para "*domesticar*" a los elementos transponibles de forma que favorezcan sus necesidades genéticas y adaptativas (exaptación). Es sabido, de hecho, que LINE1 juega un papel relevante en la embriogénesis y desarrollo temprano de mamíferos. En concreto, se ha observado que la reprogramación genética del cigoto hacia células pluripotenciales tras la fertilización coincide con una repentina activación de LINE1, lo que, teóricamente, contribuiría a mantener desplegada la cromatina para que fuese accesible a la maquinaria proteica de transcripción (Jachowicz et al. 2017). Se cree además que LINE1 juega papel fundamental en la compensación de dosis por inactivación del cromosoma X en hembras de mamíferos placentarios, tanto en la represión de la expresión génica por heterocromatinización como en el escape de algunos genes en este proceso (Chow y Heard 2009). Asimismo, los LINE1 son imprescindibles en la neurogénesis de primates (Suarez, Macia, y Muotri 2018).

USC DE COMPOSTELA

Además de la influencia en el proceso evolutivo y en la embriogénesis, la expresión de LINE1 en la línea somática puede tener un impacto perjudicial en el individuo y producir enfermedades y cáncer (Burns 2017). Por ejemplo, la sobreexpresión de la proteína ORF2 de LINE1 en conjunción a la pérdida de actividad exonucleasa 3′ TREX1 favorecen sinérgicamente la acumulación de ADNc en el citoplasma, induciendo la activación de rutas inflamatorias dependientes de interferón (Schulz 2006). Esto sucede tanto en el proceso normal de senescencia como en el síndrome de Aicardi-Goutières, una encefalopatía hereditaria caracterizada por la calcificación de ganglios basales, leucodistrofia y linfocitosis cefalorraquídea (Thomas et al. 2017; De Cecco et al. 2019). Asimismo, desde el hallazgo a finales de los ochenta por el grupo de Kazazian de que las inserciones de LINE1 pueden producir enfermedades (Haig. H. Kazazian et al. 1988), se han descrito múltiples instancias en las que LINE1 es el causante directo de diversas enfermedades, tales como la hemofilia, la Beta-talasemia, la distrofia muscular de Duchenne, la fibrosis quística, el síndrome de Apert, la neurofibromatosis y el cáncer (Goodier y Kazazian 2008; Hancks y Kazazian 2016). Algunos de estos mecanismos son descritos en (H. H. Kazazian 2004) de ellos, el mecanismo más simple e intuitivo por el cual una copia *de novo* de LINE1 puede conducir a enfermedad (Figura 10, panel a) es el truncamiento genético y la disrupción del tránscrito provocado por la propia inserción, si bien no es la única posibilidad. Del mismo modo, las movilizaciones de LINE1 concurren con reorganizaciones en los loci de inserción (Figura 10, panel b) que pueden eliminar tanto elementos reguladores de la actividad genética como "enhancers" o promotores, como exones y/o intrones, algo similar ocurre con las inversiones pues sus puntos de ruptura pueden alterar mecanismos de regulación. Estas reorganizaciones pueden producir también duplicaciones de material genético, que sí coinciden en oncogenes, pueden producir una sobreexpresión de estos. Evidencias cada vez más robustas sugieren que LINE1 media en la formación de cromosomas derivativos, ya sea por recombinación entre no homólogos (Goodier y Kazazian 2008) o por translocación directa con un puente de LINE1 (Figura 10, panel b), generando, potencialmente, genes de fusión y/o nuevas interacciones entre elementos reguladores y genes distantes (Rodriguez-Martin et al. 2020). Las transducciones, tanto hacia 3' (Figura 10, panel c) como hacia 5', pueden arrastrar moduladores de la expresión hacia los nuevos loci de inserción pudiendo interactuar éstos en cis o en trans con genes colindantes y favoreciendo bien su sobreexpresión, como es el caso de oncogenes, bien su infra expresión o inactivación completa, como en el caso de genes supresores de tumores (J. V. Moran, DeBerardinis, y Kazazian 1999). Las inserciones intrónicas de LINE1 pueden también interferir también en la expresión, (Figura 10, panel d) bien forzando la terminación de la transcripción, provocando *splicing* alternativo en el procesado del transcrito primario o interrumpiendo el marco de lectura abierto por *exonización* del propio inserto de LINE1, es decir, la incorporación de parte de su secuencia en ARNm celulares endógenos (Gonçalves et al. 2017). La disponibilidad de la maquinaria proteica de retrotransposición por activación de LINE1 provoca la movilización de elementos no autónomos como SINEs, Alus o pseudogenes. La movilización de Alus terciada por LINE1 (Figura 10, panel e) puede, además, provocar expansiones de microsatélites, lo que contribuye a generar mayor inestabilidad genómica (Dewannieux, Esnault, y Heidmann 2003). En teoría, además, la retrotransposición somática puede modular la expresión génica por la disrupción de los dominios asociados topológicamente o (TAD, por sus siglas en inglés: *Topologically*

Associated Domain), unidades de compartimentalización de la cromatina de regulación conjunta (Figura 10, panel f). Finalmente, otro mecanismo por el cual LINE1 modula la expresión génica es la extensión de la heterocromatina en torno a los loci de inserción, por lo general a través de marcas epigenéticas como modificaciones de histonas o metilación. Esta última directamente relacionada con el objetivo de esta Tesis, puesto que la alteración de los patrones de metilación de las regiones adyacentes a la inserción del LINE1 que puede tener función reguladora (Figura 10, panel g), esto es debido a que los LINE1 este contiene una región muy rica en CpG (isla CpG).

En conjunto, todos los mecanismos mencionados contribuyen de forma directa a alterar la actividad génica e, indirectamente, a generar mayor inestabilidad genómica. Ello altera la fitness celular y favorece la fundación de linajes potencialmente sujetos a procesos selectivos, lo que conduce a una expansión clonal que, eventualmente, puede desembocar en enfermedad o la génesis de un tumor.

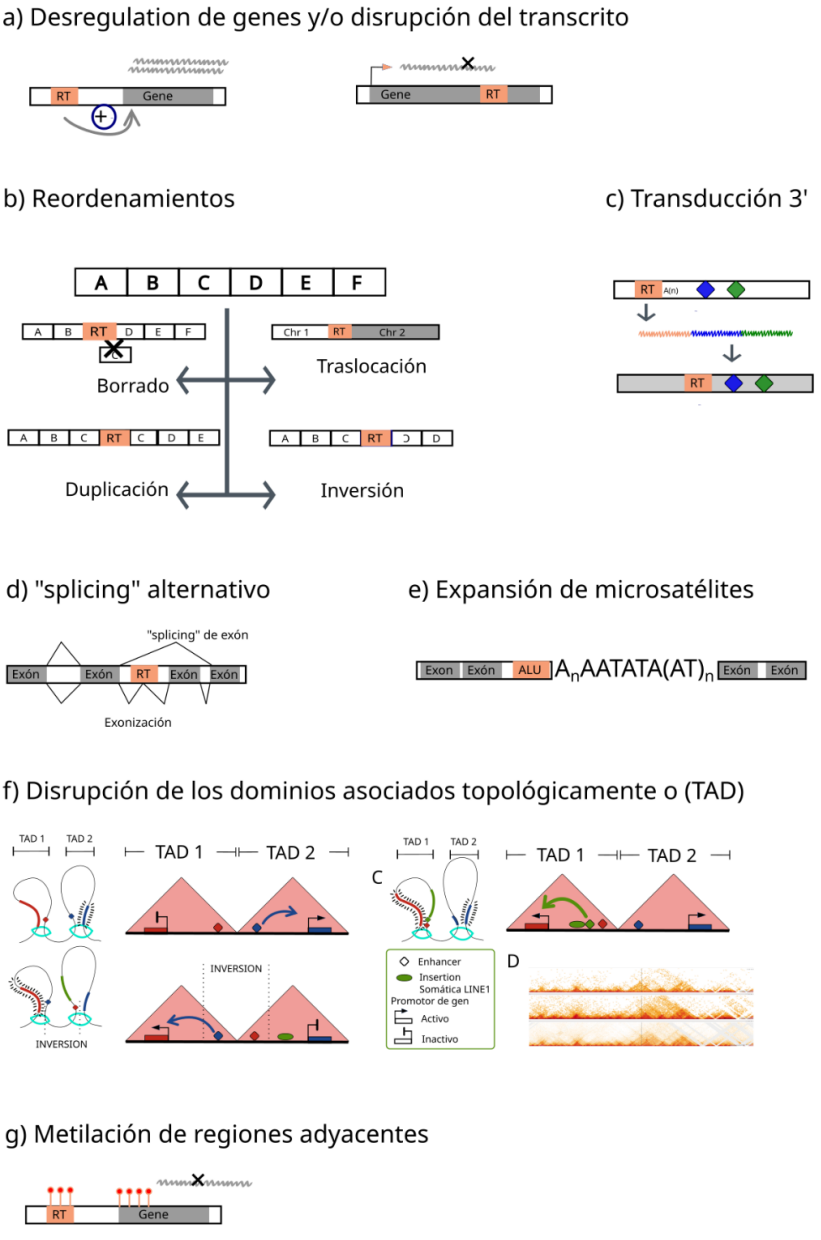


Figura 10: Posibles mecanismos de alteración de la actividad génica.

Aunque se ha demostrado que los eventos de retrotransposición en la línea germinal son relativamente frecuentes y que aproximadamente uno de cada 200 neonatos presenta una inserción *de novo* de *LINE1* no presente en el soma de sus progenitores (Cordaux y Batzer 2009; Xing et al. 2009), incrementando en algunos casos la predisposición a padecer ciertos tipos de cánceres (Qian et al. 2017; Sabatella et al. 2021; Rahrmann y Largaespada 2013), la mayor contribución de los LINE1 al proceso tumorigénico se produce por su retrotransposición en la línea somática (E. Lee et al. 2012; Helman et al. 2014; Tubio et al. 2014; Rodriguez-Martin et al. 2020). Se estima que la tasa media de movilización *de novo* de LINE1 en la línea somática es de 0.028 inserciones por año (Nam et al. 2022) si bien, al igual que la expresión génica, los patrones de expresión de LINE1 son específicos de tipos celulares, enfermedad y tumor. De hecho, es bien sabido que las movilizaciones de LINE1 son muy frecuentes en carcinomas epiteliales, especialmente en cáncer de esófago, colorrectal, de cabeza y cuello y adenocarcinoma pulmonar, frecuentes, aunque en menor medida, en cáncer de próstata y ovario e inapreciables en linfomas, leucemias y cánceres cerebrales (E. Lee et al. 2012; Helman et al. 2014; Rodriguez-Martin et al. 2020). Algunos transposones son excepcionalmente activos (Tubio et al. 2014; Rodriguez-Martin et al. 2020) y son capaces de generar muchos tipos de alteraciones genómicas que van desde sutiles mutaciones en las zonas reguladoras a grandes reorganizaciones cromosómicas. En todo caso, es difícil discernir el grado de participación de la retrotransposición somática en la carcinogénesis, ya sea como mutaciones *driver* fundadoras o consecuencia derivada de la inestabilidad genética del tumor (Burns 2022), si bien en casos puntuales se ha demostrado que las mutaciones driver inducidas por LINE1 se producen muy tempranamente en la génesis del tumor (Rodriguez-Martin et al. 2020).

3.4.1 Regulación de la retrotransposición somática

Debido a los potenciales efectos deletéreos y desestabilizantes que los transposones sobre la homeostasis celular, la presión selectiva resultante ha favorecido el desarrollo de una intrincada red de mecanismos transcripcionales y postranscripcionales que actúan sinérgicamente para reprimir la transposición somática. Por ejemplo, algunos ARN pequeños (sRNAs, *Small RNAs*), como el ARN pequeño interferente (siRNA, *small-interference RNA*)(Chen et al. 2012), ARN-PIWI (*PIWI-interacting RNAs,* piRNAs) o micro ARN, (miRNAs) (Siomi et al. 2011; Iwasaki, Siomi, y Siomi 2015; Goodier 2016; Protasova, Andreeva, y Rogaev 2021), contribuyen a modular la transposición tras la transcripción de los elementos transponibles. Los piRNAs, por ejemplo, son el principal mecanismo de represión postranscripcional de elementos transponibles en los genomas de nematodos, mosca, levaduras y plantas con flores, así como durante la gametogénesis de vertebrados. Éstos favorecen la degradación dirigida de los RNA derivados de elementos transponibles y la represión dirigida a los loci genómicos de transposones a través de varias marcas epigenéticas (Deniz, Frost, y Branco 2019). Diversos estudios han demostrado además la implicación de distintas proteínas celulares en la inhibición de LINE1, incluyendo varios miembros de la familia APOBEC (Richardson et al. 2014), proteínas estimuladas por interferón como MOV10, TREX1 o ZAP (Stetson et al. 2008), hnRNPL (Peddigari et al. 2013), SAMHD1 (Zhao et al. 2013), el supresor de tumores TP53 o la proteína de regulación de la longevidad SIRT6. En este

sentido, una ruta clave en el silenciamiento de transposones en mamíferos que ejemplifica la rápida coevolución de los mecanismos de defensa del huésped involucra a la familia de factores de transcripción con dedos de Zinc tipo Kruppel (KRAB-ZFP, *Krüppel-Associated Box Zinc-Finger protein*). La asociación de las KRAB-ZFPs a los loci de elementos transponibles recluta a otras proteínas como TRIM28 que en conjunto favorecen la formación de cromatina represiva, silenciando efectivamente los transposones en dicha región genómica (Deniz, Frost, y Branco 2019) (figura 11, panel a). Otro complejo proteico se encarga del mantenimiento del estado de metilación (figura 11, panel b).

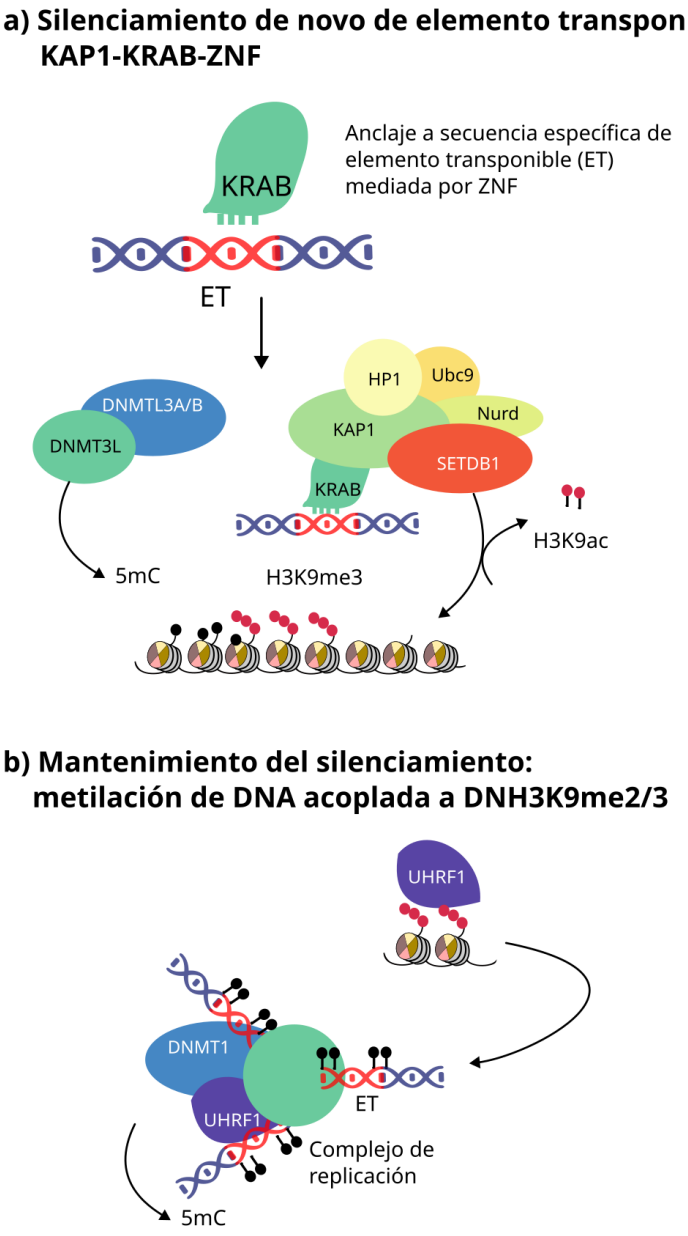


Figura 11: Mecanismos de silenciamiento del LINE1 y su mantenimiento. (Adaptada de Deniz et al 2019)

En última instancia, todos los mecanismos mencionados actúan dirigidamente sobre marcas epigenéticas a través de la adición covalente de radicales químicos (i.e.

grupos metilo (metilación), hidroximetilo (hidroximetilación), o acetilo (acetilación)) a nucleótidos o residuos de las histonas linker u octaméricas (Deniz, Frost, y Branco 2019). Variantes de histonas típicas de heterocromatina constitutiva, como por ejemplo la trimetilación en la lisina 9 de la histona H3 (*H3K9me3*), están implicadas en la compactación de la cromatina e inactivación de LINE1 (Guler et al. 2017). En todo caso, las modificaciones covalentes del ADN, y más específicamente la 5-metilcitosina (5mC), son la estrategia más ampliamente adoptada por los organismos eucariotas superiores para mantener los elementos transponibles, y por extensión los LINE1, en un estado transcripcionalmente inerte (Branco, Ficz, y Reik 2012). Se ha comprobado también que existe cierto enriquecimiento de 6-metiladenina (6mA) en las regiones 5' UTR de elementos LINE1 jóvenes en varias líneas celulares de linfoblastoma humano, presumiblemente guardando cierta correlación con su actividad, aunque con un impacto menor que la 5mC (Zhu et al. 2018).

Todos estos mecanismos son completamente reversibles y no exclusivos de la regulación de los LINE1, modulando la expresión génica en general, en lo que conocemos como epigenoma, una capa de regulación genómica superpuesta al genoma (Holliday et al. 1997; Berger et al. 2009). En conjunto, estas modificaciones epigenéticas se pueden definir como aquellos cambios en las propiedades de los ácidos ribonucleicos en general, incluyendo ADN y ARN y la cromatina que afectan a su funcionalidad sin implicar cambios en su secuencia nucleotídica (Tollefsbol 2017; Deniz, Frost, y Branco 2019). De entre ellos, el más ampliamente estudiado y, discutiblemente, relevante en la regulación de la expresión génica, es la metilación de los promotores génicos (Miousse y Koturbash 2015) y, por extensión, de las regiones promotoras de LINE1. A continuación, profundizaremos un poco en que es la metilación, y la metilación de los LINE1.

3.4.2 Metilación del ADN

La metilación del ADN es una de las modificaciones epigenéticas más importantes y quizá la más estudiada. Ésta consiste en la adición de un grupo metilo (CH_3) proveniente de una molécula de S-adenosilmetionina (SAM), en el carbono 5 de una citosina (C) de ciertos dinucleótidos CpG. La citosina se convierte entonces en una metil-citosina (C^m) (figura 12). El proceso es catalizado por las ADN metiltransferasas (*DNA methyltransferases* DNMTs). DNMT1 está implicada en el mantenimiento de los patrones de metilación tras la división celular, mientras que las DNMTs 3A y 3B se encargan de la adición de nuevos grupos metilo en zonas previamente no metiladas, es decir, generan metilación *de novo*. Este proceso de metilación afecta principalmente a la expresión génica, por norma general cuanto más metilada está la región reguladora de un gen, menos se expresa (Tollefsbol 2017; Deniz, Frost, y Branco 2019). La transición de 5mC a citosina no metilada es reversible, tanto de forma espontánea como catalizada por una serie de enzimas demetilantes TET (*Ten Eleven Translocation*) e implican el paso secuencial por los intermediarios 5 hidroximetilcitosina (5hmC), 5-formilcitosina (5fC) y 5 carboxilcitosina (5caC) y, finalmente, citosina no metilada (figura 12). En algunos casos estas modificaciones pueden ser heredadas transgeneracionalmente de células progenitoras a hijas durante la mitosis y, en ciertas instancias más excepcionales, también en la meiosis, de una generación a la siguiente (Tollefsbol 2017).

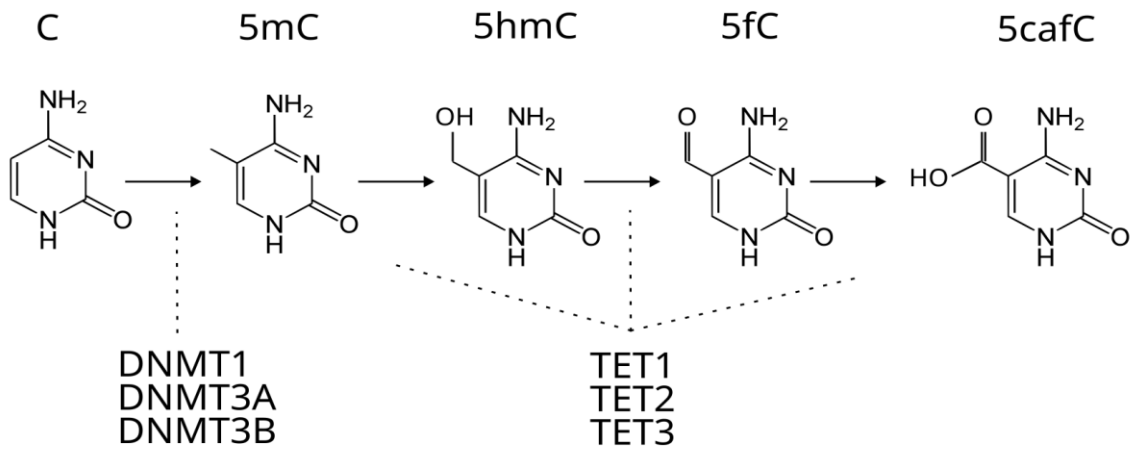


Figura 12: Fórmula estructural de la citosina, la metil-citosina y derivados.
C: citosina, 5mC: 5-metilcitosina, 5hmC: 5-hidroximetilcitosina, 5fC: 5-formilcitosina, 5cafC: 5-carboxilcitosina

En genomas de mamíferos, la mayor parte de la 5mC se localiza adyacente a residuos de guanina hacia 5', en lo que denominamos metilación de dinucleótidos CpG y, en mucha menor medida, en contextos no-CpG, como CpA, CpT y CpC. En condiciones normales, la metilación en las posiciones CpG es simétrica, es decir, las citosinas de ambas hebras de un dinucleótido CpG están simultáneamente metiladas o no metiladas, por lo que se suelen extrapolar a una única posición para su análisis. Los dinucleótidos CpG aparecen en el genoma humano con una proporción un quinto menor a la esperada por azar, seguramente por la desaminación espontánea de las citosinas metiladas hacia los dinucleótidos TpC y CpA (Simmen 2008). Además, la distribución de dinucleótidos CpG en el genoma humano no es uniforme, intercalado regiones de alta densidad de CpG con otras ampliamente deplecionadas (Gardiner-Garden y Frommer 1987). Estas áreas se conocen como "*islas CpG*" (CpG islands en su acepción inglesa) y son muy abundantes en los promotores de los genes humanos y de los LINE1.

3.4.3 Islas CpG

La definición rigurosa de las islas CpG es la siguiente: son aquellas regiones del ADN mayores de 200 pb con un contenido de GC superior al 50% donde la proporción de CpG observadas frente a CpG esperadas es mayor o igual a 0,6. Continuando la analogía con geográfica, se define como "*zonas costeras u orillas*" (shore) a las regiones de 2 KB corriente arriba y abajo (*upstream, downstream)* de las islas CpG, los dos Kb siguientes tanto corriente arriba como corriente abajo constituyen la plataforma (shelf) y el resto no contenido en las regiones anteriormente mencionadas se define como "*mar abierto*".

Figura 13: Nombres de agrupaciones de CpG.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Grosso modo, se estima que algo menos de la mitad de las más de 50,000 islas CpG en el genoma humano se encuentran en regiones de copia única o de bajo número de repeticiones, a partes iguales entre promotores génicos y regiones inter o intragénicas (Rollins et al. 2006). Muchas de estas islas CpG son importantes regiones reguladoras, ya que el 70% de los promotores de genes anotados están asociados a islas CpG, entre los que se cuenta la gran mayoría de genes *housekeeping* constitutivos (Deaton y Bird 2011). Más de la mitad de islas CpG, no obstante, se ubican en ADN repetitivo (Grandi et al. 2015). En concreto los SINEs, que constituyen el 13% del genoma humano, acaparan el 58% de las islas CpG, en tanto que los LINEs, de los que se estima abarcan poco más del 13% del genoma poseen únicamente el 8% de islas CpG, concentradas mayoritariamente sobre la región promotora de su UTR 5'. En términos generales, la metilación de las islas CpGs se considera una fuerte señal represiva, tanto de la expresión de genes endógenos como de la activación de transposones en general y de LINE1 en particular.

3.4.4 Metilación de LINE1

Bajo condiciones normales, la mayoría de LINE1 en la línea somática están fuertemente silenciados (Deniz, Frost, y Branco 2019). En términos generales se acepta que los más antiguos dentro de nuestro genoma han perdido su capacidad de movilización por acumulación de SNVs, inversiones y otro tipo de mutaciones que los inactivan *de facto*. En cambio, los más recientes en el curso de nuestra evolución como especie y todavía funcionales son reprimidos por marcas epigenéticas, principalmente por metilación de sus regiones promotoras (Deniz, Frost, y Branco 2019; Bodak, Yu, y Ciaudo 2014). Estás marcas epigenéticas son reversibles y, de hecho, se eliminan durante la movilización programada de LINE1 en el desarrollo embrionario y la neurogénesis, como ya se ha mencionado (ver figura 14). Además, los patrones de metilación de LINE1 son moderadamente específicos del tipo de tejido (Phokaew et al. 2008; Nüsgen et al. 2015). Por ejemplo, al considerar los LINE1 en masa, Chalitchagorn y colaboradores (Chalitchagorn et al. 2004) hallaron que el epitelio del colon (35.78 % de metilación promedio) y los hepatocitos (34.46 %) presentan valores de metilación de LINE1 significativamente menores que, entre otros, el epitelio esofágico (45.38 %) o la tiroides (45.07 %), lo que demuestra plasticidad en la metilación de LINE1 en tejidos adultos, presumiblemente asociada a los procesos de diferenciación celular y con funciones análogas a la reprogramación de la metilación que ocurre durante la gametogénesis y el desarrollo embrionario (Chalitchagorn et al. 2004). Estudios más recientes alcanzaron una conclusión similar al establecen que la metilación de los promotores de LINE1 es variable dependiendo de elemento, tejidos y tipos celular, estando las familias más modernas de L1HS hipermetiladas, en comparación al conjunto de subfamilias de LINE1 más antiguas (Ewing et al. 2020; Lanciano et al. 2023). Además, y si bien bajo condiciones normales la mayoría de LINE1 están constitutivamente metilados, algunos loci parecen escapar a la inactivación dependiendo de individuo y tejido (Lanciano et al. 2023). Véase, por ejemplo, el L1HS ubicado en 22q12.1 dentro de un intrón del gen TTC28, habitualmente activo en hepatomas (Pradhan et al. 2017; Schauer et al. 2018; Rodriguez-Martin et al. 2020) y hallado fuertemente desmetilado en tejido hepático de un paciente sano (Ewing et al. 2020), u otro L1HS del cromosoma 1, expresado tanto en la línea germinal como en cáncer y en fibroblastos senescentes fuertemente metilado en hipocampo, pero no en

corazón o hígado normal (Ewing et al. 2020). Todo esto implica que interpretar los resultados del metiloma de LINE1 en enfermedad y cáncer requiere un conocimiento previo de los patrones de metilación del tejido sano de origen, no pudiendo recurrir a otro tejido de referencia como la sangre, práctica habitual para discernir mutaciones somáticas de germinales en estudios de variación puntual o estructural. Además, contaminaciones cruzadas entre este "*tejido adyacente*" y el tumor, así como la presencia de subpoblaciones celulares no cancerosas en la masa tumoral, tales como estroma, vasos sanguíneos o del sistema inmunitario dificultan la interpretación directa de resultados de secuenciación masiva de masas de tejido completo (*Bulk sequencing*) (Matsunoki et al. 2012) .

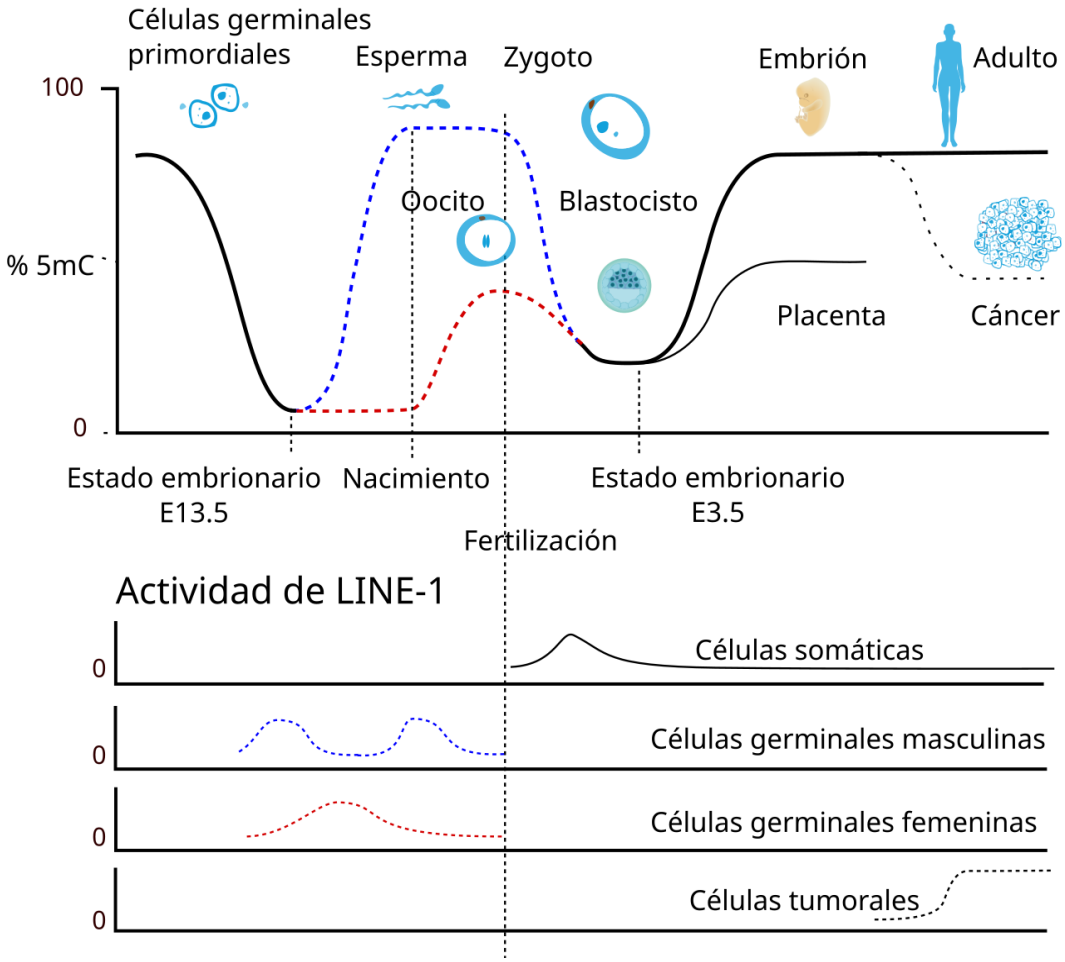


Figura 14: Actividad del LINE1.
(Adaptada de Deniz et al. 2019)

Las células tumorales sufren una profunda reprogramación epigenética, cuya principal característica es la pérdida global de metilación en su genoma, que se ha propuesto que promueven la oncogénesis por activación selectiva de proto-oncogenes. Esta hipometilación también suele abarcar a los elementos transponibles en general y a los LINE1 en particular, que se activan por desmetilación de sus regiones promotoras, particularmente en carcinomas (Thayer, Singer, y Fanning 1993; Tubio et al. 2014; Nüsgen et al. 2015; Sharma et al. 2019; Deniz, Frost, y Branco 2019). El grado de metilación de los LINE1 varía dependiendo del tipo de tumor, mostrando valores más bajos, entre otros, en los cánceres colorrectales (Lao y Grady 2011; Antelo et al. 2012),

tumores neuroendocrinos pancreáticos y carcinoides (Choi et al. 2007), carcinoma hepatocelular (Anwar et al. 2019), cáncer de mama (van Hoesel et al. 2012), carcinoma de urotelio (Jürgens, Schmitz-Dräger, y Schulz 1996), colorrectal (Ogino, Nosho, et al. 2008), de próstata (Florl et al. 2004) y de estómago (Shigaki et al. 2013). De entre todos ellos, aquellos con menores niveles de metilación en sus LINE1 son los cánceres urinarios y de vejiga, seguidos por los de próstata, colon y estómago (Nüsgen et al. 2015). La hipometilación de LINE1 sido propuesta como posible marcador pronóstico del curso de la enfermedad: Por ejemplo, (Shigaki et al. 2013) observaron una correlación positiva de los niveles de metilación en LINE1 con la supervivencia del paciente en una cohorte de más de 200 casos de cáncer gástrico. Este fenómeno ha sido subsecuentemente corroborado en cáncer de pulmón, hígado, esófago, próstata y endometrio (Kawano et al. 2014). Además, existe una correlación lineal entre la hipometilación de LINE1 y la agresividad en distintos tipos de tumores, como, por ejemplo, cáncer de colon (Ogino, Kawasaki, et al. 2008). La hipometilación de LINE1 también es habitual en los adenomas asociados a cáncer colorrectal, cuyos niveles de metilación de LINE1 son indistinguibles de los tejidos cancerosos y significativamente más bajos que el tejido normal adyacente (Shademan et al. 2020). Por lo tanto, la hipometilación de LINE1 también se puede usar como un biomarcador temprano para el cáncer. En apariencia los patrones de metilación de LINE1 son constantes dentro de una misma masa tumoral, a lo largo de su progresión e, inclusive, entre un tumor primario y sus metástasis (Matsunoki et al. 2012; Murata et al. 2013), lo que sugiere que las células tumorales hipometilan los LINE1 y su genoma en general durante sus fases tempranas de desarrollo y que esta característica se conserva a lo largo de la progresión tumoral y metástasis (Lao y Grady 2011).

Múltiples artículos han establecido que las nuevas copias surgidas de la retrotransposición somática de LINE1 son frecuentemente silenciadas, si bien las dinámicas por las cuales se produce esta represión no son del todo evidentes (Garcia-Perez et al. 2010; Kannan et al. 2017). Esos últimos (Kannan) abordan esta cuestión estudiando la retrotransposición somática de LINE1 en distintas líneas celulares de tumores humanos y embrionarias de ratón. Estos investigadores concluyen que la metilación *de novo* juega un papel fundamental en la represión de las nuevas inserciones de LINE1 en ratones, si bien los nuevos integrantes de LINE1 en las líneas tumorales humanas permanecen casi completamente desmetilados a lo largo de multitud de divisiones celulares, siendo silenciadas por desacetilación de histonas (Kannan et al. 2017). Una posible explicación sería que el silenciamiento a corto plazo de las nuevos insertos de LINE1 se produciría por actividad de enzimas deacetilasas de histonas, sin la participación inmediata de la metilación del ADN, que actuaría como una capa adicional de silenciamiento más permanente adquirida a posteriori (Garcia-Perez et al. 2010). Por otra parte, cabe destacar que estas células tumorales humanas están altamente diferenciadas y que, por consiguiente, carecen ya de la actividad metiltransferasa *de novo* necesaria (DNMT3) para ejercer la metilación de LINE1, lo que explicaría su aparente falta de implicación en el silenciamiento de los nuevos insertos (Garcia-Perez et al. 2010).

Varios estudios ponen de manifiesto el impacto de LINE1 en los perfiles epigenéticos colindantes a sus respectivos loci. En un interesante estudio en ratones, Grandi (Grandi et al. 2015) observan que la hipometilación en las inserciones somáticas de LINE1 afecta gradualmente a los patrones de metilación de las CpG en las regiones

flanqueantes (+- 2 Kb), progresivamente menos metiladas cuanto más próximas al sitio de inserción. Para ello comparan mediante BS-NGS niveles de metilación en loci heterocigotos para LINE1, es decir, aquellos que poseen un alelo (alelo lleno) con una inserción somática frente a otros que no la presentan (alelo vacío). El estudio es limitado dado que está sujeto al tamaño de inserto de Illumina y a la degradación intrínseca del ADN derivada del tratamiento con bisulfito, que sólo les permite abarcar un número limitado de CpGs adyacentes a la inserción, a lo sumo dos o tres. La reciente introducción de lecturas largas de Nanopore y la detección directa de modificaciones de nucleótidos a partir de las mismas ha permitido abordar mejor este problema. Apoyándose en esta tecnología, Lanciano (Lanciano et al. 2023) demostraron que la región genómica ubicada hasta 300 pares de bases *upstream* a una inserción somática de LINE1 adopta con frecuencia el estado de metilación de su promotor. Además, identifican factores de transcripción cuya asociación al promotor o a su flanco depende del perfil de metilación de LINE1, algunos de los cuales impulsan la expresión de transcritos quiméricos de LINE1 con genes vecinos, favoreciendo potencialmente al desarrollo del tumor (Lanciano et al. 2023). Un ejemplo se encuentra en la línea celular de carcinoma de colon HCT116, en la que la desmetilación de un LINE1 en el primer intrón del proto-oncogén cMET impulsa la transcripción en anti sentido (L1-ASP, del inglés *antisense promoter transcription*) de una isoforma truncada de su proteína, la cual se relaciona con prognosis negativa en pacientes con leucemia mieloide crónica (Cruickshanks y Tufarelli 2009; Weber et al. 2010).

En base a lo dicho, se podría suponer que las células tumorales no están sometidas a presión selectiva para compensar una pérdida de metilación en los promotores de LINE1, dado que sus inserciones somáticas en raras ocasiones comprometen la fitness del tumor e, inclusive, a veces pueden beneficiarse de ello. Por el contrario, y más allá del impacto genómico de sus inserciones, el incremento en transcritos aberrantes de LINE1 puede desencadenar una respuesta antiviral a través de interferón tipo I (IFN) y, consecuentemente, inducir la apoptosis de la célula tumoral (Guler et al. 2017; De Cecco et al. 2019). De hecho, multitud de tumores sobre expresan SETDB1 y LSD1, cuyos productos proteicos silencian epigenéticamente retrotransposones en lo que parece ser el resultado de una selección clonal en contra de la activación de ERVs y LINE1 (Lazaro-Camp et al. 2021; Lin et al. 2021). Irónicamente la activación los elementos transponibles podría estar en parte relacionado con la eficacia terapéutica de los tratamientos antitumorales basados en inhibidores de metilasas (DNMTi), sugiriendo inesperadas y emocionantes dianas terapéuticas fundamentadas en la reactivación de transposones (Johnstone et al. 2022). En definitiva, y pese a las evidentes y cada vez más crecientes implicaciones de la retrotransposición de LINE1 en cáncer y la importancia de los mecanismos epigenéticos en su regulación, nuestro conocimiento sobre los procesos que regulan la hipometilación y, por consiguiente, activación de LINE1 son todavía limitados. Desconocemos el grado de implicación de la retrotransposición somática en la transformación tumoral y cómo y cuándo se activa exactamente, ya que en última instancia sólo disponemos de una captura del proceso final de evolución cancerosa, no una sucesión de eventos que sean causa o consecuencia del advenimiento de un tumor. Abordar estas cuestiones permitiría explorar nuevas terapias antitumorales, tanto para reprimir las mutaciones conductoras derivadas de LINE1 como para reactivar LINE1 con el fin de inducir lesiones genómicas para destruir las células cancerosas.

3.5 TÉCNICAS Y HERRAMIENTAS PARA EL ESTUDIO DE LOS RETROTRANSPOSONES Y SUS PATRONES DE METILACIÓN EN CÁNCER

Como ya se ha indicado, los retrotransposones son elementos móviles dentro del genoma que pueden insertarse en diferentes lugares del mismo y afectar su estabilidad y funcionamiento, así como al de la célula que lo contiene. Por consiguiente, y ya al margen de los aspectos biológicos ya mencionados, tres son las características clave para el estudio genómico de los transposones: su secuencia, su ubicación en el genoma y su regulación. Dilucidar su secuencia y posición genómica requiere de la obtención y análisis de datos de genoma completo. Dado que a lo largo del presente manuscrito trataré en cierto detalle esta temática he creído conveniente realizar una brevísima historia de la secuenciación de los ácidos nucleicos, incluyendo las tecnologías de secuenciación de tercera generación, una de las cuales, la tecnología de los nanoporos se aplicó para estudiar las muestras usadas en este trabajo.

Un mecanismo clave en la regulación de la transcripción génica y de los LINE1 es la metilación de los dinucleótidos CpG, también estudiado en el presente trabajo. En el contexto del cáncer, la metilación de los retrotransposones puede ser un indicador importante de la progresión tumoral y un blanco potencial para el desarrollo de terapias. Por ello en el segundo apartado, describiremos con cierto detalle las técnicas, el proceso de adquisición y el análisis de la información de la metilación de los LINE1.

3.5.1 Tecnología de secuenciación de ácidos nucleicos

Al poco tiempo del descubrimiento de la estructura en doble hélice del ADN por Watson y Crick en 1953 basándose en imágenes de difracción de rayos X obtenidas por Rosalind Franklin (Watson y Crick 1953), varios grupos de investigación comenzaron el desarrollo de metodologías y protocolos para obtener la información contenida en su estructura, es decir, para determinar la secuencia de nucleótidos que componen el ADN o el ARN. Si bien la definición exacta de las generaciones de tecnologías de secuenciación de ácidos nucleicos es difusa, hoy en día se considera estas primeras técnicas como la primera generación.

3.5.1.1 Primera generación:

Los primeros intentos para secuenciar ácidos nucleicos tomaron prestada la tecnología de la química analítica, la cual permitía conocer la composición y proporción exacta de los nucleótidos, pero no desentrañar su secuencia. Debido a la facilidad de obtener grandes cantidades de material y a su reducida longitud, los primeros intentos se centraron en descifrar la secuencia de ácidos ribonucleicos (ARN). Así, en 1965, Robert Holley realizó la primera secuenciación de una cadena de ARN, más concretamente del ARN de transferencia para la Alanina de *Saccharomyces cerevisiae* haciendo uso de digestión parcial del ARN, cromatografía en columna y espectrometría de masas (Holley 1965; Heather y Chain 2016). En paralelo, Fred Sanger y sus colaboradores desarrollaron una técnica de análisis de fragmentación bidimensional del ARN con marcaje radiactivo (Sanger, Brownlee, y Barrell 1965), tecnología que les permitió secuenciar el ARN mensajero codificante para la proteína de la envoltura del fago MS2 (Jou et al. 1972). Fue

precisamente Fred Sanger quien desarrolló métodos más eficientes y precisos para la secuenciación de ADN, incluyendo la técnica de secuenciación *"más-menos"* que, junto a Alan Couson, le permitió secuenciar el genoma completo del fago φX174 (Sanger et al. 1977). Estas técnicas de 2D de electroforesis seguida de cromatografía, fueron pronto superadas por protocolos de migración del ADN marcado radiactivamente y fragmentado por nucleótidos específicos en geles de acrilamida, separando los fragmentos por su longitud lo que posibilitó la lectura de la secuencia de ADN (Maxam y Gilbert 1977). Daba comienzo la primera generación de secuenciación de los ácidos nucleicos.

El siguiente gran hito fue la técnica de terminación de cadena o didesoxi de Sanger en 1977. Esta consiste en usar la misma serie de reacciones que utilizan las células para copiar el ADN, es decir una reacción con una polimerasa de ADN. La genialidad consiste en el uso de una proporción de dideoxinucleótidos marcados radiac-tivamente (análogos de nucleótidos ddNT sin el grupo hidroxilo 3'). Estos dideoxinucleótidos hacen que la reacción de polimerización termine prematuramente, produciendo frag-mentos de ADN de todas las longitudes. Corriendo estos frag-mentos en cuatro carriles de un gel de poliacrilamida se puede leer la secuencia de ADN (Sanger, Nicklen, y Coulson 1977).

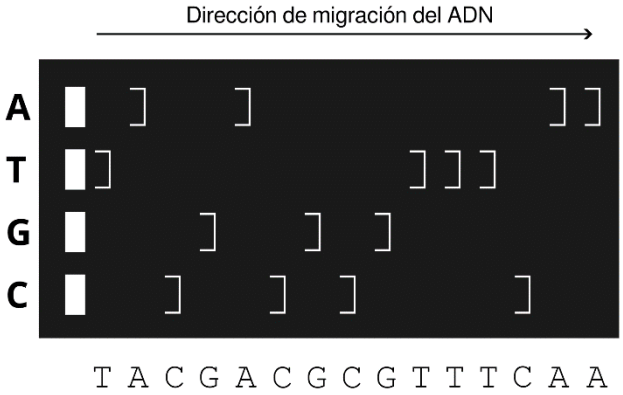


Figura 15: Determinación de una secuencia de ADN/ARN por el método de didesoxi de Sanger.

Tras el desarrollo de la técnica de Sanger, se realizaron avances que permitieron mejorar la eficiencia y la rapidez de la secuenciación del ADN, así como su automatización. Una de las mejoras más significativas fue la sustitución de los marcadores radiactivos por los fluorescentes, lo que permitió la detección automática de los fragmentos de ADN. Además, se avanzó en el diseño de los geles de electroforesis, sustituyéndolos por capilares, lo que posibilitó la aparición de máquinas de secuenciación automática de alta capacidad y precisión. Las máquinas de secuenciación automática son capaces de secuenciar grandes cantidades de ADN de cerca de 1k pb (1Kpb = kilo parece de bases = 1000 pares de base) en un corto periodo de tiempo y con una precisión muy alta. Sus versiones modernizadas siguen usándose en la actualidad, como *"Gold standard"*, debido a su elevadísima fiabilidad.

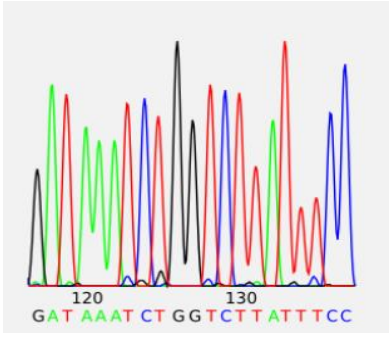


Figura 16: Ejemplo de la salida de un secuenciador capilar.

El siguiente desafío técnico en la secuenciación del ADN afrontó la secuenciación en paralelo de miles o cientos de miles de fragmentos simultáneamente o, en otras palabras, la secuenciación masiva de ácidos nucleicos. Dos fueron las principales estrategias empleadas para tal fin, el "*Hierarchical Shotgun Sequencing*" y el "*Whole Genome Shotgun Sequencing*". En la primera, se construye una librería de insertos que se

secuencia y ensambla para obtener la secuencia completa de un mismo fragmento de ADN. En el segundo, el ADN se fragmenta en infinidad de trozos, que luego se secuencian. La secuencia completa se obtenía (y se sigue obteniendo) ensamblando "*in silico*" las secuencias cuya secuencia se superpone, dando lugar largas secuencias contiguas (*contigs*). Los procesos fueron mejorados por la disponibilidad de grandes cantidades de ADN obtenido por reacciones en cadena de polimerasa (PCR), junto técnicas de ADN recombinante y nuevas polimerasas mejor adaptadas a las reacciones *in vitro*. Todas estas tecnologías y el uso de miles de máquinas de secuenciación de primera generación permitieron la secuenciación de los primeros genomas completos, lo que eventualmente llevó a afrontar el proyecto de secuenciación del genoma humano.

3.5.1.2 Segunda generación

Paralelamente al desarrollo de las técnicas de nucleótidos didesoxi, otros grupos de investigación desarrollaron técnicas como la pirosecuenciación (Ahmadian, Ehn, y Hober 2006). Esta técnica consiste en la medición de la luminiscencia producida acoplando dos procesos enzimáticos, la conversión de pirofosfato en ATP (por la ATP sulfurilasa) y el uso de este ATP como sustrato para la luciferasa. La intensidad de la luminiscencia es proporcional a la cantidad de pirofosfato.

Para inferir la secuencia de nucleótidos, en una cámara de reacción, una cadena de ADN fijada a una fase sólida se hace crecer en una reacción catalizada por una ADN polimerasa, mientras se observa la incorporación de los nucleótidos midiendo la luminiscencia y por lo tanto la producción de pirofosfato. Para identificar qué nucleótido se incorpora, se van añadiendo a la reacción los cuatro nucleótidos, uno por uno, y luego lavándolos.

A pesar de las diferencias con la técnica de secuenciación didesoxi de Sanger, la pirosecuenciación es también una secuencia por síntesis (SBS). Algunas de las ventajas son que utilizan nucleótidos naturales (en vez de dideoxinucleótidos) y que es una técnica en tiempo real. Mejoras como la PCR en emulsión de agua en aceite (emPCR), la degradación de los nucleótidos no incorporados para evitar lavados prolongados y el uso de placas de picolitros aumentó el número de reacciones que se podían realizar al mismo tiempo permitiendo una gran paralelización. Todo ello dio lugar al desarrollo de las máquinas de “secuenciación de próxima generación” (Next generation sequencing NGS).

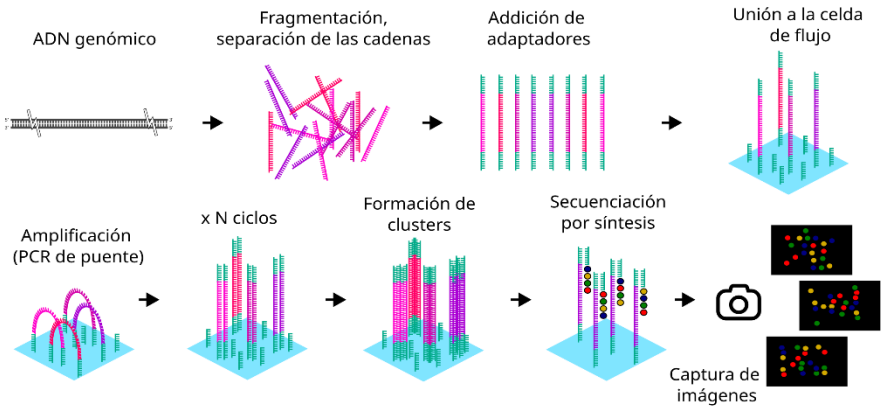


Figura 17: Secuenciación NGS.

Sin duda la tecnología que se ha impuesto es la tecnología de "amplificación por puente" (bridge amplification) de Solexa, posteriormente adquirida por Illumina.

Esta se basa en una amplificación por PCR en fase sólida de los fragmentos de ADN a secuenciar, una vez estos se unen a oligonucleótidos complementarios insertados en una celda de flujo. El último paso es la secuenciación usando una secuencia por síntesis (SBS) con terminadores reversibles fluorescentes dNTP. Estos nucleótidos especiales no se pueden unir a la cadena creciente de ADN hasta que sea eliminado el fluoróforo en la posición 3' del hidroxilo. El proceso consiste en pasos sincronizados de polimerización, lectura de los fluoróforos por una cámara CCD, y la escisión de los fluoróforos.

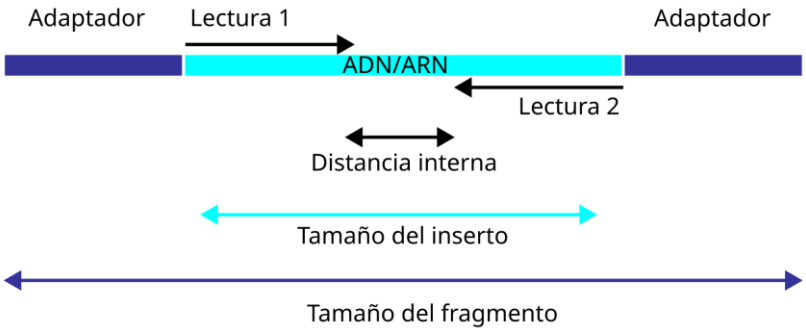


Figura 18: Términos relacionados con las lecturas NGS.

Solo pueden leer secuencias cortas de como mucho 150 pb, pero con la ventaja de que pueden producir datos '*Paired end'*, donde se secuencian unos cientos de pares de bases en los extremos de cadenas de ADN más largas, con insertos de entre 150 y 500 pares de bases.

Una tercera tecnología de esta generación fue el sistema de secuenciación por ligación y detección de oligonucleótidos (SOLiD) de Applied Biosystems. Donde la secuenciación se basa en una enzima de ligación (ADN ligasa) en lugar de una polimerasa. Aunque no puede producir la longitud de lectura y el rendimiento de otras tecnologías su coste por base la hacía competitiva en su momento.

Una última tecnología de esta generación digna de mención es la de la Ion Torrent, en la que un dispositivo semiconductor mide los protones (H^+) liberados en la reacción de polimerización.

3.5.1.3 Tercera generación

Siguiendo a James M. Heather y Benjamin Chain (Heather y Chain 2016), se considera que la secuenciación de una sola molécula (Single molecule sequencing: SMS), la secuenciación sin PCR y la divergencia de tecnologías anteriores son el sello distintivo de la secuenciación de ADN de tercera generación. Los primeros intentos de Helycos BioSciencies utilizaron un procedimiento similar al de tecnología de Illumina, pero sin el

puente de amplificación. Finalmente, las tecnologías que se consolidaron fueron la secuenciación en tiempo real de una sola molécula y la secuenciación de nanoporos.

Otras tecnologías usadas actualmente son una mezcla de esta generación con la generación anterior, como son el mapeo óptico BioNano y el sistema de lectura enlazada de 10X Genomics.

En la primera una endonucleasa crea una muesca de una sola hebra en las moléculas de ADN en un sitio de reconocimiento específico. Luego, los nucleótidos marcados con fluorescencia se incorporan y las mellas de una sola hebra se reparan con un polimerasa y ligasa. Estas cadenas etiquetadas se secuencian de manera estándar (NGS). La segunda consiste en la adición de un código de barras único a cada lectura corta de cada una de las moléculas originales individuales.

3.5.1.3.1 Secuenciación en tiempo real de una sola molécula (SMRT)

Esta tecnología fue desarrollada por Pacific Biosystems (PacBio) recientemente adquirida por la empresa Illumina y está basada en el principio de secuenciación por síntesis por polimerasas ancladas en unas microscópicas cámaras de reacción. La librería se prepara poniendo horquillas en los extremos de fragmentos de ADN, además de un cebador universal y una secuencia de inicio. Esto hace que una vez desnaturalizado forme una cadena circular. La secuenciación, tal como se comentó, se realiza en cámaras microscópicas especiales deno-minadas guía de ondas de modo cero (ZMW: zero-mode wave-guide). Estas cámaras tienen en su base anclada una polimerasa de ADN. La cadena circular de ADN se mueve a través de la polimerasa y cada vez que ésta incorpora uno de los cuatro desoxi nucleótidos trifosfato (dNTP) modificados (con un terminador reversible y marcados con un fluoróforo) se rompe un enlace de fosfato y se produce un pulso de luz, que se detecta en tiempo real y se asigna a un nucleótido determinado. Al ser la cadena circular, se pueden leer tanto la hebra positiva como la negativa varias veces, incrementando la fiabilidad de las lecturas (Pavlovic et al. 2020).

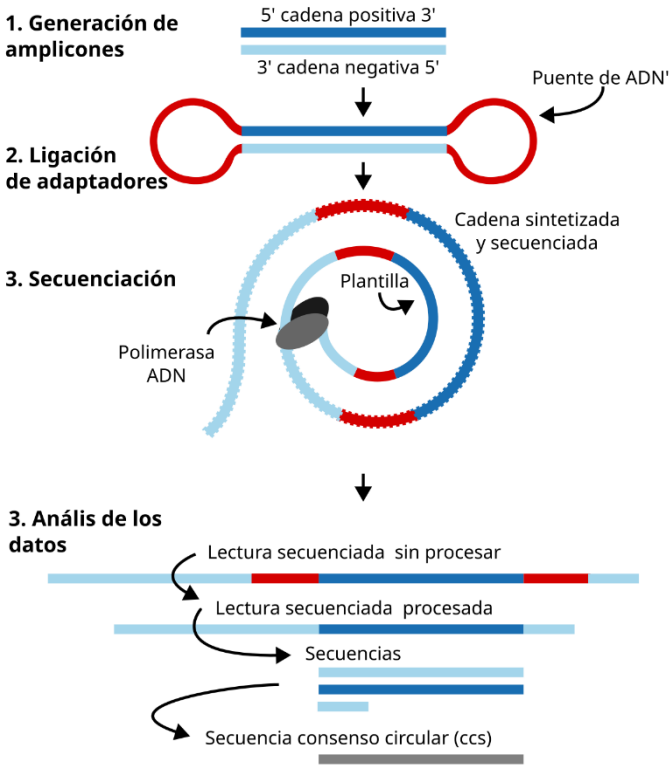


Figura 19: Preparación de la librería para la secuenciación en tiempo real de una sola molécula (SMRT).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Debido a que la cinética de incorporación de los desoxinucleótido es distinta si en la cadena original los nucleótidos están metilados o no, esto permite inferir su estado de metilación. Inicialmente presentaba el inconveniente de que se necesitaba una altísima cobertura (> 200x) (PacBio White paper *), por lo que sólo resultaba práctico con genomas pequeños (por ejemplo, bacterianos). Actualmente esta tecnología es capaz de producir la secuencia de un genoma humano con una altísima cobertura, incluyendo la información de metilación (Ni et al. 2022).

* Nota: https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf (Accedido 20.05.2023)

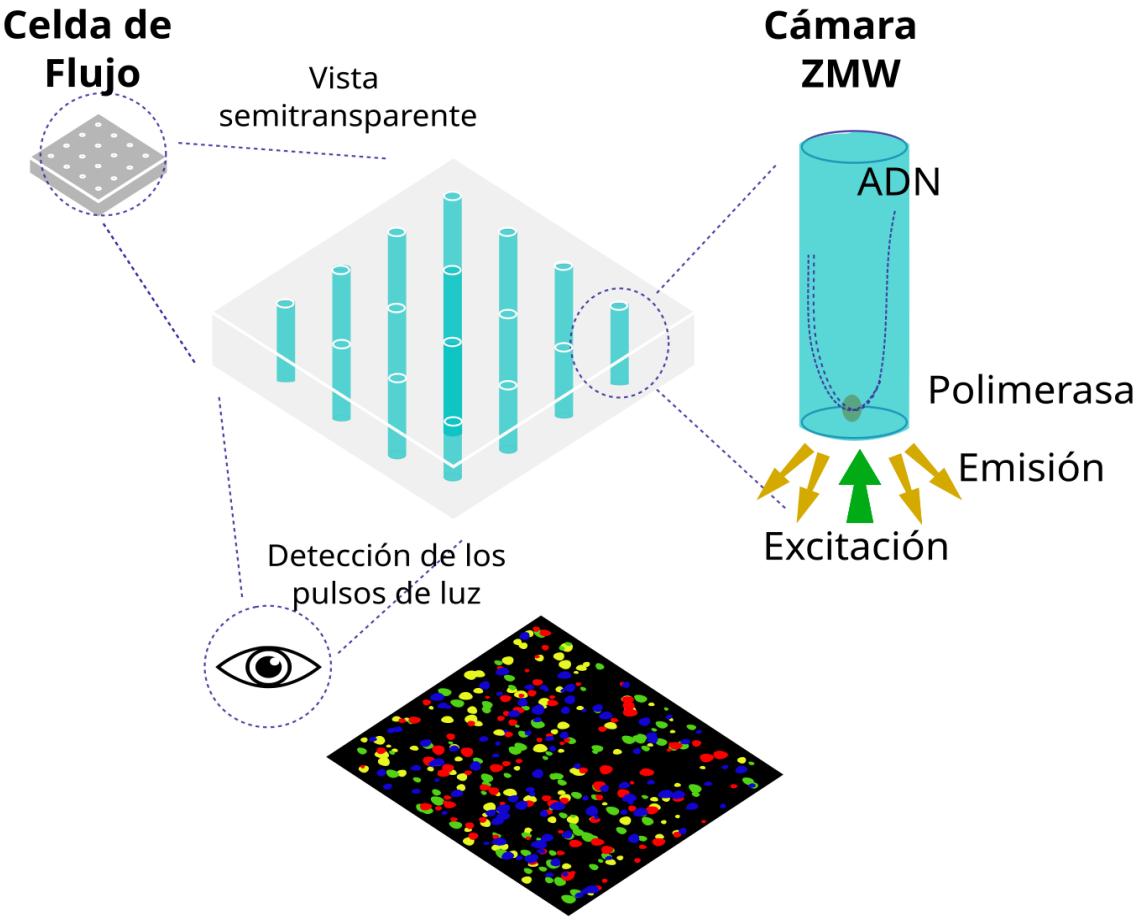


Figura 20: Esquema de secuenciación SMRT de PacBio. Secuenciación en tiempo real de una sola molécula (SMRT)

3.5.1.3.2 Secuenciación de nanoporos

La secuenciación de nanoporos de la empresa de Oxford Nanopore Technologies (ONT) consiste en la lectura directa de los nucleótidos de una cadena de ADN (o ARN) cuando ésta atraviesa unos nanoporos especialmente diseñados, por lo que es un tipo de secuenciación que no requiere síntesis. Esta es la tecnología empleada en el presente trabajo. Debido a su novedad y falta de madurez al principio de esta Tesis, hizo falta ponerla a punto y adaptarla a las necesidades de nuestro grupo de investigación y a las particularidades del estudio de los retrotransposones.

Este tipo de secuenciación permite producir lecturas de gran tamaño, y sin un tratamiento previo la identificación de modificadores epigenéticos (*e.g.* metilación) en los dinucleótidos CpG. Uno de los inconvenientes de esta tecnología es la relativa alta tasa de error por nucleótido comparada con tecnologías NGS, o la SMRT de PacBio, aunque la mejora está siendo constante y actualmente se está acercando mucho a ambas tecnologías (Weirather et al. 2017)[1]. A diferencia de otras tecnologías que se basan en la secuenciación por síntesis esta tecnología está basada en otros principios físicos, como son los canales iónicos y el contador Coulter (Feng et al. 2015).

En la década de 1990, varios grupos de investigación estudiaron la posibilidad de caracterizar polímeros a través de la interacción polímero-interfase (Patente N.º US5795782A US Church et al). De forma independiente Deamer y Akeson propusieron que era posible secuenciar el ADN usando sensores de nanoporos (Deamer, Akeson, y Branton 2016), pero fueron Kasianowicz et al. en 1996 quienes lograron por primera vez secuenciar una molécula de polinucleótido usando estos principios (Kasianowicz et al. 1996). La tecnología se basa en la medida de una corriente iónica a través de un canal nanométrico (nanoporo) cuando una molécula lo atraviesa.

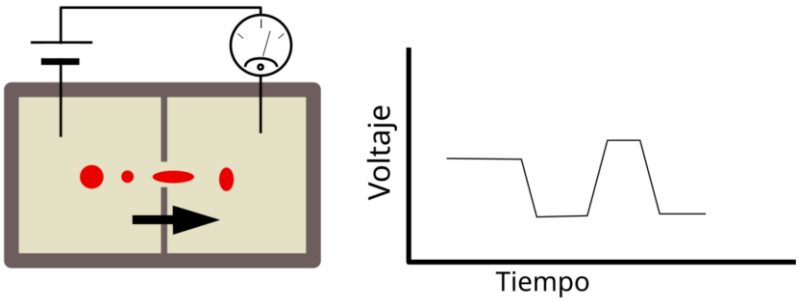


Figura 21: Contador Coulter.
Cuando diversas partículas atraviesan la separación entre dos cámaras la diferencia de potencial varia de forma mesurable.

Existen dos tipos de nanoporos:

a) Biológicos como la α-hemolisina, la porina de Mycobacterium smegmatis (MspA).

b) Nanoporos de estado sólido: Son más baratos y fáciles de producir usando técnicas de la industria semiconductores. Son superiores a los biológicos en estabilidad mecánica, química, térmica y en su integración. Además, pueden trabajar bajo una amplia variedad de condiciones experimentales.

En 2014, Oxford Nanopore Technologies (ONT) desarrolló el primer secuenciador comercial, el MinION. Este pequeño dispositivo (10 × 3 × 2 cm y 90 gramos de peso) se conecta a una computadora mediante un cable USB 3.0, utiliza celdas de flujo con 2048 nanoporos y 512 canales independientes que permiten un máximo de 512 lecturas simultáneas.

[1] Nota: A mayo de 2023, con la aparición de la química Q20+ han conseguido precisiones +99% equivalentes a los de otras tecnologías

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

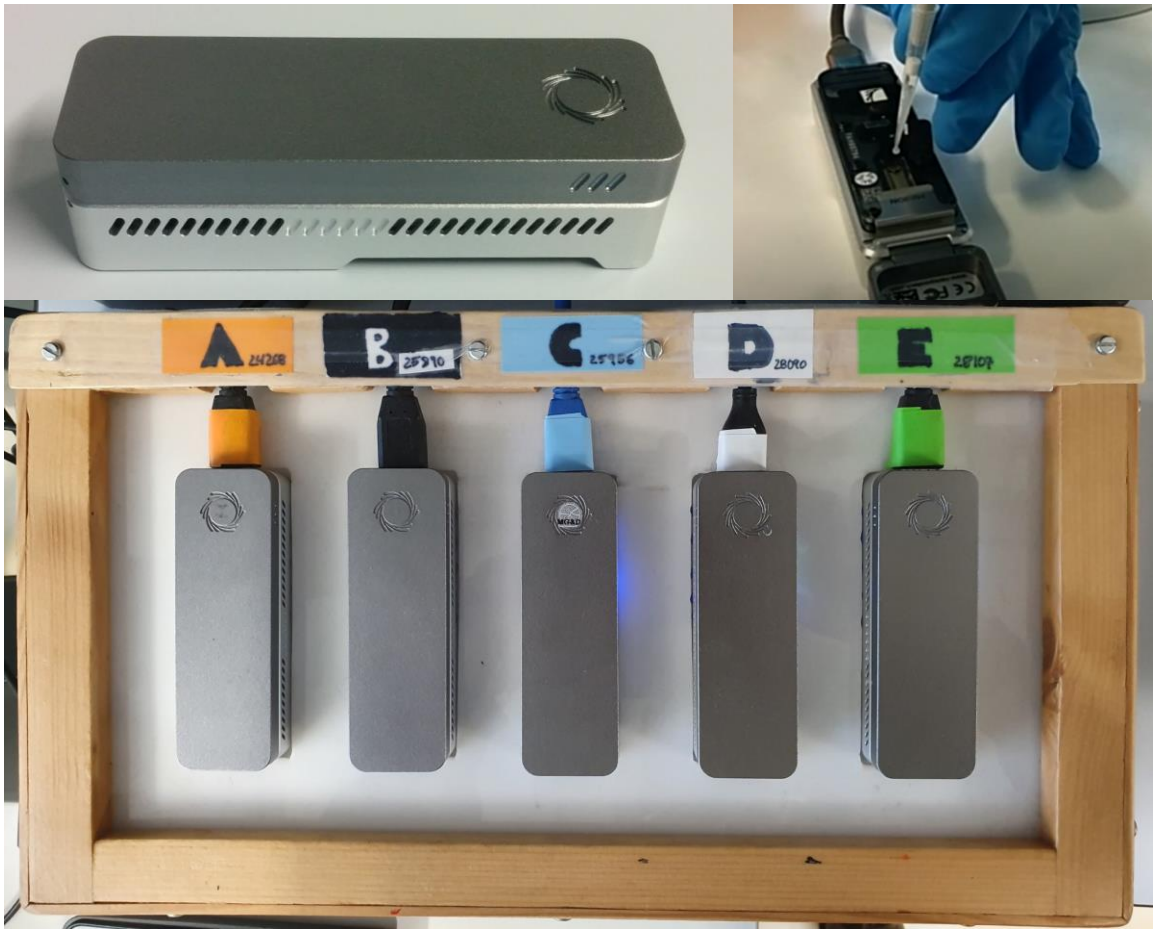


Figura 22: Diversas imágenes del dispositivo MinION.
Proceso de carga (derecha-arriba). Configuración instalada en nuestro laboratorio (abajo)

El uso del MinION es extremadamente fácil. Básicamente, se mezcla la muestra de ADN o ARN con los reactivos del kit que ONT proporciona, se coloca en la cámara de secuenciación del MINION, se conecta a un ordenador y se comienza la secuenciación en tiempo real con un software propietario. Es importante que los poros estén ocupados la mayor parte posible del tiempo para evitar la pérdida prematura de electrolitos. Por ello dependiendo del tamaño de los fragmentos de los ácidos nucleicos debemos asegurarnos de que el número de moléculas se encuentre en el rango de los 5-50 fmol.

Masa incial (ng)	Masa en la celda (ng)	Moles en la celda (fmol)
4800	2500	65
2400	850	27
1000	250	8
500	100	3
100	25	0.6
10	2	0.06

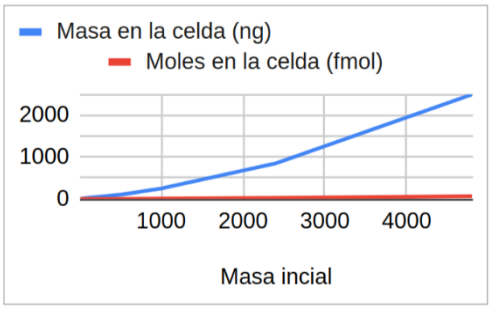


Figura 23: Relación N.º de moléculas de ADN masa de la librería.

Los reactivos contienen proteínas que hacen de motores moleculares. Una vez unidas a las moléculas de ADN y a los nanoporos de la cámara de secuenciación, despliegan las moléculas de ADN (helicasa) e insertan una de las cadenas en el nanoporo, haciendo que se mueva un nucleótido de cada vez (450 nucleótidos / seg). La conversión de la señal eléctrica en la secuencia de nucleótidos se denomina "*basecalling*" y se realiza en grupos de 5 o 6 nucleótidos (5-mers o 6-mers con 1024 ó 4096 combinaciones posibles) usando para ello un modelo de Márkov oculto (HMM), el algoritmo de Viterbi, o más recientemente una red neuronal convolucional.

Sus ventajas es que son relativamente económicos y compactos, produce lecturas excepcionalmente largas en tiempo real y permite la secuenciación en nuestro propio laboratorio con una inversión de capital inicial baja. Por contra, sus desventajas son la alta tasa de error en relación con otras tecnologías y su falta de madurez inicial, por lo que su constante evolución y mejora, hace difícil mantenerse actualizado. Algunas de sus aplicaciones más destacadas son la detección de variaciones estructurales, ensamblaje de secuencias del genoma obtenidas a partir de lecturas cortas, el ensamblaje de genomas completos y una fundamental para nuestro grupo como es la resolución de secuencias repetitivas, haplotipos y regiones ambiguas, características de los transposones en general. Cada celda de flujo de su dispositivo más modesto (MinION) puede generar de 10 a 15 Gb de datos de secuencia de ADN con longitudes de lectura ultra largas por un precio que ronda los 700 € (celda y reactivos). Recientemente se logró la lectura continua de una secuencia de ADN de más de un par de millones de bases.

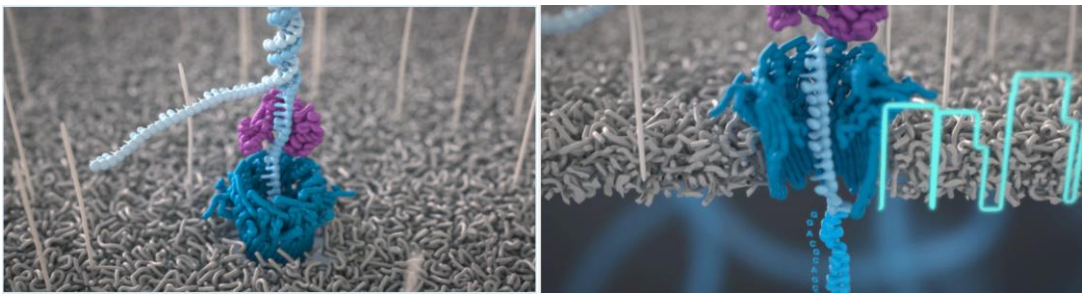


Figura 24: Secuenciación por nanoporos.
Detalle del ADN desenrollándose y entrando en el nanoporo (izquierda) y recreación de la variación del voltaje a medida que los nucleótidos atraviesan el poro (derecha). https://nanoporetech.com/media (accedida 2022.05.30)

Actualmente la empresa dispone de una amplia gama de dispositivos, desde secuenciadores de alto rendimiento, hasta dispositivos para el móvil.

Figura 25: Dispositivos de secuenciación de Oxford Nanopore Technologies.
(https://nanoporetech.com/media, accedida ene 2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.5.1.4 Aplicación de la secuenciación al estudio de los retrotransposones y cáncer

El desarrollo de las tecnologías de secuenciación masiva en paralelo de ácidos nucleicos, ya sean ADN o ARN, habitualmente referida por su nomenclatura inglesa, *next generation sequencing* (NGS) revolucionó nuestra comprensión sobre el genoma humano y las dinámicas mutacionales que lo perfilan, entre otras la retrotransposición. Frente a la secuenciación Sanger, las tecnologías de NGS abarcan una representación completa o parcial del genoma sin limitarnos a un número concreto de fragmentos amplificados mediante PCR. Este hecho, unido al abaratamiento del coste de secuenciación por nucleótido (Figura 26) y el desarrollo de nuevos programas y algoritmos informáticos, está permitiendo que nuestro conocimiento de los genomas, y en concreto del genoma humano se haya incrementado de forma exponencial, afectando a todas las áreas de conocimiento, no solo biológico, sino a muchos otros campos de la ciencia, incluyendo el de la salud humana, permitiendo que la medicina personalizada sea una realidad.

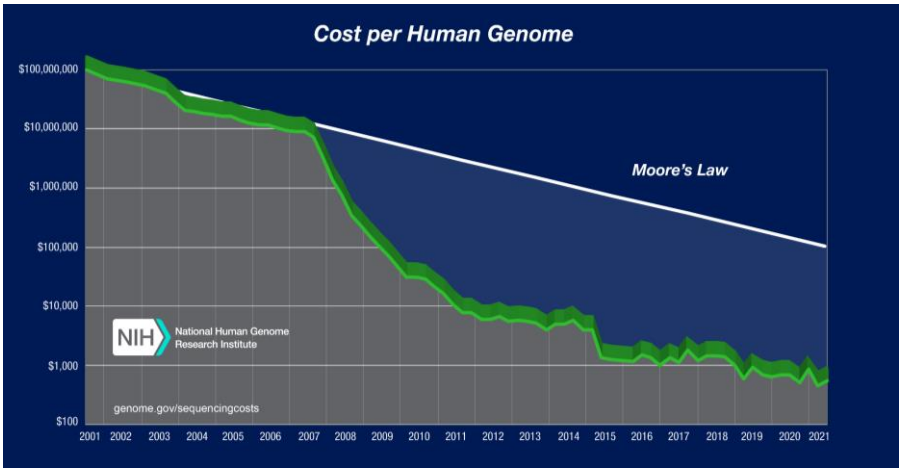


Figura 26: "The Cost of Sequencing a Human Genome". https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (accedido 09.07.2022)

En la actualidad, el "*Gold standard"* de referencia para la mayoría de los estudios genómicos es la tecnología de NGS de paired-ends desarrollada por Illumina (Illumina PE), de elevada fidelidad para la detección inequívoca de variantes mutacionales pequeñas, tales como SNVs o pequeñas Indels. Es por ello por lo que la inmensa mayoría de las herramientas informáticas de análisis genómico se han centrado en esta tecnología. En sí misma, esta tecnología fue suficiente para detectar y cuantificar el grado de retrotransposición somática y para deducir los mecanismos de inserción de LINE1 en el genoma del cáncer (E. Lee et al. 2012; Helman et al. 2014; Tubio et al. 2014; Rodriguez-Martin et al. 2020), mejorando notablemente nuestro entendimiento sobre esta materia a través de herramientas informáticas específicas para su detección tales como *Trafic* (Tubio et al. 2014) desarrollado en nuestro grupo de investigación.

Pese a ello, una de las grandes contrapartidas de las librerías de Illumina paired-ends es el tamaño de los fragmentos secuenciados, por lo general de 150 a 500 pares de bases de inserto, que resulta insuficiente para resolver completamente variantes estructurales en el genoma de las muestras en estudio (Rodriguez-Martin et al. 2020), entre otras, aquellas derivadas de eventos de retrotransposición somática. Las copias hijas de LINE1 insertadas *de novo* somáticamente en algún punto del genoma poseen

normalmente una longitud mayor (teóricamente de 6 Kb, si bien los truncamientos son habituales) que el tamaño de inserto de las librerías de Illumina PE. En consecuencia, no es posible determinar la naturaleza de la nueva copia de LINE1 más allá del rango promedio de la librería en torno a los puntos de inserción (o *breakpoints*). Por esto mismo, y salvo en los casos en los que se produzca una inserción canónica en la que se formen dos grupos (*clusters*) de lecturas pareadas en torno al *breakpoint*, tampoco es posible determinar inequívocamente translocaciones mediadas por LINE1 o deducir el elemento fuente del cual procede la nueva copia. Asimismo, dada su elevada repetitividad y homología entre las copias a lo largo del genoma, las lecturas con secuencias correspondientes a LINE1 mostrarán multitud de alineamientos suplementarios sobre la secuencia del genoma de referencia, lo que hace que un buen número sean no informativas.

Para paliar estos problemas es por lo cual recientemente se han desarrollado secuenciadores de molécula única (tercera generación) que permiten recuperar secuencias de fragmentos de ADN de mayor longitud. Estas secuencias están limitadas únicamente por la integridad del ADN inicialmente empleado para construir la librería. Los principales representantes en la actualidad de esta tecnología son las librerías single-molecule real-time sequencing (SMRT) de PacBio y las de Oxford Nanopore Technologies (mencionadas en el apartado anterior), cuyas lecturas, aunque por ahora un poco más imprecisas en la identificación de variaciones puntuales, resultan extremadamente útiles para detectar y reconstruir por completo eventos de reorganizaciones genómicas complejas y esclarecer haplotipos de amplio rango (Kraft y Kurth 2019). Todo ello supuso un punto de inflexión en el campo de la retrotransposición y una gran oportunidad para revisar conjeturas otrora inabarcables (Rodriguez-Martin et al. 2020).

En trabajos previos (Tubio et al. 2014; Rodriguez-Martin et al. 2020) y otros se demostró la importancia de los retrotransposición en tumores, siendo la principal variante estructural en adenocarcinomas de esófago, la segunda en tumores de cabeza y cuello, cánceres de colon y una fracción muy importante en cánceres de pulmón.

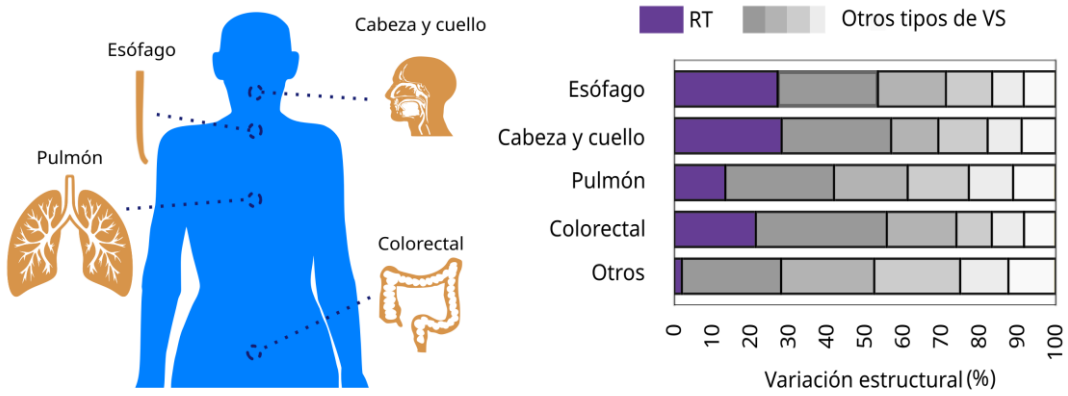


Figura 27: Cánceres con alta tasa de inserciones somáticas de retrotransposones. (Figura derecha adaptada de Rodriguez-Martin et al., 2020)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En el artículo (Tubio et al. 2014) además se estudió la movilización de los transposones y su relación con el estado de metilación de los promotores de estos, este artículo constituye el punto de partida del presente trabajo.

3.5.2 Técnicas y herramientas para el estudio de la metilación del ADN

Las técnicas y herramientas para el estudio de la metilación es un campo de estudio amplísimo, que recientemente ha evolucionado a gran velocidad. Hasta hace pocos años era frecuente estudiar la proporción de citosinas metiladas por medio de HPLC o espectroscopia de masas, el uso de microarrays (massARRAY), el análisis de cortes con enzimas de restricción sensibles a metilación (junto con hibridación southern blot, electroforesis en gel, electroforesis digital o PCR en tiempo real), métodos de captura por anticuerpos específico de secuencias metiladas o proteínas de unión a ADN metilado (Methylated DNA-binding protein or MBD), PCR específicas para metilación (MSP), MSP en tiempo real, MethyLight, análisis de derretimiento de alta resolución (high-resolution melting analysis), o métodos como el *"combined bisulfite restriction analysis (COBRA)"*, (Phokaew et al. 2008; Tollefsbol 2017; Hattori y Ushijima 2017; Rauch y Pfeifer 2017).

No obstante todos estos métodos o no tienen la resolución necesaria, o no abarcan todo el genoma, características necesarias para el desarrollo este trabajo, donde necesitamos conocer el estado de metilación de citosinas individuales a lo largo de todo el genoma, Para ello son necesarias metodologías más actuales como el pretratamiento del ADN con bisulfito y su posterior secuenciación con técnicas estándar (Illumina, Sanger, pirosecuenciación (Shigaki et al. 2013)), o las tecnologías de secuenciación de tercera generación.

3.5.2.1 Tratamiento con bisulfito y secuenciación

Esta técnica es hoy en día el "*Gold standard*" para el estudio de la metilación del ADN a nivel de citosinas individuales a lo largo de todo el genoma (Adusumalli et al. 2015). El sistema se basa en forzar la deaminación a uracilo de las citosinas no metiladas (posteriormente se convierten en timina en el ADN). Por el contrario, las citosinas metiladas están protegidas de la deaminación. Posteriormente el ADN tratado es secuenciado de forma estándar ya bien sea por secuenciación Sanger tradicional, por Illumina (HST: high-throughput sequencing), o incluso por pirosecuenciación (Shigaki et al. 2013). Este sistema no está exento de problemas. El tratamiento con bisulfito es bastante agresivo con el ADN, reduciendo de forma considerable la longitud de las cadenas de ADN. Por otro lado, el alineamiento de las secuencias producidas es mucho más difícil debido a la pérdida de complejidad en las secuencias al convertirse las citosinas no metiladas en timinas. La cobertura debe ser altísima para cubrir de forma homogénea todo el genoma, lo que implica miles de millones de lecturas.

La reducción del coste de esta técnica, experimentada desde hace unos años, posibilita la secuenciación completa de genomas completos de gran tamaño (WGBS whole-genome bisulfite sequencing), permitiendo proyectos como el atlas de metilación

de las células normales de los seres humanos (Loyfer et al. 2023). En caso de que el interés de un estudio sean regiones concretas, o para presupuestos bajos, se puede recurrir a la secuenciación por bisulfito con representación reducida RRBS-Seq "Reduced-representation bisulfite sequencing" en su acepción en inglés, o su versión para células únicas (single cell) scRRBS (Meissner et al. 2005; Guo et al. 2015). Estos protocolos usan una o múltiples enzimas, para cortar el ADN en secuencias específicas, que luego son tratadas con bisulfito y secuenciadas de forma similar al WGBS.

Nota: Un protocolo muy similar al anterior, en el que se realizan una serie adicional de reacciones permite la detección de 5 hidroximetilcitosina (5hmC) (Gao et al. 2013).

3.5.2.2 Detección directa de bases modificadas con secuenciadores de tercera generación

En el apartado (3.5.1.3) dedicado a las tecnologías de secuenciación, se describían las de tercera generación, cuyas principales características eran la capacidad de producir secuencias de gran longitud. Otra de las características de ambas tecnologías de secuenciación de molécula única, es que son capaces de inferir, de forma nativa, sin ningún tipo de tratamiento previo, el estado de metilación (y otras modificaciones) de las cadenas de ADN y ARN.

En el caso del sistema SMRT de PacBio, la detección de las modificaciones se basa en el análisis de las diferencias cinéticas en la incorporación de los nucleótidos. Las bases modificadas, por ejemplo, las metiladas, tienen una cinética ligeramente distinta que las citosinas no metiladas. Para distinguir ambas se requiere una secuenciación a muy alta cobertura, por lo que inicialmente solo era posible secuenciar genomas pequeños como los bacterianos. Mejoras recientes en el análisis de la cinética de incorporación, ha permitido extender su utilidad a genomas más grandes.

El sistema de nanoporos de ONT (Xu y Seki 2020) es capaz de detectar el estado de metilación de cada nucleótido de las cadenas de ADN o ARN analizando la señal eléctrica cuando una base modificada atraviesa el nanoporo, ya que el patrón de la señal eléctrica (*squiggle*) es distinto al patrón de la señal eléctrica del nucleótido sin modificar. El sistema utiliza un algoritmo basado en redes neuronales para el reconocimiento de los patrones. Aunque en principio el sistema es capaz de identificar cualquier tipo de modificación, los esfuerzos se han dirigido principalmente al entrenamiento de la red neuronal para la detección de la metilación e hidroximetilación del carbono 5 de las citosinas en el contexto de dinucleótidos de CpG (m5C, hm5C). El campo de estudio está evolucionando de forma vertiginosa, y ya existen redes neuronales experimentales capaces de detectar, metil 6 adeninas (m6A), y otro tipo de modificaciones y diferentes contextos. Esta es la tecnología usada en el presente trabajo, debido a que su puesta a punto requirió un esfuerzo considerable de investigación y adaptación se describe con más detalle en el apartado de resultados (6.1 Adaptación e implementación en nuestro laboratorio de las tecnologías de secuenciación de molécula única) y (6.2.2 Methyl-Meiga).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Estas nuevas tecnologías presentan la ventaja de que no hace falta tratar el ADN con bisulfito, por lo que el ADN es más íntegro, pudiéndose secuenciar fragmentos de gran longitud (de miles de decenas de pares de bases), ideal para el estudio de variantes

estructurales en general, y de los transposones en particular. Otra ventaja es que en cada secuenciación se obtienen los datos de secuenciación genómica y los datos de metilación, evitando tener que hacer secuenciaciones específicas para cada tipo de datos, con el consiguiente ahorro tanto del normalmente escaso y preciado material genético, como material de laboratorio y tiempo. Por otro lado, los inconvenientes son una mayor tasa de error y la falta de madurez de las herramientas utilizadas para su análisis.

3.5.2.3 Flujo de trabajo para el análisis de la metilación

Cualquier estudio bioinformático consiste en una serie de pasos, desde la señal producida por los detectores hasta el resultado final. El análisis de la metilación no es diferente, pero dependiendo de si se usa el método tradicional de bisulfito + secuenciación o la secuenciación de tercera generación, el flujo de trabajo cambia

3.5.2.3.1 Flujo de trabajo para pretratamiento con bisulfito + secuenciación Illumina

Los datos resultantes de este tipo de técnica consisten en una serie de ficheros con datos de secuenciación (normalmente ficheros FASTA). Estos datos tienen la particularidad de que las citosinas no metiladas se han convertido en una timina si se encuentran en la cadena positiva o en una adenina si se encuentran en la cadena negativa.

Por ejemplo, supongamos que tenemos esta secuencia que contiene una citosina metilada y otra sin metilar en ambas cadenas.

5’ --- A T G G C C^m --- 3’
↕ ↕ ↕ ↕ ↕ ↕
3’ --- T A C C^m G G --- 5’

Tras el tratamiento con bisulfito las citosinas no metiladas se deaminan por lo que se convierten en

5’ --- A T G G T C^m --- 3’ si la secuencia es del a cadena positiva

ó 3’ --- T A T C^m G G --- 5’ si es la cadena negativa,

(siendo su complementario ATAGCC que es lo que se refleja en la secuencia de datos).

Existen varios programas que alinean estas secuencias contra el genoma de referencia y así infieren si un determinado nucleótido está modificado o no. Algunos de los programas más utilizados para realizar estas tareas son BSMAP (Xi y Li 2009), BS-Seeker3 (Huang, Huang, y Chen 2018), Bismark (Krueger y Andrews 2011). En el presente trabajo la metodología de pretratamiento con bisulfito + Illumina se utilizó para los análisis preliminares y para corroborar la fiabilidad de la secuenciación con las nuevas tecnologías.

Una vez se conoce la secuencia con los nucleótidos modificados se puede proceder al análisis correspondiente. Los análisis pueden realizarse entre posiciones individuales, pero debido a las peculiaridades de este tipo de modificaciones, donde la metilación se suele producir por regiones, en la mayor parte de los análisis se suele estudiar *“regiones diferentemente metiladas”* (DMR) entre dos muestras, por ejemplo,

tejido sano frente a tumoral, de un tejido con un tratamiento determinado frente a otro sin él.

3.5.2.3.2 Flujo de trabajo para datos de secuenciadores de 3ª generación

Debido a su reciente implantación, los flujos de trabajo con estas tecnologías no están estandarizados, y siguen siendo altamente experimentales. A continuación, se describe a modo de ilustración el flujo de trabajo para los secuenciadores de ONT, el cual seleccionamos para este trabajo de Tesis, y por lo tanto conocemos con detalle.

La señal o los datos producidos por las máquinas de secuenciación son convertidos a una secuencia de nucleótidos (incluidos los modificados) por los denominados *basecallers*. Este tipo de programas se encarga de asignar nucleótidos a la información en "bruto" producida por las máquinas de secuenciación. La mayor parte de las empresas que producen máquinas de secuenciación poseen sus propios *basecallers* con código propietario. La empresa ONT tomó una aproximación más abierta creando e implicando a una comunidad de investigadores alrededor de sus máquinas de secuenciación de nanoporos, lo que propició la aparición de multitud de *basecallers* de código abierto para este tipo de dispositivos, aunque al final se impuso el proporcionado por la propia empresa, inicialmente el denominado Albacore, substituido por el Guppy, con varios modos de funcionamiento: 1) rápido, 2) de alta calidad (usado en nuestra cohorte), y 2) super alta calidad. Recientemente (febrero 2023) ha sido sustituido por el Dorado, mucho más rápido y de mayor precisión, aunque ya no ha dado tiempo a usarlo en el presente trabajo.

Algo parecido está sucediendo con los *basecallers* que devuelven información de las modificaciones de los nucleótidos, en concreto de su estado de metilación. Estos programas han sido analizados profusamente en varias publicaciones (Yuen et al. 2021; Q. Liu et al. 2019; Y. Liu et al. 2021) algunos de ellos son Nanopolish, Tombo, DeepSignal, DeepMod, deepMP, methBERT y los creados por la propia empresa ONT, Megalodon, usado en esta Tesis, y su versión renovada denominada Remora. Ambos programas, están basados en redes neuronales. La constante mejora de los modelos y los algoritmos subyacentes está propiciando que se estos programas se estén imponiendo al resto, aunque no es infrecuente el uso de varios de ellos.

Megalodon produce una serie de ficheros con la información de las bases modificadas, agrupadas por lectura y posición, (en ficheros de texto planos y también en un base de datos mySQL), agrupadas solo por posición, o directamente en ficheros *".bam"*. Inicialmente los ficheros contenían el resultado de la emulación de una secuenciación con bisulfito. Actualmente el programa produce ficheros *".bam"* con unas nuevas etiquetas MM y ML, específicas de la modificaciones de las bases con una información mucho más detallada. A partir de la información de los ficheros de texto planos, o de los ficheros *".bam"* se pueden comenzar los análisis de metilación, en algunos casos usando las mismas herramientas que para los datos obtenidos con bisulfito + secuenciación NGS, en otros casos usando herramientas específicas, como Methylartist, NanoMethViz, PycoMeth, CpGTools, Nanomethphas, Methplotlib, o usando nuestra propias herramientas (Methyl-Meiga y notebooks de Python).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.5.3 Herramientas para la búsqueda y análisis de retrotransposones

Las nuevas tecnologías de secuenciación NGS y de tercera generación han permitido describir y cuantificar con precisión los genomas y sus relaciones. No obstante, sigue siendo difícil su estudio en las extensas y numerosas regiones repetitivas de los genomas, compuestas en un altísimo porcentaje por transposones. En el caso concreto de los genomas humanos estos transposones son mayoritariamente retrotransposones como se ha comentado en apartados previos.

Para el estudio específico de los retrotransposones una de las mayores dificultades es poder asignar correctamente el locus de origen a partir de las lecturas, sobre todo si son cortas (Sexton y Han 2019) (conocido como el problema de la *"mapabilidad"* [2]) . Esto es debido a la altísima similitud entre los elementos móviles, mayor cuanto más jóvenes son estos pues no han tenido tiempo para acumular suficientes mutaciones. Estos elementos jóvenes son los más interesantes pues suelen ser los que retienen la capacidad para movilizarse. El paso previo para cualquier análisis genómico de los retrotransposones es alinear las secuencias facilitadas por los secuenciadores. Este paso consiste en averiguar la posición de cada una de las lecturas en relación a un genoma de referencia dado.

Existen infinidad de alineadores con sus ventajas e inconvenientes, en el presente trabajo se usaron el BWA (Li y Durbin 2010), para el alineamiento de las lecturas cortas de Illumina (short read paired-end), el Minimap2 (Li 2018) para las lecturas larga de Oxford Nanopore Technologies (ONT) y el Bismarq (Krueger y Andrews 2011) para el alineamiento de lecturas NGS + bisulfito.

Para los transposones anotados en el genoma de referencia (germinales), se puede proceder directamente a su estudio en las muestras de interés. No obstante, una porción importante de ellos es polimórficos en la población, o son inserciones somáticas de "*novo*", lo que requiere de programas específicos para su detección. Existen varios procedimientos y técnicas para ello. Unos consisten en la búsqueda por similitud de secuencias ya conocidas de retrotransposones: RepeatMasker, RepeatModeler, RepeatExplorer, RepeatHMM, RepeatClassifier, Transposome, RetroTector, TE-locator, los cuales por lo general usan las secuencias de transposones que se encuentran en bases de datos especializadas como "Rep base" (Bao, Kojima, y Kohany 2015), Dfam (Hubley et al. 2016), L1base (Penzkofer et al. 2017).

Otros métodos se basan en la búsqueda de alteraciones de los patrones y secuencias del genoma. Ejemplos de ellos son Trafic (desarrollado por nuestro grupo de investigación), xTea, TIDAL, Retroseq, y muchos otros. Esto programas se basan en alteraciones de la cobertura, el estudio de lecturas con "*clipping reads"* (segmentos de una lectura que no se corresponden con la región del genoma contra el cual se está alineando), el estudio de inserciones, deleciones, o reorganizaciones y en el caso de "*paired reads"* (Illumina) lecturas discordantes "*discordant reads*".

[2] Nota: La palabra mapabilidad (Sexton y Han 2019) no existe en la lengua castellana, y constituye un anglicismo. No obstante, su uso es tan frecuente en el campo que se ha decidido incluirlo en el presente trabajo. Tiene el significado de calidad de un alineamiento o de una lectura con respecto al genoma de referencia, algo equivalente a estar en el "mapa".

Una vez localizados los transposones de interés hace falta caracterizarlos y anotarlos, (si no lo están ya). Para ello se comparan con bases de datos existentes, normalmente usando toda la secuencia o nucleótidos diagnóstico, pudiéndose caracterizar el tipo, familia y subfamilia del transposón.

Como se comentaba anteriormente existe una gran cantidad de programas para analizar y descubrir transposones germinales *"de novo"*, pero quedan fuera del ámbito de este trabajo, puesto que estamos principalmente interesados en los retrotransposones LINE1 germinales e íntegros ya descritos, pues son los que retienen teóricamente la capacidad de movilizarse, y además, estamos sobre todo interesados en las inserciones somáticas de los mismos, ambos en el contexto del cáncer. Para la búsqueda específica de inserciones somáticas de retrotransposones existen también varios programas, algunos usan lecturas cortas como por ejemplo: jitterbug, transposec, TIGER, TEMP, TEFlon, tea, xTea (Chu et al. 2021) , Melt (Gardner et al. 2017), o Trafic (Tubio et al. 2014) desarrollado por nuestro grupo. Recientemente han aparecido programas que aprovechan las lecturas largas de los secuenciadores de tercera generación, como xTea-lr (Chu et al. 2021), TLDR(Ewing et al. 2020), o MEIGA-LR (pendiente de publicación) también desarrollado por nuestro grupo. Este último programa es el que se ha ejecutado en toda nuestra cohorte de muestras, para la búsqueda de inserciones somáticas de retrotransposones. Para el presente trabajo se seleccionaron los datos de las inserciones somáticas de LINE1.

3.5.4 Técnicas y herramientas para el estudio de la metilación de los LINE1

Trabajos de investigación previos hacen uso de tecnologías que engloban el estado general de todas las CpG de diversos LINE1 pues usan el índice de metilación (MI) (Marzese et al. 2012; Marzese y Hoon 2015), Southern blots hipersensibles para metilación (Moore 2002), o ensayos de extensión de cebadores dependientes de metilación (SIRPH methylation-dependent primer extension assays) basados en HPLC (Sharma et al. 2019).

Más recientemente el uso de técnicas como la secuenciación de la metilación por pirosecuenciación (Poulin et al. 2018), y otras tecnologías como las de los estudios de (A. S. Yang et al. 2004; England y Pettersson 2005; Ogino, Kawasaki, et al. 2008; Goel et al. 2010; Antelo et al. 2012) han permitido acceder a la información de las primeras CpG en la zona 5' donde se encuentra el promotor de los LINE1. En ellos se analiza un número muy reducido de dinucleótidos de CpG (4-20), además al utilizar amplicones con cebadores genéricos no es posible distinguir entre los distintos LINE1 del genoma. En otros ensayos como el de (Nüsgen et al. 2015) al usar cebadores específicos, sí que son capaces de estudiar algunos LINE1 individuales, pero con las mismas limitaciones del número de bases analizadas.

La secuenciación con NGS tras el tratamiento con bisulfito (Tubio et al. 2014) también permite el individualizar los LINE1 al poder alinear parte de la lectura o su lectura pareada correspondiente a localizaciones exactas del genoma. La aparición de tecnologías de secuenciación de tercera generación, con lecturas de gran longitud y datos de metilación de forma nativa, permiten superar parte de las limitaciones anteriormente comentadas. Al comienzo de este trabajo de investigación no había ninguna aplicación

específica que hiciese uso de esta información por lo que los estudios de metilación de LINE1 se basaban en soluciones *ad hoc*.

3.6 MOTIVACIÓN DE LA TESIS

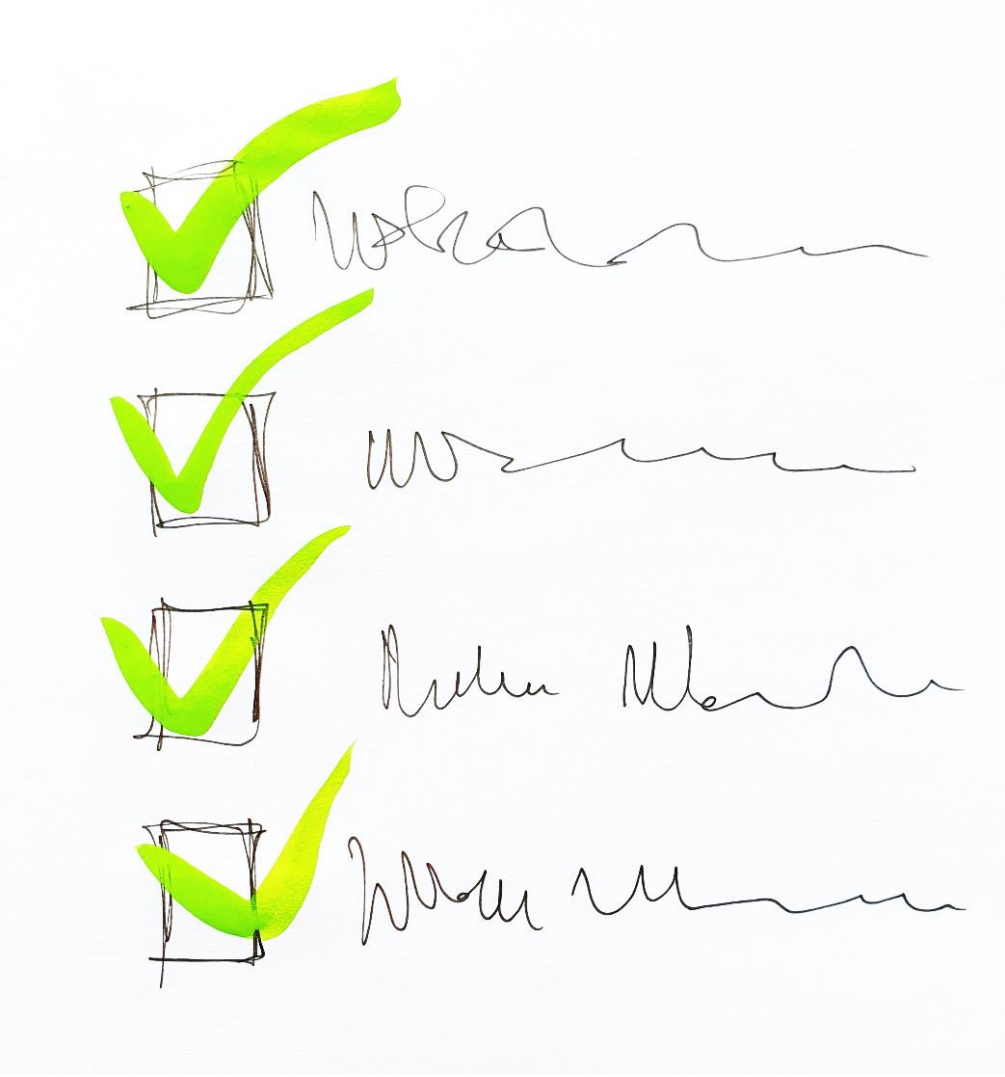
A pesar de lo anteriormente comentado, los patrones de metilación de los retrotransposones en tumores siguen estando mal caracterizadas, quedando pendientes cuestiones como:

- El estudio de los patrones de metilación de todo el promotor y el cuerpo del LINE1 de los LINE1 íntegros, no limitándose a estudiar solo las primeras posiciones como ocurre con la secuenciación NGS después del tratamiento con bisulfito
- ¿Influye el patrón de metilación del resto del retrotransposón en su capacidad de movilización?
- ¿Varían los patrones de metilación de los retrotransposones en los tumores con respecto al tejido sano?
- ¿Cómo afecta la inserción de retrotransposones al patrón de metilación de la región donde se inserta?
- Las copias hijas de los retrotransposones activos en el ser humano, ¿heredan el patrón de metilación de los elementos fuente de los cuales provienen cuando se movilizan en los tumores?

El objetivo del presente trabajo de Tesis es tratar de responder algunas de estas preguntas, para ello se puso a punto, en nuestro laboratorio, la tecnología de secuenciación de nanoporos de Oxford Nanopore Technologies y se desarrollaron los programas y algoritmos necesarios para para la extracción y análisis de la información de metilación e hidroximetilación a partir de los datos proporcionados por el mencionado secuenciador.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4. OBJETIVOS



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

El presente proyecto tiene los siguientes objetivos:

4.1 OBJETIVOS BIOLÓGICOS

4.1.1 Caracterizar los patrones de metilación de los LINE1 germinales (íntegros) y las regiones adyacentes en los genomas tumorales y su correspondiente tejido no tumoral.

4.1.2 Determinar la correlación entre el estado de metilación de los promotores de los LINE1 germinales (íntegros) y su actividad.

4.1.3 Caracterizar los patrones de metilación de las inserciones somáticas de los LINE1 y el tejido adyacente en el genoma tumoral y su correspondiente tejido no tumoral.

4.1.4 Determinar si las copias hijas de los LINE1 activos en el ser humano, heredan el patrón de metilación de los elementos fuente de los cuales provienen cuando se movilizan en los tumores.

4.2 OBJETIVOS METODOLÓGICOS

Debido a la dificultad de estudiar estos elementos altamente repetitivos con las tecnologías NGS actuales, se va a:

4.2.1 Adaptar e implementar en el laboratorio la tecnología de secuenciación de lecturas largas “single-molecule sequencing” MinION de Oxford Nanopore Technologies.

4.2.2 Desarrollar novedosos algoritmos informáticos para el estudio de patrones de metilación de elementos móviles.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5. METODOLOGÍA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Carga de una librería con ADN tumoral en un MinION” Acuarela y rotulador sobre papel (feb 2023)

5.1 PROTOCOLOS DE LABORATORIO

5.1.1 Electroforesis en gel de agarosa

Este tipo de electroforesis se empleó de forma rutinaria para la comprobación de los productos de PCR (a partir de este momento amplicones de ADN genómico), para comprobaciones de la integridad del ADN para las librerías de secuenciación y para los experimentos de obtención de ADN de cadena larga.

La concentración de agarosa (Fisher Scientific) fue de 0.8% en tampón TAE 1X, pre-teñidos con SYBR-Safe. Los geles se corrieron en tampón de TBE 1x.

La visualización del DNA se realizó bajo luz ultravioleta (UV) en un transiluminador LM-20 con el sistema de documentación de geles GelDoc (Bio Rad). Como marcadores de peso molecular se utilizó el "Estándar de ADN 1 Kb Plus" Invitrogen™ mostrada en la figura 28(facilitada por el fabricante).

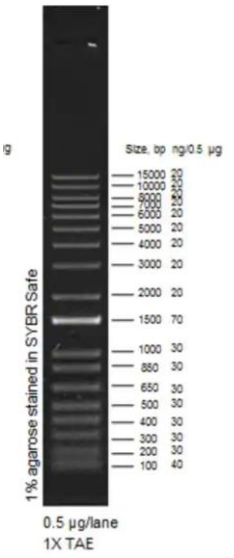


Figura 28: Escalera de pesos moleculares del ADN.

5.1.2 Reacción en cadena de polimerasa (PCR)

Esta técnica se empleó para amplificar regiones de DNA de interés para su posterior secuenciación. Las amplificaciones por PCR consistieron en un paso inicial de desnaturalización a 95 °C durante 7 min, seguido por 35 ciclos a 95 °C por 30 s, 60 °C por 30 s, y 72 °C por 45 s., una fase final de extensión a 72 °C durante 7 min y luego se mantuvo a 4 °C. Todas las PCR se realizaron en un volumen final de 25 µL que contiene 2 µL de ADN (10 ng/µL), 1 µL de cebadores directos e inversos a 10 µM (Sigma), 2,5 µL de dNTP (Invitrogen) a 2 mM, 0,5 µL de Taq polimerasa (Sigma) a 5 unidades/µL, 2,5 µL de tampón PCR 10X (Sigma) y 15,5 µL de agua (agua de grado PCR). Los productos de PCR fueron visualizados por electroforesis en gel de agarosa al 0.8% + TAE 1X y se migraron en un buffer de TBE 1X tal como se indica en el protocolo anterior.

5.1.3 Extracción de ADN a partir de tejido

Se usó el kit AllPrep(r) DNA/RNA mini kit de QIAGEN siguiendo las indicaciones del fabricante. Básicamente son las siguientes:

Trabajando en una campana de seguridad, se pone en hielo 20 mg de tejido ultracongelado (-80º C), cortando con escalpelo y fórceps. Se introduce en un eppendorf de 1,5 ml con buffer de lisis formado por RLT 1ml (Quiagen) + β-Mercaptoetanol 0.01 ml. Se saca del tubo y se pone en un portaobjetos, se corta, homogeniza y se vuelve al eppendorf. Se centrifuga el lisado 3 minutos a velocidad máxima. Con cuidado se retira el sobrenadante pipeteando y se transfiere a la columna de ADN. Se centrifugan las columnas 30 segundos a 11.000 rpm, asegurándose de que no queda líquido en la

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

membrana de la columna. Se guarda la columna de ADN en nuevos eppendorfs y se almacena a 4º C para su posterior purificación.

5.1.4 Extracción de ADN a partir de sangre y/o líneas celulares

Se usó el kit "QIAamp® DNA Mini and Blood Mini Handbook" de QIAGEN siguiendo las indicaciones del fabricante. Con la única particularidad de que el ADN se eluyó en agua, porque las librerías de lecturas largas de MinION así lo requieren. El tamaño del ADN suele ser de hasta 50 kb (con alguna molécula de hasta 200 kb), con predominio de fragmentos de aproximadamente de 20–30 kb.

5.1.5 Purificación del ADN

Se usa el mismo kit descrito en el apartado anterior. Se añaden 0,5 ml de buffer AW1 a la columna obtenida en el paso de extracción de ADN. Se centrifuga 15 minutos a 11.000 rpm para lavar la columna. Descartar. Se añaden 0,5 ml de buffer AW2 a la columna de ADN y se centrifuga 2 minutos a máxima velocidad para lavar y secar la columna. Se coloca la columna de en un nuevo eppendorf de 1,5 ml con 0,1 ml de buffer EB y se incuba a temperatura ambiente a 1 minuto. Se centrifuga otro minuto a 11.000 rpm para eluir el ADN. Se vuelve a poner el eluido en la membrana y se repite el proceso para maximizar la recuperación de ADN. Para eluir más ADN de la membrana, se pone la columna en otro eppendorf de 1,5 ml con otros 0,1 ml de buffer EB y se centrifuga 1 minuto 11.000 rpm. Se transfieren 3 µl a un tubo de 0,2 ml para el control de calidad (QC).

5.1.6 Limpieza del ADN

Puesto a punto por Daniel García-Souto 04/03/2019 V0.0

1) Mezclar el MgCl2 (3M; 1µl por cada 50µl de ADN) y el acetato amónico (o potásico) (10M; 2µl por cada 50µl de ADN) con el ADN. Ambas sales hacen que el ADN sea más estable. (Esterilizar en autoclave o filtrar con 0,2 µM el acetato de amonio y el MgCl2 en viales de 2 ml para evitar contaminaciones).

2) Agregue al menos el doble de la cantidad de etanol puro. Por lo general, cuanto más etanol, mejor. Es mejor si el etanol está frío (congelador -20).

3) Mezclar por inversión varias veces.

4) Girar para que todo el líquido regrese al fondo del eppendorf.

5) "Incubar" a -20ºC durante al menos media hora. Cuanto más tiempo, mejor también. Se puede hacer durante la noche para maximizar la recuperación.

6) Centrifugar a máxima velocidad durante cinco minutos (ejemplo 5424 13000rpm).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7) Eliminar CUIDADOSAMENTE el sobrenadante. Si la cantidad de ADN es suficiente, será visible en la parte inferior del eppendorf. De lo contrario, pipetee desde la pared opuesta a la rotación de la centrífuga.

8) Resuspender el ADN con 500 µl de etanol al 70%. RESUPENSIÓN SUAVE. La idea aquí es separar el ADN del fondo del tubo. Vierta el etanol dentro del vial, luego sumerja la punta, extraiga un poco de líquido y fuerce suavemente el ADN para que se separe del fondo. Este paso permite que el ADN se rehidrate un poco el ADN.

9) Centrifugar a máxima velocidad durante cinco minutos. (Microcentrífuga a 12500 rpm)

10) Eliminar CUIDADOSAMENTE el sobrenadante Si la cantidad de ADN es suficiente será visible en el fondo del eppendorf. De lo contrario, pipetee desde la pared opuesta a la rotación de la centrífuga. En este paso lo mejor es sacar la mayor parte del líquido para dejarlo secar sin etanol.

11) Los pasos 6 a 8 pueden repetirse tantas veces como desee para tratar de eliminar las sales. Aunque perderás algo de ADN cada vez.

12) Deje que el ADN se seque en la poyata a temperatura ambiente. El Speedvac debería funcionar igualmente. La idea es no dejar que se seque demasiado, de lo contrario se agrietará y no se resuspenderá bien.

13) Resuspender el ADN en una cantidad adecuada de agua sacudiendo el tubo.

14) Dejar resuspendiendo toda la noche a 4 grados.

5.1.7 Cuantificación y control de calidad del ADN

Antes de proceder a la secuenciación del ADN obtenido en las extracciones, se realizó el control de calidad y su cuantificación, puesto que para la secuenciación del ADN debe estar libre de proteínas y otros contaminantes. Su integridad debe ser óptima para obtener las moléculas más largas posibles.

5.1.7.1 Medición de la cantidad de ADN y contaminación por proteínas:

Se utilizó el espectrofotómetro One de la casa NanoDrop. Todas las comprobaciones se realizaron con 1µl de muestra, tras establecer el nivel base (“blanco”) con agua libre de nucleasas (ThermoFisher, cat # AM9937). Se comprobaron las absorbancias a 260 nm (DNA) y la 280nm (proteína) y a 230 nm (otros contaminantes como EDTA, guanidina, fenol), siendo la relación 260/280 entre 1.8-2.0 y la relación 260/230 mayor que 2. (ver figura 43 fago λ)

Figura 29: Espectrofotómetro One.

Ejemplo de concentraciones y contaminación de las líneas celulares usadas en la Tesis.

USC UNIVERSIDADE DE SANTIAGO DE

Muestra	Concentración supuesta (ng/ul)	Alícuota (ul)	NFW (ul)	NanoDrop [] (ng/ul)	260/280	260/230	Qbit [] muestra (ng/ul)	Qbit [] original (ng/ul)
H2009-B	547	4	4	35.7	1.84	1.73	25.6	51.2
H2087-A	138	2	6	85	1.87	2.35	74.2	296
BL 2009-B	26.2	8	0	43.3	1.88	2.56	32.8	32.8
BL 2087-A	53	4	4	28.2	1.87	1.99	24.0	48

Tabla 1: Contaminación de las líneas celulares (NanoDrop)

Debido a que este tipo de dispositivos no distingue entre ADN de cadena sencilla y cadena doble también se usaron otros métodos para cuantificar el ADN como un fluorímetro.

5.1.7.2 Medición de la cantidad de ADN de doble cadena

Para medir la cantidad y concentración del ADN de cadena doble, necesario para proceder a la secuenciación se usó el fluorímetro Qubit (Invitrogen) usando el protocolo de la marca. Brevemente, tras preparar la recta patrón con dos muestras de concentración conocida se procede a calibrar el aparato, A cada una de las muestras (1 ul) se le añade 1ul del fluoróforo, tras un minuto de reposo se procede a su medición, obteniendo directamente del aparato la concentración del ADN. Para la secuenciación de moléculas es necesario que el ADN se encuentre muy íntegro, información que no proporciona este dispositivo. Por lo que todas las muestras se analizaron mediante un sistema de electroforesis automatizado.

Figura 30: Fluorímetro Invitrogen Qubit.

5.1.7.3 Medición de la calidad y longitud del ADN.

Para comprobar la integridad del ADN de las muestras se utilizó una TapeStation 4200 (Agilent Technologies). Este dispositivo realiza una electroforesis de forma automatizada a partir de 1-2 µL de muestra. Es capaz de analizar hasta 15 muestras al mismo tiempo, requiriendo el uso de una escalera (A1 en la figura inferior) para marcar los pesos moleculares. Existen distintos kits para ARN, ADN.

En la figura 31, derecha se muestra la salida típica de varias muestras de ADN usadas en esta Tesis. Se puede observar que todas ellas tienen buena integridad y un alto peso molecular.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

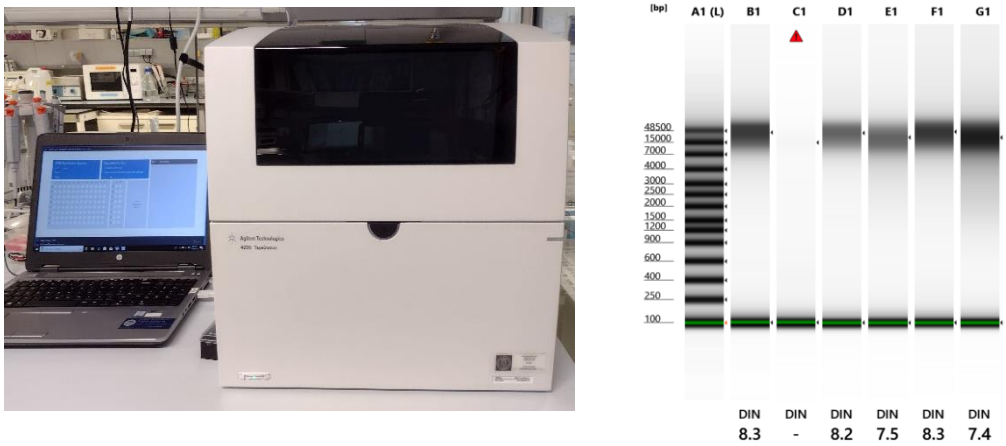


Figura 31: TapeStation 4200 (Agilent Technologies) y salida de la medición de la calidad del ADN de varias muestras.

5.1.8 Conservación del ADN

El ADN para la preparación de librerías para ONT debe estar resuspendido en agua ultrapura (agua milli-Q), por lo que no se usó ningún tipo de buffer para resuspender el ADN, como por ejemplo TE (10 mm de Tris-HCl (pH 8,0), 0,1 mm de EDTA). Cuando se tenía previsto usar el ADN en los días siguientes, se almacenó en la nevera a 4ºC. Si por el contrario se trataba de almacenarlo por largos períodos de tiempo se almacenaba en el congelador a -20º C.

5.1.9 Fragmentación del ADN

Tal como se comenta en el apartado (5.1.12 Secuenciación de nanoporos), las celdas de secuenciación contienen 512 poros, que deben estar ocupados durante la secuenciación para evitar la pérdida prematura de electrolitos. Por ello el número de moléculas en las librerías de ONT deben ser como mínimo 5-50 fmol tal como se indica en la figura 23. Por ello en ciertos casos cuando el ADN es escaso, es necesario fragmentarlo de una forma predecible, sin dañarlo en exceso.

Figura 32: Tubo de fragmentación Covaris.

Se hicieron pruebas haciendo pasar el ADN por jeringuillas con agujas muy finas, y mediante un sistema de agitación de alta intensidad, pero el resultado era la degradación del ADN, sin romperlo en fragmentos íntegros (ver apartado de resultados 6.1.1.5 Pruebas de fragmentación del ADN).

El método recomendado por la propia empresa ONT es la fragmentación por medio de tubos COVARIS g-TUBE, los cuales permiten obtener fragmentos entre 6 y 20 k bases, de una forma reproducible. El funcionamiento consiste en hacer pasar la muestra dos veces por un orificio a una velocidad controlada. Para realizar la fragmentación se sigue el protocolo proporcionado por la casa. Muy brevemente:

1- Se Coloca el g-TUBE en la posición de CARGA del soporte, atornille hacia arriba

USC UNIVERSIDADE DE SANTIAGO DE

2- Se Retira la tapa roscada del g-TUBE y se carga 150 µl de muestra en la parte superior del tubo

3- Se cierra el g-TUBE y se atornilla la tapa firmemente

4- Se carga el g-TUBE (tapón de rosca hacia arriba) en la centrífuga. Se usa un g-TUBE adicional para equilibrar el rotor

5- Se ajusta la centrífuga (Eppendorf® 5424) a la velocidad especificada en la tabla aportada por el fabricante, En nuestro caso para conseguir fragmentos de 10K pb 6.000 rpm, 60 segundos. (es lo que denominaremos corte "estándar")

6-. Se retira el g-TUBE de la centrífuga; se verifica visualmente para determinar si toda la muestra se ha drenado de la cámara superior del g-TUBE.

7- Se invierte el g-TUBE, colocándolo en el rotor (con la tapa de rosca hacia abajo) y se centrifuga nuevamente usando la misma velocidad y duración.

8- Se abre el g-TUBE y se extrae el ADN fragmentado.

5.1.10 Selección de ADN de cadena larga (Circulomics SRE)

Para la realización de librerías para ONT, es necesario el ADN lo más íntegro posible dado que la secuenciación es muy sensible a cualquier defecto en la cadena del ADN, pudiéndose producir la terminación prematura de la lectura. Debido a la naturaleza de los estudios realizados, era necesario que las secuencias fuesen lo más largas posible.

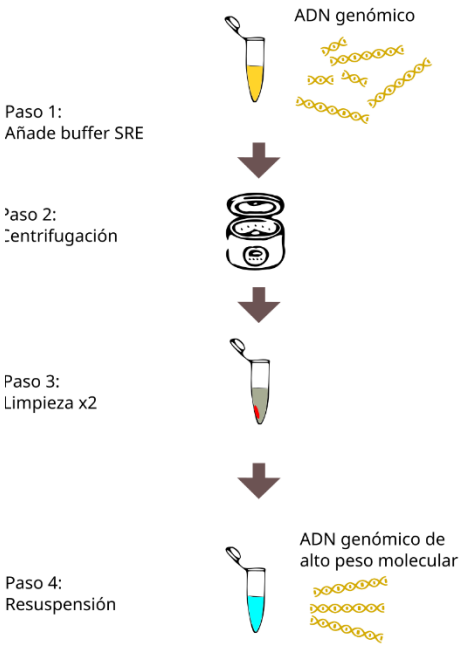


Figura 33: Esquema de selección de cadenas de ADN largas usando el protocolo de Circulomics SRE.
(obtenido del protocolo del fabricante)

Tal como se comentará en el apartado de resultados se hicieron varios experimentos que no resultaron óptimos para la selección de ADN de cadena larga. La aparición de una solución comercial vino a solucionar el problema. Esta solución es el kit" Short read eliminator (SRE)" de la casa Circulomics, recientemente adquirida por PAC BIO (Illumina). El kit produce la precipitación selectiva del ADN. Basado en nuestros propios experimentos, (datos no mostrados), suponemos que podría consistir en un buffer a base de polietilenglicol y sales para estabilizar el DNA (Acetato amónico o sódico, MgCl2) y, posiblemente, algún colorante para ver mejor el ADN al precipitarlo.

El protocolo es extremadamente sencillo.

- Ajustar la muestra de ADN a un volumen total de 60 µL y una concentración de ADN Qubit (HR) de entre 50 – 150 ng/µL. Pipetear la muestra en un tubo

Eppendorf DNA LoBind de 1,5 ml. (Diluir la muestra con tampón TE (pH 8), tampón EB o en nuestro caso agua, ya que el tampón puede interferir en el proceso de secuenciación)

- Agregar 60 µL de Buffer SRE o Buffer SRE XL a la muestra. Mezclar bien con suavidad, golpeando el tubo o pipeteando suavemente hacia arriba y hacia abajo.
- Cargar el tubo en la centrífuga con la bisagra mirando hacia el exterior del rotor.
- Centrifugar a 10 000 x g durante 30 min a temperatura ambiente (RT).
- Retirar con cuidado el sobrenadante del tubo sin perturbar el sedimento de ADN. Colocar la punta de la pipeta en el borde del lado de la bisagra del tubo. (El sedimento de ADN se habrá formado en el fondo del tubo debajo de la región de la bisagra)
- Agregar 200 µL de la solución de lavado de EtOH al 70 % al tubo y centrifugar a 10,000g durante 2 minutos a temperatura ambiente. (No golpear ni mezclar después de agregar EtOH al 70 %. Colocar el tubo directamente en la centrífuga)
- Retirar con cuidado la solución de lavado del tubo sin perturbar el sedimento de ADN. Colocar la punta de la pipeta en el borde del lado la bisagra del tubo
- Agregar 50 - 100 µL de Buffer EB al tubo e incubar a temperatura ambiente durante 20 minutos. El volumen del tampón se puede ajustar para lograr la concentración deseada.
- Después de la incubación, golpear suavemente el tubo para asegurarse de que el ADN se vuelva a suspender correctamente.
- Analizar la recuperación y pureza del ADN mediante NanoDrop y Qubit.

5.1.11 Secuenciación NGS

Figura 34: Secuenciador NGS miseq de Illumina

Aunque nuestro laboratorio dispone de acceso a un secuenciador miseq de Illumina, debido al volumen de las muestras secuenciadas se recurrió a un servicio externo, en concreto la empresa Magrogen (Seúl, Corea del Sur) donde se secuenciaron utilizando la plataforma Illumina NovaSeq6000, con fragmentos pareados (PE) de 150 pb y un tamaño de inserto de 500 pb. Las librerías de genoma completo se realizaron siguiendo el protocolo de extracción del ADN 5.1.3 Extracción de ADN a partir de tejido y 5.1.4 Extracción de ADN a partir de sangre. Para cada muestra se envió como mínimo 200 ng diluidos en 50 µl de agua ultrapura (milli-Q)

Dependiendo del propósito se obtuvieron diferentes cantidades de datos. Para un análisis preliminar de las muestras se secuenciaron a 15 GB. Para las muestras seleccionadas se secuenció a 90 GB por muestra.

5.1.12 Secuenciación de nanoporos mediante la tecnología de Oxford Nanopore Technologies.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Los principios teóricos y la descripción de esta técnica se describieron con detalle en 3.5.1.3.2 Secuenciación de nanoporos, A continuación, mostramos los protocolos y

procedimientos necesarios para secuenciar el ADN con esta tecnología. Partimos para ello de un ADN de gran integridad y alto peso molecular obtenido conforme a los protocolos anteriormente descritos

5.1.12.1 Preparación de la librería de ADN para secuenciación de nanoporos ONT (protocolos SQK-LSK108 y 109)

La secuenciación de ácidos nucleicos mediante nanoporos, es un campo en constante y rápida evolución, y los protocolos para la secuenciación no son menos. Durante el desarrollo de esa Tesis los protocolos de secuenciación usados han sido SQK-LSK108, SQK-LSK109, (ONT Ltd) para el dispositivo MinION, que utilizando celdas de secuenciación FLO-MIN106D, R9.4 ONT Ltd y con los kits de reactivos “9 chemistry”. Tanto los protocolos, y reactivos son a día de hoy considerados obsoletos por la propia empresa, que ahora propone el protocolo SQK-LSK114, el kit de reactivos “14 chemistry”. Además, está haciendo énfasis en los secuenciadores de alto rendimiento PromethION, que usan unos protocolos, celdas de secuenciación y kits de reactivos ligeramente distintos. (información obtenida de la web de la empresa **https://nanoporetech.com**/. Accedida 16. enero.2023)

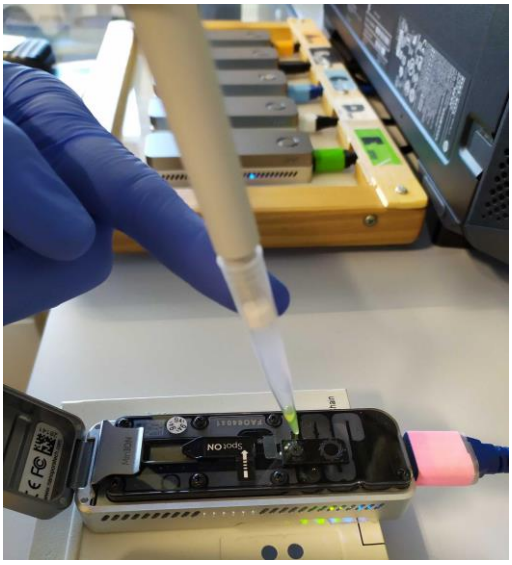


Figura 35: Carga de una librería en un MinION de ONT

Nota: Todos los protocolos se encuentran disponibles en la web del fabricante, previo registro gratuito.

Conceptualmente la preparación de la librería es muy sencilla. Partiendo del ADN que nos interesa (ya sea por fragmentación para obtener más moléculas, o por selección de tamaños para obtener fragmentos de la mayor longitud posible) el primer paso es reparar las posibles mellas que pueda contener, para evitar que la hebra de ADN se rompa una vez desenrollado. En este mismo paso se reparan los extremos del ADN, para que en el siguiente paso se puedan ligar los adaptadores, que contienen la helicasa que desarrollará el ADN e irá introduciendo una de las hebras en el nanoporo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

También se le une el "tether", (No mostradas en la figura) que son unas bolitas microscópicas que hacen que el ADN vaya al fondo de la cámara de secuenciación, lo que facilita que el ADN adaptado encuentre un poro.

Debido a la importancia de este protocolo para la puesta en funcionamiento y adap-tación de la secuenciación de nanoporos en nuestro labora-torio, y por formar parte importante del desarrollo por mí realizado se reproduce a continuación íntegro y traducido.

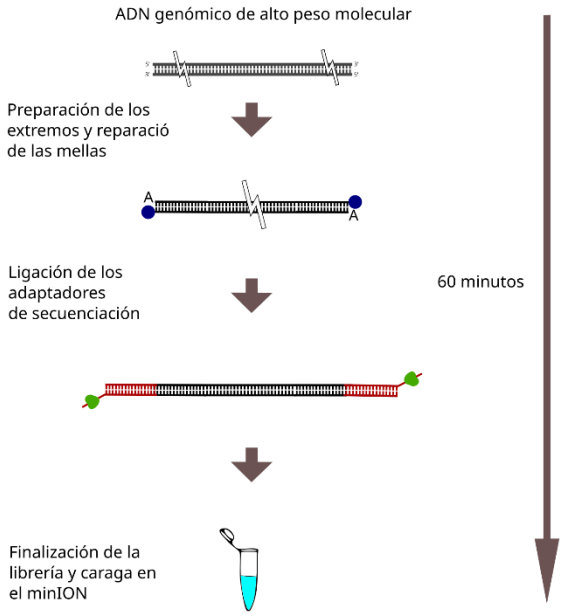


Figura 36: Esquema de una librería de secuenciación SQK-SLK109 ONT

Traducción basada en el protocolo: GDE_9063_V109_REVE_23MAY2018 (ONT)

- Materiales
  - 1 µg (o 100-200 fmol) gDNA o 100+ ng genómico de alto peso molecular si se realiza fragmentación de ADN
  - Kit de secuenciación de ligación (SQK-LSK109)
  - Kit de cebado de celdas de flujo (EXP-FLP002)
- Consumibles
  - Agencourt AMPure XP bolitas magnéticas (Beckman Coulter™, N.º de cat. A63881)
  - NEBNext FFPE Repair Mix (M6630) Microcentrífuga
  - NEBNext Ultra II Módulo de reparación de extremos/adición de dA (E7546)
  - Módulo de ligación rápida NEBNext (E6056)
  - Tubos Eppendorf DNA LoBind de 1,5 ml
  - Tubos de PCR de pared delgada de 0,2 ml Agua libre de nucleasas (p. ej., ThermoFisher, n.º de cat. AM9937)
  - Etanol al 70 % recién preparado en agua libre de nucleasas
- Equipos
  - Mezclador Hula (mezclador giratorio suave) o el equivalente diseñado y construido por mí (ver apartado 6.1.2).
  - Separador magnético, apto para 1,5 ml.
  - Mezclador Vortex
  - Termociclador
  - Temporizador
  - Pipetas y puntas de pipeta P2, P10, P20, P100, P200, P1000
  - Recipiente con hielo

Reparación de ADN y preparación de bordes romos.

- Descongelar la muestra de control de ADN (DCS) a temperatura ambiente, centrifugar, mezclar pipeteando y colocar en hielo.
- Preparar los reactivos NEBNext FFPE DNA Repair Mix y NEBNext Ultra II End Repair / dA-tailing Module de acuerdo con las instrucciones del fabricante y colocar en hielo.
- Descongelar todos los reactivos en hielo.
- Mover y/o invertir los tubos de reactivos para asegurarse de que estén bien mezclados.

Nota: No agitar en vórtex la mezcla de reparación de ADN FFPE ni la mezcla de enzimas de preparación de extremos Ultra II.

USC UNIVERSIDADE DE SANTIAGO DE

- La primera vez, cada día que se realiza este protocolo, girar hacia abajo los tubos antes de abrirlos
- El tampón de preparación de extremos Ultra II y el tampón de reparación de ADN FFPE pueden tener un poco de precipitado. Permitir que la mezcla llegue a temperatura ambiente y pipetear el tampón arriba y abajo varias veces para romper el precipitado, seguido de agitación del tubo durante 30 segundos para solubilizar cualquier precipitado.

  Nota: Es importante que los tampones se mezclen bien mediante agitación vorticial.

- El tampón de reparación de ADN FFPE puede tener una coloración amarillo, esto es correcto y se puede usar.
- Preparar el ADN en agua libre de nucleasas:
- Para celdas de flujo R9.4.1, transfiera 1 µg (o 100-200 fmol) de ADN genómico a un Eppendorf DNA LoBind de 1,5 ml
- Ajustar el volumen a 47 µl con agua libre de nucleasa
- Mezcle bien moviendo el tubo
- Girar brevemente en una microcentrífuga
- En un tubo PCR de pared delgada de 0,2 ml, mezclar lo siguiente:
    - 47 µl de ADN del paso anterior
    - 1 µl de CS de ADN (ADN de control fragmento terminal del ADN del fago Λ)
    - 3,5 µl de tampón de reparación de ADN NEBNext FFPE
    - 2 µl de mezcla de reparación de ADN NEBNext FFPE
    - Tampón de reacción de preparación final Ultra II de 3,5 µl
    - 3 µl de mezcla enzimática de preparación final Ultra II
- Asegurarse de que los componentes se mezclen completamente pipeteando y centrifugando.
- Incubar en el termociclador, a 20 °C durante 5 minutos y a 65 °C durante 5 minutos.
- [**IMPORTANTE]** Limpieza con bolitas magnéticas AMPure XP
- Volver a suspender las bolitas magnéticas AMPure XP mediante agitación vorticial.
- Transferir la muestra de ADN a un tubo limpio Eppendorf DNA LoBind de 1,5 ml.
- Agregar 60 µl de microesferas AMPure XP resuspendidas a la reacción de preparación final y mezclar moviendo el tubo.
- Incubar en un mezclador Hula (mezclador rotatorio) durante 5 minutos a temperatura ambiente.
- Preparar 500 l de etanol fresco al 70% en agua libre de nucleasas.
- Hacer descender la muestra al fondo del eppendorf y sedimentar en un imán hasta que el sobrenadante sea transparente e incoloro. Mantenga el tubo en el imán y pipetear el sobrenadante.
- Mantener el tubo en el imán y lave las bolitas magnéticas con 200 µl de etanol al 70% recién preparado sin perturbar las bolitas. Retirar el etanol con una pipeta y desechar.
- Repetir el paso anterior.
- Hacer descender la muestra al fondo del eppendorf y volver a colocar el tubo en el imán. Pipetear cualquier etanol residual. Deje secar durante ~30 segundos, pero que no se seque el pellet hasta el punto de agrietarse.
- Retirar el tubo de la gradilla magnética y volver a suspender el sedimento en 61 µl de agua libre de nucleasas.
- Incubar durante 2 minutos a temperatura ambiente.
- Precipitar las bolitas magnéticas en un imán, para formar un pellet hasta que el eluido sea transparente e incoloro, (al menos 1 minuto).
- Retirar y conservar 61 µl del eluido en un tubo Eppendorf DNA LoBind limpio de 1,5 ml.
- Cuantificar 1µ l de muestra eluida utilizando un fluorímetro Qubit.

  Nota: este punto es también es posible almacenar la muestra a 4°C durante la noche.

  Ligación y limpieza de los adaptadores

- **[IMPORTANTE]**Aunque la ligasa de terceros recomendada se suministra con su propio tampón, la eficiencia de ligación del Adapter Mix (AMX) es mayor cuando se utiliza el tampón de ligación suministrado con el kit de ligación.
- Centrifugar el Adapter Mix (AMX) y la Quick T4 Ligase, y colocarlos en hielo.
- Descongelar el Ligation Buffer (LNB) a temperatura ambiente, centrifugar y mezclar pipeteando. Debido a la viscosidad, agitar este tampón en un vórtex es ineficaz. Colocar en hielo inmediatamente después de descongelar y mezclar.
- Descongelar el tampón de elución (EB) a temperatura ambiente, mezclar con vórtex, centrifugar y colocar en hielo.
- **[IMPORTANTE]** Según el tampón de lavado (LFB o SFB) utilizado, el paso de limpieza después de la ligación del adaptador está diseñado para enriquecer fragmentos de ADN de >3 kb, o purificar todos los fragmentos por igual.
- Para enriquecer con fragmentos de ADN de 3 kb o más, use Long Fragment Buffer (LFB)
- Para retener fragmentos de ADN de todos los tamaños, utilice el tampón de fragmentos cortos (SFB)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- Para enriquecer los fragmentos de ADN de 3 kb o más, descongele un tubo de Long Fragment Buffer (LFB) a temperatura ambiente, mezclar por agitación, girar hacia abajo y colocar en hielo.
- Para retener fragmentos de ADN de todos los tamaños, descongele un tubo de tampón de fragmentos cortos (SFB) a temperatura ambiente, mezclar por agitación, girar hacia abajo y colocar en hielo.
- En un tubo Eppendorf DNA LoBind de 1,5 ml, mezclar en el siguiente orden:
    - 60 µl de muestra de ADN del paso anterior
    - 25 µl de tampón de ligación (LNB)
    - 10 µl NEBNext Quick T4 ADN ligasa
    - Mezcla de adaptadores de 5 µl (AMX)
- Asegurarse de que los componentes se mezclen completamente pipeteando y centrifugando.
- Incubar la reacción durante 10 minutos a temperatura ambiente.
- **[IMPORTANTE]** Si ha omitido el paso de purificación AMPure después de la reparación del ADN y la preparación final, no incube la reacción durante más de 10 minutos.
- Volver a suspender las bolitas magnéticas AMPure XP mediante agitación vorticial.
- Agregar 40 µl de bolitas magnéticas AMPure XP resuspendidas a la reacción y mezclar sacudiendo el tubo.
- Incubar en un mezclador Hula (mezclador rotatorio) durante 5 minutos a temperatura ambiente.
- Hacer descender la muestra al fondo del eppendorf y volver a colocar el tubo en el imán, para que se forme un pellet. Mantenga el tubo en el imán y pipetear el sobrenadante.
- Lave las bolitas magnéticas agregando 250 µl de tampón de fragmentos largos (LFB) o 250 µl de tampón de fragmentos cortos (SFB). Mover las bolitas magnéticas para volver a suspenderlas, y hacerlas descender en el recipiente, luego devuelva el tubo a la gradilla magnética y deje que las bolitas magnéticas sedimenten formando un pellet. Retirar el sobrenadante con una pipeta y desechar.
- Repita el paso anterior.
- Hacer descender la muestra al fondo del eppendorf y volver a colocar el tubo en el imán, para que se forme un pellet. Pipetear cualquier sobrenadante residual. Deje secar ~30 segundos, pero sin permitir que seque el pellet hasta el punto de agrietarse.
- Retirar el tubo de la gradilla magnética y volver a suspender el sedimento en 15 µl de tampón de elución (EB).
- Centrifugar e incubar durante 10 minutos a temperatura ambiente. Para ADN de alto peso molecular, la incubación a 37°C puede mejorar la recuperación de fragmentos largos.
- Sedimentar las bolitas magnéticas en un imán hasta formar un pellet y que el eluido sea transparente e incoloro, (al menos 1 minuto).
- Retirar y conservar 15 µl de eluido que contiene la librería del ADN en un Eppendorf DNA LoBind limpio de 1,5 ml.
- Cuantifique 1 µl de muestra eludida utilizando un fluorímetro Qubit.
- **[IMPORTANTE]** Recomendamos cargar 5-50 fmol de la librería final preparada en una celda de flujo.
- La librería preparada se utiliza para cargarla en la celda de flujo. Guarde la librería en hielo hasta que esté lista para cargar.

Cebado y carga de la celda de flujo SpotON

- **[IMPORTANTE]** Tener en cuenta que el tubo Sequencing Tether (SQT) NO se utilizará en este protocolo.
- Descongelar el tampón de secuenciación (SQB), las bolitas magnéticas de carga (LB), el Flush Tether (FLT) y un tubo de Flush Buffer (FB) a temperatura ambiente antes de mezclar los reactivos mediante agitación (vórtex) y centrifugar a temperatura ambiente.
- Para preparar la mezcla de cebado de la celda de flujo, agregar 30 µl de Flush Tether (FLT) descongelado y mezclado directamente al tubo de Flush Buffer (FB) descongelado y mezclado, y mezcle con vórtex a temperatura ambiente.
- Abrir la tapa del dispositivo MinION y deslizar la celda de flujo debajo del clip.
- Deslizar la tapa del compartimento de cebado de la celda de flujo en el sentido de las agujas del reloj para abrir el compartimento de cebado.
- **[IMPORTANTE]** Tenga cuidado al retirar el tampón de la celda de flujo. No extraiga más de 20-30 µl y asegúrese de que la matriz de poros esté cubierta por tampón en todo momento. Introducir burbujas de aire en la matriz puede dañar irreversiblemente los poros.
- Después de abrir el compartimento de cebado, busque una pequeña burbuja de aire debajo de la cubierta. Extraiga un pequeño volumen para eliminar cualquier burbuja:
- Poner una pipeta P1000 a 200 µl
- Insertar la punta en el compartimento de cebado
- Girar la rueda hasta que el dial muestre 220-230 ul, para absorber 20-30 ul, o hasta que pueda ver un pequeño volumen de tampón entrando en la punta de la pipeta

USC UNIVERSIDADE DE SANTIAGO DE

Nota: Verificar visualmente que haya un búfer continuo desde el compartimento de cebado hasta la matriz de sensores.

- Cargar 800 µl de la mezcla de cebado en la celda de flujo a través del compartimento de cebado evitando la entrada de aire burbujas. Esperar 5 minutos. Durante este tiempo, preparar la librería para la carga siguiendo los pasos a continuación.
- Mezclar bien el contenido de las microesferas de carga (LB) pipeteando.
- **[IMPORTANTE]** El tubo Loading Beads (LB) contiene una suspensión de bolitas de alta densidad. Estas bolitas se asientan muy rápidamente. Es vital que se mezclen inmediatamente antes de su uso.
- En un tubo nuevo, preparar la librería para la carga de la siguiente manera:
    - 37,5 µl de tampón de secuenciación (SQB)
    - 25,5 µl de microesferas de carga (LB), mezcladas inmediatamente antes de su uso
    - librería de ADN de 12 µl
- Completar el cebado de la celda de flujo:
- Levantar con cuidado la cubierta del compartimento de muestra SpotON para que el compartimento SpotON sea accesible.
- Cargar 200 µl de la mezcla de cebado en el compartimento de cebado de la celda de flujo (no en el compartimento de muestra SpotON), evitando la introducción de burbujas de aire.
- Mezclar suavemente la librería preparada pipeteando hacia arriba y hacia abajo justo antes de la carga.
- Agregar 75 µl de muestra a la celda de flujo a través del compartimento SpotON gota a gota. Asegurarse de que cada gota fluye hacia el puerto antes de agregar la siguiente.
- Volver a colocar con cuidado la cubierta del compartimento SpotON, asegurándose de que el tapón entre en el puerto SpotON, cerrar el puerto de cebado y volver a colocar la tapa del dispositivo MinION.
- Realizar la secuenciación de la muestra con el programa MinKnow.

Finalizando el experimento

- Una vez finalizado el experimento de secuenciación, si se desea reutilizar la celda de flujo, seguir las instrucciones del kit de lavado y almacenar la celda de flujo lavada a 2-8 °C.
- Seguir el procedimiento de devolución para limpiar la celda de flujo y dejarla lista para enviarla a Oxford Nanopore.

5.1.12.1.1 Preparación de una librería de fragmentos amplificados (amplicones) para secuenciación por nanoporos

Este protocolo es muy similar al de preparación de una librería estándar como la mencionada en los apartados anteriores, con la única peculiaridad de que hay que ajustar previamente el número de moléculas de los distintos amplicones para que todos ellos estén igualmente representados y a la hora de secuenciar se obtenga un número similar de lecturas para cada amplicón.

5.1.12.1.2 Preparación de una librería con marcadores secuenciales (*barcoding)* para secuenciación por nanoporos

Hay ciertos experimentos donde hay que secuenciar una gran cantidad de muestras, pero la cantidad de ADN no es suficiente como para correr en una celda en separado. En este caso no se pueden mezclar varias muestras ("*pool*") como en el caso anterior de los amplicones, pues no es posible distinguir unas muestras de otras cuando se alinean contra un genoma de referencia. Para estos casos se puede utilizar un marcado previo de cada una de las muestras de tal forma que cada una las moléculas del ADN de cada muestra están unívocamente identificadas. Este protocolo se denomina de "barcoding" y está basado en el protocolo de ONT NBD_V109_REVB

Brevemente se describe a modo de ejemplo, el protocolo usado para secuenciar varias muestras similares de levaduras, llevado a cabo por la doctoranda Vanesa Hurtado del grupo de Miguel González Blanco, y que posteriormente fue usado de referencia.

"Se partió del ADN de 12 muestras de ADN de levadura (S. Cerevisiae) cada una con 250 ng de gDNA, en vez de 1 µg. Se escalaron los volúmenes de las reacciones consecutivas a la cantidad de gDNA de partida. El primer paso fue la reparación del gDNA (NEBNext FFPE DNA Repair Mix, M6630) y la generación de extremos romos mediante la adición de una cola de poli-A al extremo 3´ (NEBNext End repair/dA-tailing Module, E7546). La ligación del código de barras (Native Barcoding Expansion 1-12, EXP-NBD104)) se llevó a cabo mediante NEBNext Quick Ligation Module (E6056) y la ligasa Blunt (NED Blunt/TA Ligase Master Mix, M0367) llevado a una cantidad total de 700 ng de gDNA en un vol. final de 65 µL. A la combinación de muestras obtenida se le añadió la mezcla de adaptadores (adapter mix II, AMII) y la enzima T4 DNA ligasa (Quick T4 DNA Ligase) para la ligación de los adaptadores, así como el tampón L. Fragment Buffer (LFB) para la eliminación de fragmentos genómicos pequeños. Tras la preparación (Flow cell priming kit, EXP-FLP001) y comprobación del correcto funcionamiento de la celda de flujo (modelo FLO-MIN 106D R9-RevD, Oxford Nanopore Technologies), se cargó la librería de DNA."

5.1.12.2 Preparación de la librería para secuenciación de nanoporos (protocolo Rapid Sequencing gDNA SQK-RAD004).

Si la preparación de las librerías de secuenciación por los protocolos anteriores era relativamente sencilla, la siguiente basada en una transposasa, es incluso más sencilla, e increíblemente rápida de preparar. De hecho, este tipo de librerías son las que se han usado para secuenciar en el espacio (estación espacial) o en regiones inhóspitas del planeta como en los polos, selvas tropicales o desiertos.

Este kit de secuenciación rápida (SQK-RAD004) es el recomendado por la empresa para usuarios que requieren un tiempo de preparación corto y tienen acceso limitado a equipos de laboratorio. Además, el ADN de partida es de tan solo 400 ng de ADN en vez de 1µg como en los protocolos anteriores

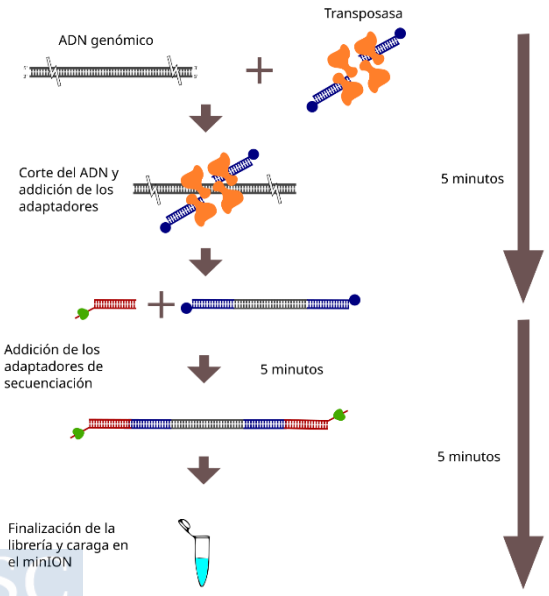


Figura 37: Esquema de una librería de secuenciación rápida SQK-RAD004 ONT

Los pasos en el flujo de trabajo de forma muy resumida serían los siguientes:

1. Preparación del experimento: extracción el ADN, verificación de su longitud, cantidad y pureza (si es posible)
2. Preparación de la biblioteca: fragmentación, "etiquetado" del ADN, y agregación de los adaptadores de secuenciación. Este proceso lo realiza una transposasa, que en una sola reacción química corta el ADN y lo coloca unos adaptadores a los cuales en un paso posterior se le añaden los adaptadores de secuenciación y el "tether"
3. Por último, solo queda cargar la

librería y arrancar la secuenciación (de forma idéntica a los protocolos anteriores (SQK-LSK108 y 109)

Este protocolo fue probado en nuestro laboratorio con resultados prometedores, siendo muy deseable, por su absoluta sencillez y velocidad de preparación. Lamentablemente el rendimiento por celda y librería era alrededor de un 40 % menor que con las librerías SQK-LSK109. Debido a esto desestimamos su utilización de forma rutinaria en el laboratorio.

5.1.12.3 Ejecución del programa MinKnow para la secuenciación de la librería.

Basado en el protocolo: **experiment-companion-minknow-MKE_1013_v1_revCP_11Apr2016-any.pdf**

- El programa MinKnow se ejecutó con los siguientes parámetros:
- El kit seleccionado fue el SQK-LSK 108-109,
- 78 horas y -180 mV como opciones de ejecución,
- Archivos en formato FASTQ (.fastq files) y, FAST5 (.fast5 files) configurados a 4000 lectura por archivo. En las últimas versiones del programa se forzó que los ficheros fueran comprimidos (.gz)
- Los datos se descargaron del ordenador de secuenciación y subidos al CESGA, para su almacenamiento y posterior análisis.

5.1.12.4 Limpieza de las celdas de secuenciación

Tras finalizar la secuenciación las celdas deben limpiarse, para su posterior uso, en caso de que tenga un buen número de poros intactos, o para devolver a ONT para su reciclado, cuando todos los poros están dañados.

En el último caso el proceso consiste en hacer pasar agua destilada (2 ml) por los conductos de la celda, para eliminar cualquier resto de la librería (la cual puede tener material biológico peligroso para la salud). Una vez la celda está limpia se almacena en unos paquetes especiales proporcionados por la propia empresa y se le envían para su reciclado.

En el primer caso, de que la celda se pueda reutilizar, es necesario destruir cualquier resto de nucleótidos de la librería secuenciada para evitar la contaminación entre muestras. Para ello se usa una solución con nucleasas facilitada por ONT en su kit de limpieza. En caso de que se vaya a utilizar inmediatamente se procederá como en el paso final del protocolo de secuenciación mencionado anteriormente. En caso de que se quiera almacenar para su posterior uso se debe rellenar con el buffer de almacenamiento que viene en el kit de limpieza (ONT).

5.2 PIPELINES Y PROGRAMAS BIOINFORMÁTICOS

Para llevar a cabo las tareas de manipulación de datos genómicos se usaron una serie de "*pipelines* informáticas". Usamos este neologismo, pues no hay un término aceptado con las mismas implicaciones de una serie de programas de ordenador encadenados unos detrás de otros para realizar una tarea determinada.

Estas pipelines fueron desarrolladas por nuestro grupo, con mi participación, con el objetivo de que fuesen reproducibles, fiables, y aprovechasen al máximo los recursos que el Centro de Supercomputación de Galicia (CESGA) puso a nuestra disposición. El CESGA, Infraestructura Científico-Tecnológica Singular de España (ICTS), cuenta con el superordenador FinisTerrae, con un sistema operativo Red Hat y un gestor de tareas SLURM.

FinisTerrae II era un sistema de computación de altas prestaciones (HPC) basado en procesadores Intel Haswell e interconectado mediante red Infiniband (FDR@56Gbps) con un rendimiento pico de 328 TFlops y sostenido en Linpack de 213 Tflops. Destacamos sus 306 nodos de cómputo "Thin" cada uno con 2 procesadores Haswell 2680v3, 24 cores, 128 GB memoria y un disco de 1 disco de 1TB. También disponía de 4 nodos con GPUs, cada uno con: GPUs NVIDIA Tesla K80 (2 GPUs por cada tarjeta, 4 GPUs por nodo en total), 2 procesadores Haswell 2680v3, 24 cores , 28 GB memoria y 2 discos de 1TB
FinisTerrae III es el nuevo sistema de computación de altas prestaciones (HPC) que sustituye al anterior. Dispone de una potencia de cómputo total de 4,36 PetaFLOPS. Se compone de 357 nodos interconectados a través de una red Infiniband HDR. Posee 714 procesadores Intel Xeon Ice Lake 8352Y de última generación con 32 núcleos a 2,2Ghz (22.848 núcleos) y 157 GPU (141 Nvidia A100 y 16 Nvidia T4).

Tabla 2: Características del computador de altas prestaciones del CESGA.
Obtenido de la web https://cesga-docs.gitlab.io/ft3-user-guide/overview.html, accedido 18.02.2023

Algunas de las tareas realizadas por ellas son, el alineamiento de las secuencias obtenidas con Illumina (lecturas cortas de todo el genoma WGS) y las obtenidas con los secuenciadores MinION (ONT) (lecturas largas), comprobación de la contaminación de las librerías/datos de secuenciación, búsqueda de mutaciones puntuales y estructurales, búsqueda de alteraciones en el número de copias, estadísticas, análisis y búsqueda de retrotransposones.

Se describen a continuación algunas de las más importantes. Como se comenta más adelante alguno de los datos usados en el presente trabajo (inserciones somáticas de retrotransposones, inferencia de los elementos fuente a partir de nucleótidos diagnóstico) fueron obtenidas por otros miembros del grupo, para sus propios trabajos de investigación o para la publicación de un artículo científico (pendiente de publicación). Para ello hicieron uso de otras pipelines distintas cuya descripción queda fuera de los objetivos de esta Tesis.

5.2.1 Alineación de datos de secuenciación de Illumina (lecturas cortas "pair end")

El conjunto de datos de secuenciación de extremos emparejados ("*Pair end"*) se alinearon contra el genoma de referencia Hg38 usando BWA-mem 0.7.17-r1188 (Li 2013) con configuraciones predeterminadas. Se utilizó samtools v.1.9 (Li et al. 2009) para ordenar e indexar los archivos. Las lecturas duplicadas se marcaron usando el paquete biobambam/bammark (Tischler y Leonard 2014).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.2.2 Basecalling de los datos sin procesar (raw) de lecturas largas ONT

El programa usado para controlar el MinION y la secuenciación se denomina MinKnow. Este programa por defecto lanza el basecaller para convertir en tiempo real los datos de secuenciación sin procesar (raw) en secuencias de nucleótidos en formato FASTQ. Las primeras versiones del programa usaban el Albacore v2.0.1, que fue rápidamente substituido por el Guppy. Este programa fue usado en el presente trabajo, desde la versión inicial v2.1.3 hasta la versión 6.0.6.

En ciertas ocasiones cuando el número de dispositivos MinION secuenciando concurrentemente era muy alto (>5) o la GPU no estaba funcionando correctamente, se optaba con ejecutar el MinKnow en modo sin "*basecalling*", de tal forma que se generaban los datos en bruto, por lo que era necesario hacerlo posteriormente, ejecutando para ello los programas Albacore o Guppy. En las primeras versiones del programa (Albacore) este generaba millones de ficheros, uno por cada lectura, haciendo inmanejable su procesado, ya que se colapsaba el sistema de ficheros tanto de los equipos de secuenciación como luego en el HPC del CESGA. Por eso fue necesario desarrollar un programa para su gestión, (ver apartado 6.2.1.1 MGD_agente: el problema de los millones de ficheros). Posteriores mejores de los programas de ONT, volvieron obsoleto este programa, puesto que directamente generaban ficheros multilectura (4000 por fichero). Tiempo después los programas de *basecalling* empezaron a generar los ficheros comprimidos en formato .gz, evitando que tuviésemos que hacerlo nosotros.

5.2.3 Alineación de datos de secuenciación de lecturas largas MinION (ONT)

Esta pipeline consiste en el alineamiento de los ficheros de secuencias .fastq.gz de lecturas largas obtenidos de la secuenciación de con los dispositivos MinION de la empresa Oxford Nanopore Technologies. Para ello se usa el programa minimap2 (versión 2.20), contra un genoma de referencia, (en nuestro caso el Hg38). Los ficheros *.sam* resultantes se comprimen en ficheros *.bam* con Samtools 1.12, y posteriormente son ordenados e indexados con la misma herramienta. En todo momento la pipeline deja un log del resultado de los programas, permitiendo así su trazabilidad. Por último, se generan estadísticas tanto de la alineación (mosdepth 0.2.6), como de la calidad de la secuenciación (NanoPlot 1.39.0).

5.2.4 Alineación de datos de metilación. Pretratamiento con bisulfito + Illumina + Bismark

Uno de los programas más usados para alinear secuencias obtenidas tras el pretratamiento del ADN con bisulfito y su secuenciación con Illumina es el programa Bismark del instituto Babraham. (https://www.bioinformatics.babraham.ac.uk/projects/bismark/) (versión 0.23.0)

Indicamos a continuación los pasos (y *scripts*) necesarios para correrlo

1. Preparación del genoma de referencia.

```
En la carpeta que contiene el genoma de referencia (Hg38)
```

```
#creation of a folder with a simlink to the reference genome of interest.
cd /mnt/netapp2/mobilegenomes/0/0_reference/0_homo_sapiens
mkdir 1_bismark_prepared_genome
cd 1_bismark_prepared_genome
ln -s  ../hg_38/hg38.fa hg38.fa

screen -S bismark
compute -c 24
module load gcccore/6.4.0 python/3.8.1
module load gcccore/6.4.0 bowtie2/2.4.1
#launch preparation of the ref genome for bismark
/mnt/netapp2/uscmg_aplic/0_external_tools/32_bismark/Bismark-0.23.0/Bismark-
master/bismark_genome_preparation --verbose .
```

Código 1: Preparación del genoma de referencia para Bismark.

2. Recorte de los adaptadores de la librería: Para ello se usó el programa trim_galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ Versión 0.6.6 / Cutadapt versión: 2.10) con los parámetros por defecto.
3. Alineamiento de las secuencias con Bismark, deduplicación, extracción de la información de la metilación y generación de estadísticas

```
/mnt/netapp2/uscmg_aplic/0_external_tools/53_sbatch_bismark/b_bismark_0.23.0.sh
#!/bin/bash
#SBATCH --mail-type=all #Envía un correo en cada evento
#SBATCH --mail-user=fjtemes@gmail.com #Dirección a la que se envía
#SBATCH -J j-job         # Job name
#SBATCH -o job.%j.out    # Specify stdout output file (%j expands to jobId)
#SBATCH -p thinnodes
#SBATCH -n 1                    # Total number of tasks
#SBATCH -c 24
#SBATCH -t 30:00:00

# exit when any command fails
set -e

module load gcccore/6.4.0 python/3.8.1
module load gcccore/6.4.0 bowtie2/2.4.1
module load samtools

program=/mnt/netapp2/uscmg_aplic/0_external_tools/32_bismark/Bismark-
0.23.0/Bismark-master/bismark

#fastq 1 2 trimed with trim_galore
# /mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2087_trimmed/DNAH2087-bis_1_val_1.fq.gz \
# /mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2087_trimmed/DNAH2087-bis_2_val_2.fq.gz

reference_genome=$1

fastq_file1=$2
fastq_file2=$3

output_directory=$4

#hg19
/mnt/netapp2/mobilegenomes/0/0_reference/0_homo_sapiens/1_bismark_prepared_genome
/
```

```
#hg38
/mnt/netapp2/mobilegenomes/1/0_reference/0_homo_sapiens/1_bismark_prepared_genome
/

echo "**********************************************************************"
echo $0
echo reference genome $reference_genome
echo input fastq1 $fastq_file1
echo input fastq2 $fastq_file2
echo output directory $output_directory
date +"%T"
echo "**********************************************************************"

cd $output_directory

#call bismark
time $program \
--parallel 4 \
-n 1 $reference_genome \
-1 $fastq_file1 \
-2 $fastq_file2

time deduplicate_bismark -p  $base/$base"_1_bismark_bt2_pe.bam"
time bismark_methylation_extractor --gzip --bedGraph
$base/$base".2_1_bismark_bt2_pe.deduplicated.bam"
time bismark2report
time bismark2summary

date +"%T"

echo "process went well (OK)"
```

Código 2: Script SLURM para la preparación de datos para Bismark.

Comandos de lanzamiento

```
 sbatch -A uscmg
/mnt/netapp2/uscmg_aplic/0_external_tools/53_sbatch_bismark/b_bismark_0.23.0.sh
/mnt/netapp2/mobilegenomes/0/0_reference/0_homo_sapiens/1_bismark_prepared_genome
/ /mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2087_trimmed/DNAH2087-bis_1_val_1.fq.gz
/mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2087_trimmed/DNAH2087-bis_2_val_2.fq.gz
/mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/2_alignment_bismark/H2087
sbatch -A uscmg
/mnt/netapp2/uscmg_aplic/0_external_tools/53_sbatch_bismark/b_bismark_0.23.0.sh
/mnt/netapp2/mobilegenomes/0/0_reference/0_homo_sapiens/1_bismark_prepared_genome
/ /mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2009_trimmed/DNAH2009_1_val_1.fq.gz
/mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/1_sequencing/DNAH2009_trimmed/DNAH2009_2_val_2.fq.gz
/mnt/netapp2/mobilegenomes/0/1_projects/1027_CELL-LINES-METILOME-TRANSCRIPTOME-
PE/2_alignment_bismark/H2009
```

Código 3: Ejemplo de lanzamiento del script de preparación de datos para Bismark.

obtención de los datos de metilación.

```
methyl_extract.sh
#!/bin/bash
#-----------------------------------------------------------------
#Launche bismark
#
# FT-III system.
#
# This script requests one dedicated core on one shared node
#-----------------------------------------------------------------

#SBATCH -J meth_ex           # Job name
#SBATCH -o meth_ex.%j.out    # Specify stdout output file (%j expands to jobId)
#SBATCH -c 32                     # Total number of tasks
#SBATCH -t 20:00:00              # Run time (hh:mm:ss) - 1.5 hours
#SBATCH --mem=20G                  # Total memory demanded - 20 GB RAM

# Load any necessary modules
# Loading modules in the script ensures a consistent environment.
#module load cesga/2020
#module load samtools

# check the hostname
hostname

/mnt/netapp2/uscmg_aplic/0_external_tools/32_bismark/Bismark-0.23.0/Bismark-
master/bismark_methylation_extractor --bedGraph --cytosine_report   --parallel 10
--genome_folder /mnt/netapp2/mobilegenomes/1/0_reference/hg38_no_alt DNAH2087-bis
_1_val_1_bismark_bt2_pe.bam
```

Código 4: Lanzamiento de extracción de datos de metilación de Bismark.

5.2.5 Alineación de datos de metilación. Lecturas largas obtenidas con ONT+Megalodon

Obtener los datos de alineación de metilación con la información de metilación es tan sencillo como llamar al programa Megalodon con los siguientes parámetros.

```
#!/bin/bash
#-----------------------------------------------------------------
#Launch Megalodon whole genome
#
#/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_javi/2_meme
# FT-III system.
#
#-----------------------------------------------------------------

#SBATCH -J megaWG           # Job name
#SBATCH  -o  09_logs/Megalodon_WG.%j.out       #  Specify  stdout  output  file  (%j
expands to jobId)
#SBATCH -c 64                      # Total number of tasks
#SBATCH --gres=gpu:a100:2
#SBATCH -t 100:00:00               # Run time (hh:mm:ss) - 100 hours
#SBATCH --mem=128G                   # Total memory demanded - 128 GB RAM


# Load any necessary modules
# Loading modules in the script ensures a consistent environment.
module load cesga/2020

# check the hostname
hostname
sample=$1

base_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_javi/
2_meme/
output_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_jav
i/2_meme/04_Megalodon_WG/"$sample"/"$sample"_Megalodon_WG/
mkdir -p  $output_dir

source
/mnt/netapp2/uscmg_aplic/3_environments/python/16A_Megalodon_2.5/bin/activate

Megalodon \
/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/01_projects/01_MEIGA-
LR/0/1_projects/3051_MEIGA_SCREENING-
MN/1_sequencing/"$sample"/1_symbolic_link_fast5_pass/ \
--guppy-server-path  /mnt/netapp2/uscmg_aplic/0_external_tools/33_guppy/ont-guppy-
gpu_6.1.3/bin/guppy_basecall_server \
--guppy-timeout 50 \
--guppy-config dna_r9.4.1_450bps_hac.cfg \
--remora-modified-bases dna_r9.4.1_e8 hac 0.0.0 5hmc_5mc CG 0 \
--outputs basecalls mod_mappings mappings mods per_read_mods \
--reference
/mnt/netapp2/genomas_enfermedad/1/0_reference/0_hg38/GCA_000001405.15_GRCh38_no_a
lt_analysis_set.fna \
--processes 20 \
--devices 0 \
--mod-motif m CG 0 \
--verbose-read-progress 3 \
--write-mods-text \
--mod-min-prob 0 \
--overwrite \
--output-directory
/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_javi/2_meme/04
_Megalodon_WG/"$sample"/"$sample"_Megalodon_WG/ \
--sort-mappings \
```

Código 5: Lanzamiento de Megalodon.

El programa genera los siguientes ficheros de salida

1. mappings.bam	33G
2. mappings.summary.txt	2,7G
3. modified_bases.5hmC.bed	3,4G
4. modified_bases.5mC.bed	3,5G
5. mod_mappings.bam	32G
6. mod_mappings.sorted.bam	25G
7. mod_mappings.sorted.bam.bai	42M
8. per_read_modified_base_calls.db	121G
9. per_read_modified_base_calls.txt	188G

Tabla 3: Ficheros de salida de Megalodon.

Algunos archivos contienen la información de la metilación de citosinas individuales por lectura (8, 9) y su media agrupada por posición (2), así como la información desglosada en datos de metilación (3) y de hidroximetilación (4). También genera varios ficheros de alineamiento ".bam", uno con el formato estándar para comprobar el alineamiento de las lecturas al genoma de referencia (1) y otro con la información de metilación (etiquetas mm, ml) (5,6,7).

5.2.6 Búsqueda de retrotransposones con Meiga

Existen gran cantidad de programas para la búsqueda y caracterización de inserciones somáticas de retrotransposones (incluido los LINE1) (ver apartado 3.5.3 Herramientas para la búsqueda y análisis de retrotransposones), no obstante no hay demasiados que se basen en lecturas largas, por ello en nuestro grupo, Genomas y enfermedad, con gran experiencia en el desarrollo de programas de este tipo (Trafic, Meiba, Viga), se decidió crear un programa específico. El resultado es MEIGA-LR, (pendiente de publicación) proyecto liderado por la estudiante de doctorado Sonia Zumalave con la colaboración del resto del grupo, entre los cuales me incluyo. MEIGA-LR es un método para detectar inserciones de elementos móviles y otros tipos de integraciones en los genomas usando datos de secuenciación de lecturas largas.

Junto con este programa se desarrolló la versión para lecturas cortas MEIGA-SR (no publicado), programa que sustituye a Trafic también desarrollado por nuestro grupo para correr en el genoma de referencia Hg19 usando lecturas cortas.

MEIGA-SR se utilizó para la selección de las muestras con mayores tasas de retrotransposición de una cohorte de más de 150 tumores, de los cuales disponíamos los datos de secuenciación de Illumina.

MEIGA-LR tiene varios modos de ejecución: *"Single"* para buscar todos los elementos móviles que no se encuentran en el genoma de referencia, y *"paired"* que permite identificar qué elementos son inserciones somáticas al comparar los de una muestra y una referencia (generalmente sangre o tejido adyacente). En nuestro caso los datos del tumor se filtraron con los datos del tejido adyacente o por un método alternativo explicado más adelante. El programa tiene multitud de parámetros, algunos de los más interesantes son los cuales permiten seleccionar cromosomas o regiones, seleccionar el tipo de elementos que nos interesa buscar: inserciones y/o traslocaciones, filtrar por calidad y/o por número de lecturas que soportan un evento, o seleccionar el tipo de

tecnología de secuenciación de los datos. Para el presente trabajo se utilizó el modo *"paired"*, con todos los cromosomas, con el modo de inserciones, con un mínimo de 3 lecturas por evento, y el tipo de tecnología *"ont"*.

A continuación, se muestra una comparativa con datos simulados de la puntuación F1 de MEIGA-LR comparado con programas similares. Como se puede comprobar MEIGA-LR identifica de forma muy notable las inserciones de los LINE1, que son el objeto de estudio del presente trabajo.

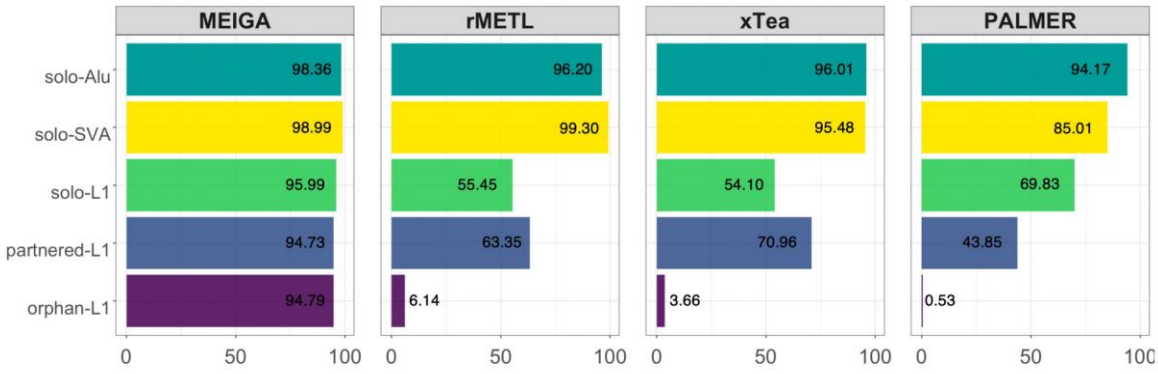


Figura 38: Comparativa de MEIGA-LR con otros programas de descubrimiento de retrotransposición somática.

$$Precisión = \frac{verdaderos\ positivos}{verdaderos\ positivo\ +\ falsos\ positivos}$$

$$Sensibilidad = \frac{verdaderos\ positivos}{verdaderos\ positivos\ +\ falsos\ negativos}$$

$$Puntuación\ F1 = 2 * \frac{precisión\ *\ sensibilidad}{(precisión\ +\ sensibilidad)}$$

Figura 39: Matriz de confusión

El programa se corrió en usando el tejido adyacente para filtrar los transposones germinales en las muestras tumorales en caso de disponer de él (10) o sangre (2 tumores y 2 líneas celulares). Para el resto de las muestras se corrió el programa usando la muestra PD0277c para realizar el filtrado de los transposones germinales. Se utilizó esta muestra debido a que tiene una buena cobertura y un buen N50, permitiendo filtrar inserciones somáticas comunes en la población. Para estas muestras se realizó un segundo filtrado usando las inserciones obtenidas usando el programa xTea sobre los datos de lecturas cortas de los tejidos adyacentes de cada muestra tumoral.

Nota: los datos resultantes se encuentran en el repositorio:

<../5_datos/*_INS_MEIGA.somatic_PASS.tsv.gz >

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.2.7 Búsqueda de retrotransposones con xTea, rMelt, PALMER

xTea(Chu et al. 2021) fue desarrollado por el grupo de investigación de J. Park para la búsqueda de retrotransposones utilizando lecturas cortas (Illumina). En el presente trabajo se utilizó para obtener las inserciones somáticas y germinales de retrotransposones en las muestras de tejido adyacente. Esta información fue posteriormente usada para el filtrado de transposones germinales en las muestras tumorales de aquellas muestras para las cuales no disponíamos de datos de secuenciación largas de ONT del tejido adyacente.

Tanto Melt (Gardner et al. 2017) usado para la búsqueda de elementos móviles en el proyecto 1000 genomas, como PALMER (Zhou et al. 2020) específicamente diseñado para el descubrimiento de LINE1 específicos de humanos, fueron usados como referencia a la hora de evaluar MEIGA-LR.

5.2.8 Inferencia de los haplotipos (phasing) con whatshap

Varios son los motivos por los cuales es muy interesante conocer a qué alelo pertenecen cada una de las lecturas de los datos de secuenciación, es decir tratar de reconstruir los haplotipos. Uno de ellos es para realizar comprobaciones de los métodos. Las inserciones somáticas de elementos móviles solo se producen en una de las cromátidas, por lo que tener información del haplotipo nos permite filtrar falsos positivos. Es muy interesante comprobar si los patrones de metilación se producen solo en la cromatina donde se inserta el elemento móvil, o si la inserción tiene alguna preferencia por cromátidas con un patrón determinado de metilación. También es interesante saber si ciertas regiones del genoma tienen imprimación de los progenitores o saber de qué progenitor proviene un determinado rasgo genético o mutación.

La aparición de lecturas largas ha permitido distinguir los diferentes haplotipos a los que pertenecen la mayor parte de las secuencias. Para ello los programas para hacer "*phasing*" que es el anglicismo de la segregación por alelos, se basan en variaciones puntuales diagnóstico, las cuales permiten distinguir a que haplotipo pertenece una determinada lectura. En nuestro grupo el programa seleccionado es whatshap 1.2.1 (Martin et al. 2016; Garg, Martin, y Marschall 2016). Este tipo de programas tienen problemas para obtener los haplotipos de tumores, debido a las variaciones de número de copias, por ello hemos usado una aproximación híbrida en la que usamos los SNP de los datos de la muestra de tejido normal, secuenciados con Illumina. Con esta información obtenemos los haplotipos de la muestra tumoral de los datos de lecturas largas secuenciados con los MinION de ONT. El flujo de trabajo consiste inicialmente en obtener las variaciones puntuales (SNP) de la muestra para lo que usamos FREEBAYES v1.3.2-40-gcce27fc (Garrison y Marth 2012) un programa para detectar variantes genéticas usando probabilidad bayesiana, diseñado para encontrar pequeños polimorfismos(se forzó el parámetro -C 5 para requerir al menos 5 lecturas del haplotipo alternativo). Como comentábamos anteriormente, la información que se usa es la de las lecturas cortas de Illumina del tejido adyacente de la muestra que nos interesa. Una vez obtenidos los SNP whatshap asigna un haplotipo a cada una de las lecturas largas de ONT (no siempre es posible). Para ello se ejecuta el programa en modo *"PHASE"* y posteriormente en modo *"HAPLOTAG"*. Al final obtenemos un fichero "*.bam"* con la

información del haplotipo en cada lectura. Esta información es procesada por un par de programas desarrollados por nuestro grupo (por Sonia Zumalave (no publicado)) para extraer la información en un formato usable por Methyl-Meiga.

5.2.9 Comprobación de la integridad de los ficheros fast5

En ciertas ocasiones el programa Megalodon detenía su ejecución con una serie de errores relativos a la imposibilidad de abrir un fichero fast5. En la secuenciación de una muestra puede haber miles de ellos. Por lo que se necesitaba un programa para comprobar su integridad.

```
Traceback (most recent call last):
  File "/home/usr-adm/bin/miniconda3/lib/python3.7/threading.py", line 926, in _bootstrap_inner
    self.run()g: 852reads [00:18, 45.51reads/s, samples/s=6.58e+6]
  File "/home/usr-adm/bin/miniconda3/lib/python3.7/threading.py", line 870, in run                                    | 1866/10000
    self._target(*self._args, **self._kwargs) 0%|                                                                    | 9/10000
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/megalodon/fast5_io.py", line 296, in _fill_files_file_enum_worker
    file_read_ids = get_read_ids(fast5_fn):   0%|                                                                     | 7/10000
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/megalodon/fast5_io.py", line 31, in get_read_ids
    with ont_get_fast5_file(fast5_fn, "r") as fast5_fp:
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/ont_fast5_api/fast5_interface.py", line 12, in get_fast5_file
    if is_multi_read(filepath):
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/ont_fast5_api/fast5_interface.py", line 40, in is_multi_read
    with MultiFast5File(filepath, mode='r') as f5_file:
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/ont_fast5_api/multi_fast5.py", line 15, in __init__
    self.handle = h5py.File(self.filename, self.mode, driver=driver)
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/h5py/_hl/files.py", line 408, in __init__
    swmr=swmr)
  File "/home/usr-adm/bin/megalodon/megalodon_env/lib/python3.7/site-packages/h5py/_hl/files.py", line 173, in make_fid
    fid = h5f.open(name, flags, fapl=fapl)
  File "h5py/_objects.pyx", line 54, in h5py._objects.with_phil.wrapper
  File "h5py/_objects.pyx", line 55, in h5py._objects.with_phil.wrapper
  File "h5py/h5f.pyx", line 88, in h5py.h5f.open
OSError: Unable to open file (truncated file: eof = 290127872, sblock->base_addr = 0, stored_eof = 924387722)
```

Figura 40: Ficheros .fast5 estropeados, detectados por Megalodon.

Recurrimos a un pequeño *script* de Ryan Wick

https://github.com/rrwick/Fast5-to-Fastq/blob/master/fast5_integrity_check.py (descargado 2020.09.16)

para averiguar si había ficheros fast5 corruptos entre los datos de la muestra que estaba causando el error. Efectivamente se encontraron un par de ellos. Una vez eliminados el programa funcionó correctamente.

```
for file in *.fast5;do echo $file;h5dump -nH $file;done > fast5.log
h5dump error: unable to open file "FAO60919_pass_c8a82ede_42.fast5"
h5dump error: unable to open file "FAO60919_pass_c8a82ede_70.fast5"
```

Código 6: Lanzamiento de fast5_integrity_check.

5.2.10 Visualización de los datos de secuenciación con IGV

Trabajar con datos genómicos en forma de tablas resulta muy arduo debido al gran tamaño de estas y la abstracción de los datos. Por ello su visualización es fundamental. Existen varias alternativas, como por ejemplo la visualización en línea en el "*UCSC Genome Browser*". https://genome-euro.ucsc.edu/

Este visualizador, contiene una gran cantidad de información acerca del genoma de referencia, incluida la información de repeticiones y elementos móviles (indicado por una flecha roja en la figura 41).

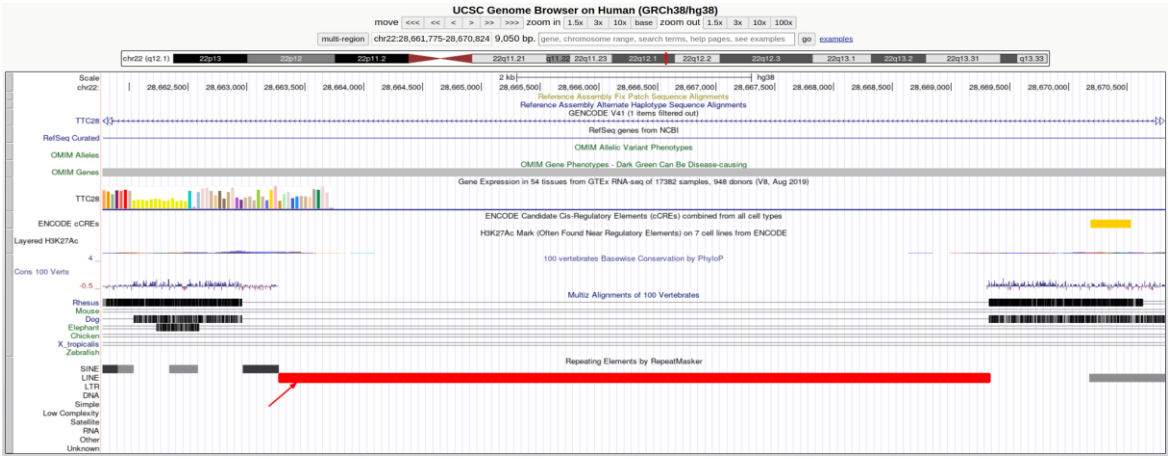


Figura 41 UCSC Genome Browser.

No obstante a la hora de visualizar información de datos propios de forma local (*.bam, .bed, .vcf*) y sobre todo de información de metilación los visualizadores en línea son limitados, por loque se recurrió a visualizadores instalados en nuestros ordenadores y en el CESGA. Existen varias opciones, de las cuales hemos probado alguna como por ejemplo Jbrowse 2 o IGV. La herramienta que mejor se adecúa a nuestras necesidades es el visualizador **IGV, (Integrative Genomics Viewer)** del Broad Institute. Es un programa en Java que se puede instalar en prácticamente en cualquier plataforma y sistema operativo. En él se puede indicar un genoma de referencia de dominio público (por ejemplo, el Hg38), o cargar uno generado por nosotros, funcionalidad utilísima, ya que para el estudio de la metilación generamos genomas de referencia "artificiales" que contienen los LINE1 y las inserciones somáticas de los elementos móviles, todos ellos con orientación positiva para facilitar su interpretación y análisis, así como 20,000 pares de bases corriente arriba y abajo para situarlos en el contexto donde se encuentran. Otra importante funcionalidad es que se pueden cargar las lecturas de varios genomas distintos para su comparación (.bam) así como diversos ficheros con anotaciones (*.bed, .vcf, .gff, .gff3*)

Además, permite visualizar la información de la calidad del alineamiento, las variaciones puntuales y si existen regiones que no alinean correctamente con el genoma de referencia, las cuales son posibles indicaciones de alteraciones genómicas, inserciones, deleciones, inversiones, duplicaciones, translocaciones, o inserciones de elementos móviles.

Por último y no menos importante para el desarrollo de esta Tesis es la posibilidad de visualizar la información de la metilación. Tanto si las muestras fueron tratadas con bisulfito, (o su emulación) como la información contenida en "etiquetas" en las lecturas largas generadas por los "*basecallers*" como *Guppy* , *Megalodon* o *Remora*. (en los ficheros *.bam*)

En la figura siguiente (del visualizador IGV) se muestra cómo se visualiza el mismo elemento que mostramos en el *UCSC Genome Browser*, pero usando un genoma de referencia creado por nosotros. En la parte superior se muestra la visualización en

modo *"bisulfito" ("colour visualization by bisulfite mode")*. En la parte inferior, en el panel superior se puede observar la visualización de los nucleótidos estándar y en el inferior el nuevo sistema de visualización de las "etiquetas" generadas por los *basecallers* ("colour visualization by tag").

Para el presente trabajo se decidió correr el *basecaller* de metilación con el modo "etiquetas" (tags) (además de ser el recomendado por el fabricante) en vez de la emulación bisulfito a pesar de que inicialmente la visualización en el IGV no era tan clara. Por ello en el programa Methyl-Meiga implementé una visualización al estilo bisulfito a partir de la información contenida en las etiquetas de los *ficheros .bam* generados. Posteriormente el IGV (a partir de la versión de febrero de 2023) permite la visualización de la metilación de forma clara (similar a la emulación con bisulfito) a partir de los ficheros ".bam" con etiquetas.

En la figura siguiente (42) se muestra una captura de pantalla de un genoma artificial constituido por un LINE1 (recuadro naranja) y las regiones adyacentes (20,000 pb) corriente arriba y corriente abajo . En el panel superior se muestra un fichero al que se le ha realizado un basecalling con Megalodon en modo "bisulfito" donde los cuadrado rojos representan citosinas metiladas y los azules citosinas sin metilar. En el panel intermedio se muestra otro LINE1 mostrando un alineamiento y visualizaciones convencionales. En el panel inferior se muestra el nuevo modo de visualización, "de etiquetas" con las variaciones epigenéticas, donde cuadrados rojos de distinta intensidad muestran la metilación (más intenso cuanto mayor sea la probabilidad de la citosina de estar metilado) y en azul la hidroximetilación.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

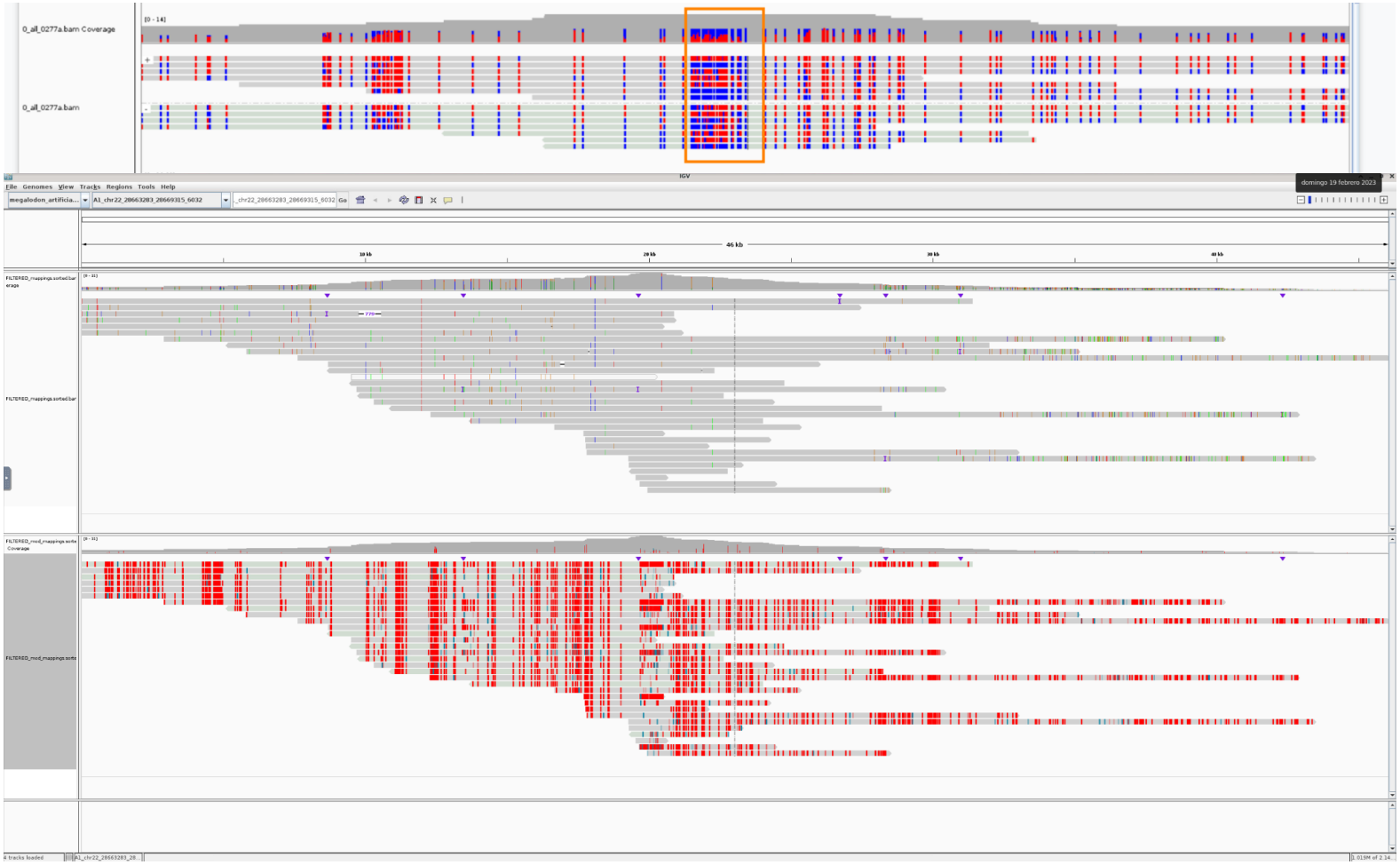
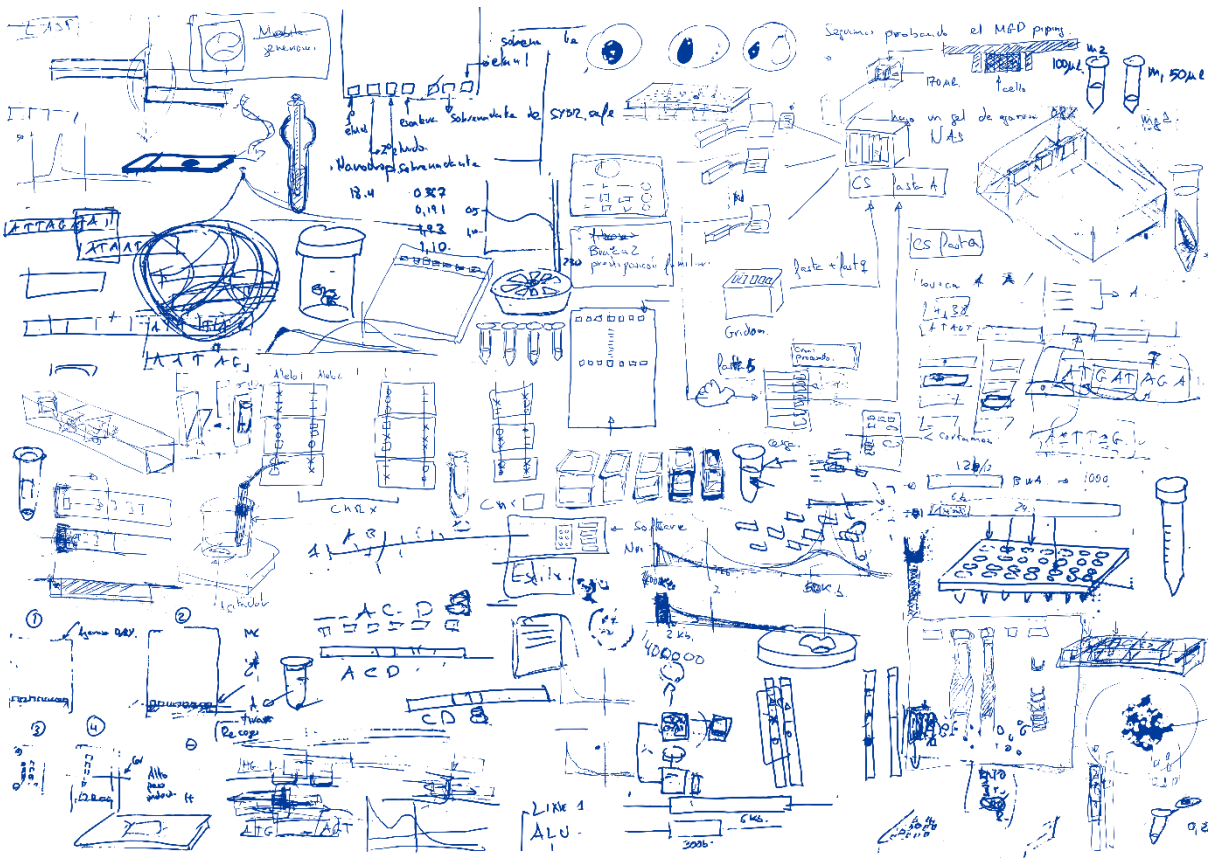


Figura 42: Visualización de la metilación de un LINE1 con IGV.

6. RESULTADOS Y DISCUSIÓN I. RESULTADOS METODOLÓGICOS



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Recortes de esbozos y diagramas del libro de notas de laboratorio”, bolígrafo sobre papel (durante la Tesis 2018-2023)

Este proyecto de Tesis se planteó tras investigaciones previas realizadas en nuestro grupo por el Dr. José C. Tubío et al, relativas a la metilación de los promotores de los LINE1, y plasmadas en el artículo "**Mobile DNA in Cancer. Extensive Transduction of Nonrepetitive DNA Mediated by L1 Retrotransposition in Cancer Genomes**" (Tubio et al. 2014). Estos trabajos seminales abrieron una nueva vía de investigación, que, debido a las limitaciones de la tecnología existente en su momento, dejaban una serie de cuestiones sin resolver, algunas de las cuales se tratan aquí. Para ello hizo falta poner a punto herramientas y tecnología que proporcionasen información más detallada que la disponible al inicio de esta Tesis, consistente en datos de secuenciación genómica de 2º generación (lecturas cortas). Parte de los esfuerzos de este trabajo de investigación fueron el adaptar e implementar la tecnología de secuenciación de tercera generación, de molécula única (lecturas largas), de Oxford Nanopore Technologies (ONT) (ver apartado 6.1), para el estudio de los LINE1 en general y de sus patrones de metilación en concreto.

Una vez seleccionada una cohorte de muestras de tumores con alta tasa de retrotransposición (ver apartado 7.), se procedió a su secuenciación con esta nueva tecnología, inicialmente por mí, y una vez conseguimos estandarizar el proceso por nuestros técnicos de laboratorio. Al mismo tiempo que se iban procesando los datos de secuenciación se fue desarrollando el programa para analizar los patrones de metilación de los LINE1 en cáncer (ver apartado 6.2.2 Methyl-Meiga). Una vez ejecutado el programa (Methyl-Meiga) obtuvimos un conjunto de tablas e imágenes que fueron posteriormente analizadas para tratar de responder a las cuestiones biológicas planteadas al inicio de la Tesis. Para estos análisis se crearon programas y algoritmos específicos usando libros (*Notebooks*) de Python.

Para facilitar la compresión de los resultados se describen primero los procedimientos, test y experimentos de laboratorio realizados para la puesta a punto y adaptación de la tecnología de secuenciación de molécula única de ONT a las necesidades de nuestro grupo de investigación y para la realización de este trabajo. Se describen también de varios dispositivos creados específicamente para solventar varios problemas que se nos presentaron. A continuación, describo los programas usados para generar y analizar los datos. Dedico un capítulo aparte a las muestras y datos usados y otros a los resultados biológicos, la discusión general y a las conclusiones.

6.1 ADAPTACIÓN E IMPLEMENTACIÓN EN NUESTRO LABORATORIO DE LAS TECNOLOGÍAS DE SECUENCIACIÓN DE MOLÉCULA ÚNICA.

6.1.1 Trabajo de laboratorio

Nuestro grupo de investigación cuenta con una amplia experiencia en el campo de la genómica y la secuenciación de primera y segunda generación, tanto en el trabajo de laboratorio como en el procesamiento de datos. Sin embargo, al comienzo de este proyecto de investigación, no teníamos experiencia en la tecnología de secuenciación de lecturas largas de Oxford Nanopore Technology (ONT), tecnología de secuenciación de tercera generación que en aquel momento estaba emergiendo. En este capítulo, describimos detalladamente los experimentos, test y procesos de adaptación y optimización realizados para la generación los datos genómicos indispensables para la consecución de esta Tesis.

USC DE COMPOSTELA

6.1.1.1 Secuenciación del fago λ

Este apartado cumple dos objetivos: por un lado, describir con un ejemplo sencillo el proceso de secuenciación y por otro el de mostrar los pasos y comprobaciones que se realizaron y se realizan para secuenciar las muestras usadas en esta Tesis (y otros trabajos), con la tecnología de ONT, en concreto con los dispositivos MinION. También se describen los gráficos y estadísticos que se generan en el proceso.

Siguiendo las recomendaciones del fabricante (ONT) para comprobar que el equipo funcionaba correctamente MinION (hardware) + MinKnow (Software) y que las librerías estaban bien construidas, se procedió a secuenciar el ADN proporcionado por el propio fabricante. Este proviene del ADN del fago λ, también denominado bacteriófago Lambda, o simplemente fago Lambda. Este virus es un bacteriófago complejo con ADN lineal bicatenario de aproximadamente 48 Kpb descubierto en 1952 por Esther Lederberg (Lederberg y Lederberg 1953).

El ADN proporcionado, (con una concentración de 50 ng/µl) es de altísima calidad. La longitud de sus cadenas de ADN es muy grande, abarcando muchas de ellas la longitud íntegra del virus (48 kb), siendo por lo tanto ideales para la puesta a punto del sistema. En el momento en que se realizó este experimento sólo disponíamos de un espectrofotómetro (NanoDrop), con el cual se realizó la comprobación de la calidad del ADN suministrado, que como se puede comprobar en la figura 43 es de pureza a aceptable. También se migró en multitud de ocasiones en geles de agarosa como el de la figura 78, donde se puede comprobar su alto peso molecular y su integridad.

#	Sample ID	User name	Date and Time	Nucleic Acid Conc.	Unit	A260	A280	260/280	260/230	Sample Type	Factor
1	20180322_lamda	Administrador	22/03/2018 10:16:11	17,0	ng/µl	0,341	0,183	1,86	1,59	DNA	50,00

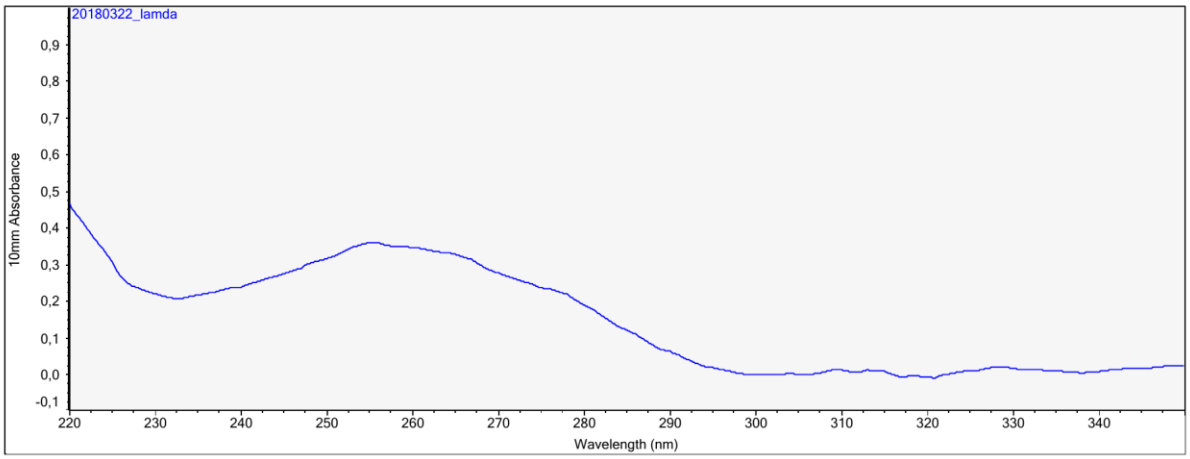


Figura 43: Comprobación de calidad del ADN del fago λ mediante un espectrofotómetro (NanoDrop). Detección de contaminación de proteínas y otros contaminantes (ratios 260/280 y 260/230)

Posteriormente el laboratorio adquirió un fluorímetro QBIT que nos permitió cuantificar con precisión la cantidad de ADN de doble cadena, permitiéndonos comprobar que el ADN facilitado por el fabricante tiene una concentración de alrededor de 50 ng/µl. Con posterioridad el laboratorio adquirió un sistema de electroforesis automatizado TapeStation, lo que nos permitió comprobar que el ADN suministrado contenía moléculas de gran longitud.

Se procedió a la secuenciación siguiendo un protocolo similar al explicado en (5.1.12 Secuenciación de nanoporos mediante la tecnología de Oxford Nanopore Technologies). La secuenciación se realizó durante 6 horas y se obtuvieron 42,989 lecturas, leyéndose 273,966,586 nucleótidos (con una calidad > 7)

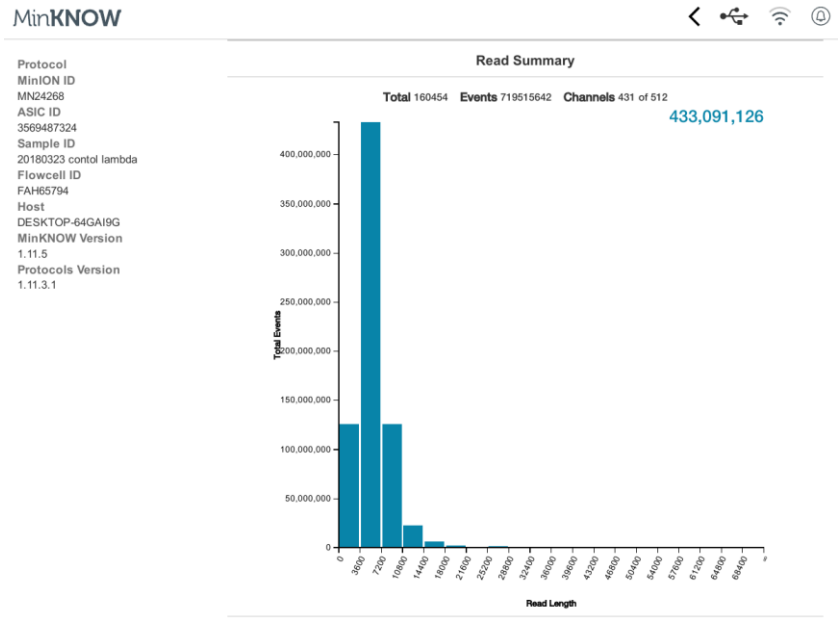


Figura 44: Salida del MinKnow de la primera secuenciación en nuestro laboratorio, un fago λ.

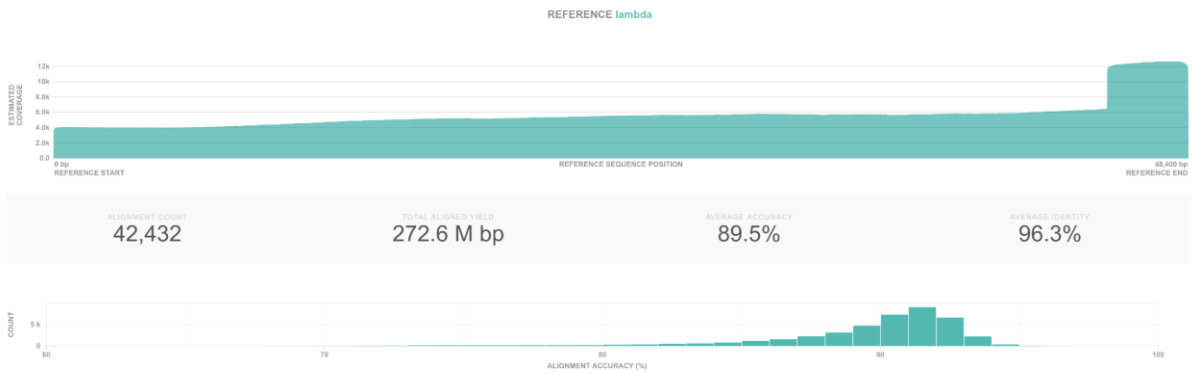


Figura 45: Salida del programa NanoPlot de la primera secuenciación en nuestro laboratorio, un fago λ. Cobertura por nucleótido (panel superior). Número de secuencias vs precisión del alineamiento (panel inferiror).

Los ficheros obtenidos (fastq) se alinearon de forma manual (minimap2, samtools sort, samtools index) de forma similar a como lo realiza la pipeline automatizada (ver apartado 5.2.3)

Las estadísticas de esta primera secuenciación se muestran a continuación. Me gustaría destacar el asombro que nos produjeron semejantes resultados. Aunque actualmente obtener lecturas largas es algo habitual, en su momento (marzo 2018) obtener lecturas con una media de 6,000 pares de bases, muchas con la longitud completa del fago 48 kb) y una de 71,621 pb fue algo realmente increíble, máxime teniendo en cuenta que los secuenciadores de

Illumina eran capaces de secuenciar solo 150 pb, o 800-1000 pb los secuenciadores capilares Sanger.

El resultado del alineamiento se puede ver en la figura 47 obtenida del IGV.

Es de destacar (la numeración corresponde con los números en la figura)

1. La gran longitud de las lecturas, muchas de ellas cubriendo por completo el genoma del virus. Incluso existen algunas lecturas equivalentes al doble del genoma viral, que bien pudieron pueden ser naturales, puesto que el virus se replica mediante un sistema de replicación en círculo rodante, o porque al confeccionarse la librería, la ligasa unió dos moléculas de ADN
2. La asombrosa cobertura de más de 12,000x. Para este rango de tamaño de genomas (48 Kpb) no es práctico manejar semejante volumen de datos. Pero en genomas más grandes como el humano (3000 Mpb) es muy conveniente tener semejante rendimiento.
3. Errores de lectura: Uno de los inconvenientes de esta tecnología era una relativa alta tasa de error por nucleótido, de entre un 5-8 % inicialmente, alrededor de un 1-2% en las últimas versiones de las librerías / *basecallers* usados en el presente trabajo. Las nuevas versiones de las librerías y programas de ONT alcanzan tasas de error que se están acercando mucho a las de otras tecnologías como la SMRT de PacBio-Illumina.

Metric	Value
Mean read length	6,735.8
Mean read quality	10.5
Median percent identity	90.6
Median read length	6,330.0
Median read quality	10.8
Number of reads	42,989.0
Read length N50	8,565.0
STDEV read length	4,319.4
Total bases	289,563,747.0
Total bases aligned	273,966,586.0
Number of reads >Q5	42,989 (100.0%)
Megabases of reads >Q5	289.6Mb
Number of reads >Q7	42,981 (100.0%)
Megabases of reads >Q7	289.5Mb
Number of reads >Q10	32,827 (76.4%)
Megabases of reads >Q10	220.3Mb
Number of reads >Q12	612 (1.4%)
Megabases of reads >Q12	3.5Mb
Number of reads >Q15	0 (0.0%)
Megabases of reads >Q15	0.0Mb

Tabla 4: Estadísticas de secuenciación fago λ

4. 5. 6. Al igual que otro tipo de librerías (Illumina) también se producen inserciones (4), deleciones (5) y lecturas desapareadas (6) (clip reads).

7. Este aumento de la cobertura nos tuvo algún tiempo desconcertados, al final descubrimos que corresponde al ADN de control que se añade a la librería en el paso de "**Reparación de ADN y preparación de bordes romos**", y que consiste precisamente en un fragmento del propio fago λ.

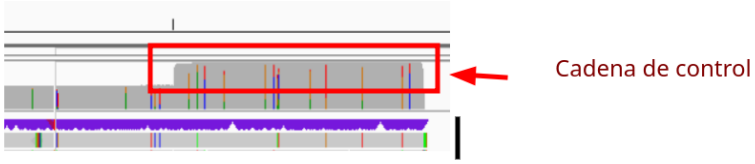


Figura 46: Aumento de cobertura debido a la cadena de control.

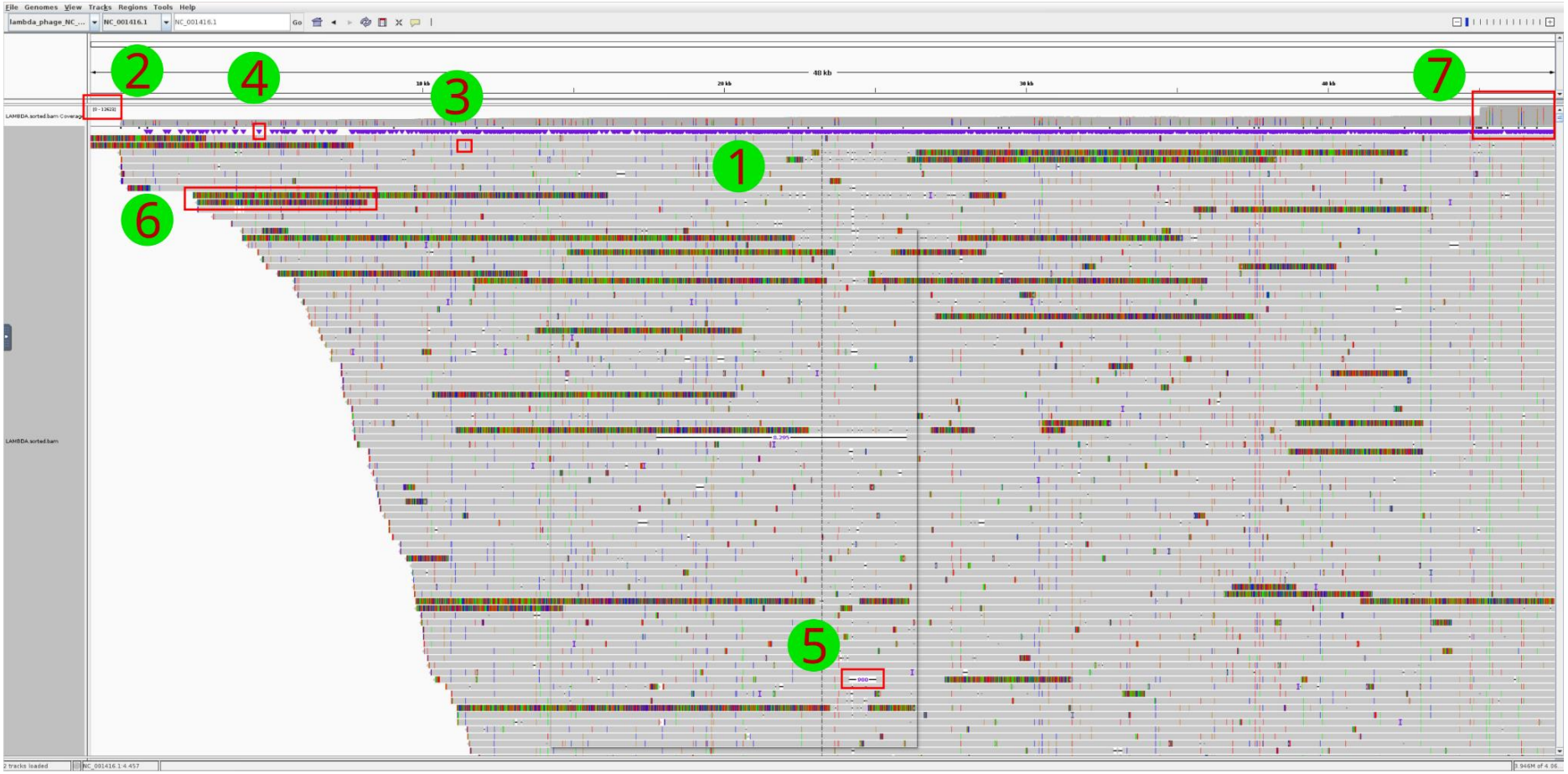


Figura 47: Salida del IGV mostrando un alineamiento de lecturas largas (MinION de ONT).

Con posterioridad y como parte de las pipelines de alineamiento se introdujo un paso en el procesado de los datos para comprobar la calidad de la secuenciación y que consiste en la generación de una serie de gráficos y estadísticas, mostrados a continuación.

En la siguiente gráfica se muestra el número de lecturas por longitud. En el recuadro rojo se señalan las lecturas correspondientes a la cadena de control. Se puede observar que existen lecturas que cubren toda la secuencia del virus, e incluso alguna de 70Kpb, aunque para esta muestra la media y mediana están alrededor de 6-7 Kpb. Varios cambios en la química y la forma de preparación de la librería nos permitieron aumentar esta media de la longitud de las lecturas a más de 20Kpb en algunas muestras con ADN de gran integridad.

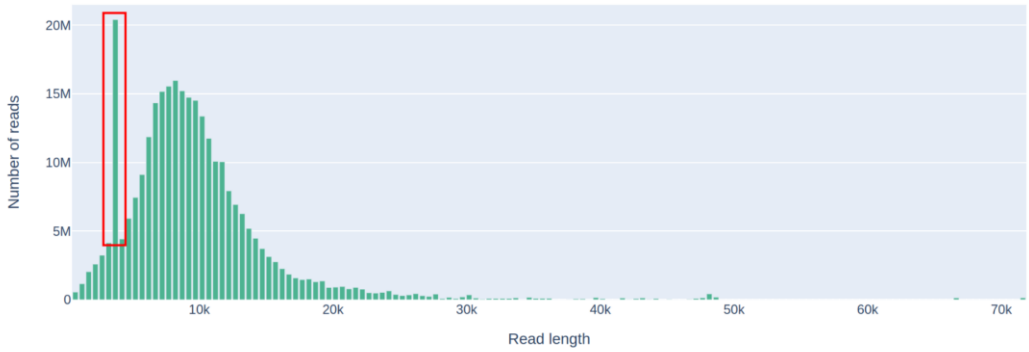


Figura 48: Salida de NanoPlot mostrando el número de lecturas por longitud de la secuencia leídas (Fago λ). En el cuadrado rojo se muestran las lecturas correspondientes a la cadena de control

En la siguiente figura (49) se muestran la calidad de las lecturas frente a su longitud

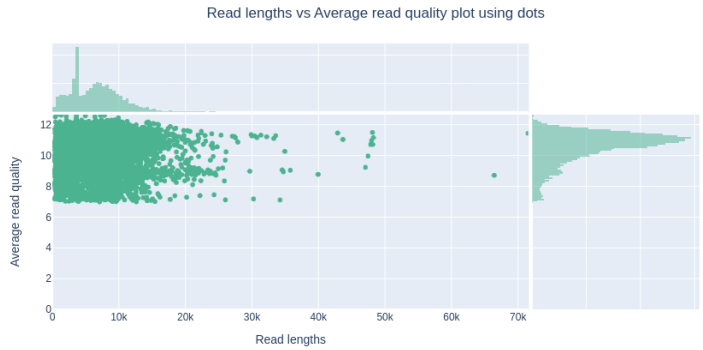


Figura 49: Salida de NanoPlot mostrando la calidad de las lecturas frente a su longitud (Fago λ).

y en ésta (figura 50) el porcentaje de identidad de los alineamientos que como vemos tiene su pico en un 90 % de identidad

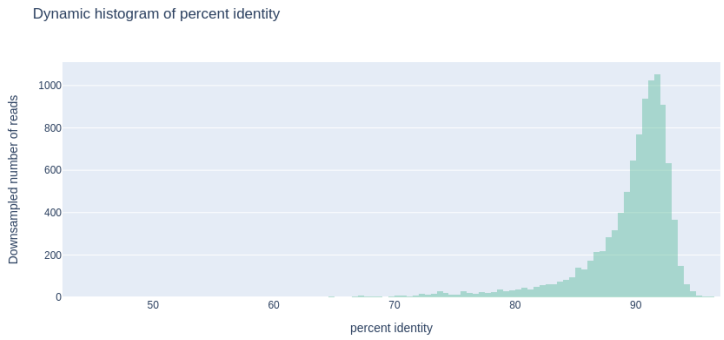


Figura 50: Salida de NanoPlot mostrando el porcentaje de identidad de las lecturas (Fago λ).

6.1.1.2 Secuenciación de un genoma humano.

Tras haber conseguido secuenciar el ADN del fago λ, el siguiente paso lógico era secuenciar el ADN de un ser humano, pues en última instancia nuestro objetivo era el de estudiar los LINE1 en tumores humanos.

Para ello se usó el ADN de un voluntario sano "**sujeto 1**" (con su consentimiento expreso por escrito, figura 51). Se le extrajo sangre periférica mediante punción en uno de los dedos. El ADN se obtuvo mediante el protocolo (5.1.4 Extracción de ADN a partir de sangre y/o líneas celulares) y se purificó mediante el protocolo (5.1.5 Purificación del ADN). Los resultados de analizarlo con el espectrofotómetro son los siguientes:

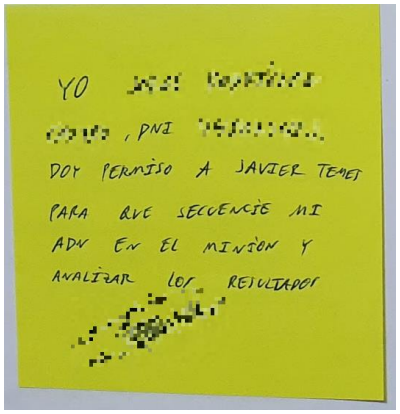


Figura 51: Autorización del "sujeto 1" para la secuenciación de su ADN.

#	Sample ID	User name	Date and Time	Nucleic Acid Conc.	Unit	A260	A280	260/280	260/230	Sample Type	Factor
1		Administrador	26/03/2018 12:35:33	35,0	ng/µl	0,699	0,373	1,87	1,95	DNA	50,00
2		Administrador	26/03/2018 12:39:48	15,2	ng/µl	0,304	0,166	1,83	0,49	DNA	50,00

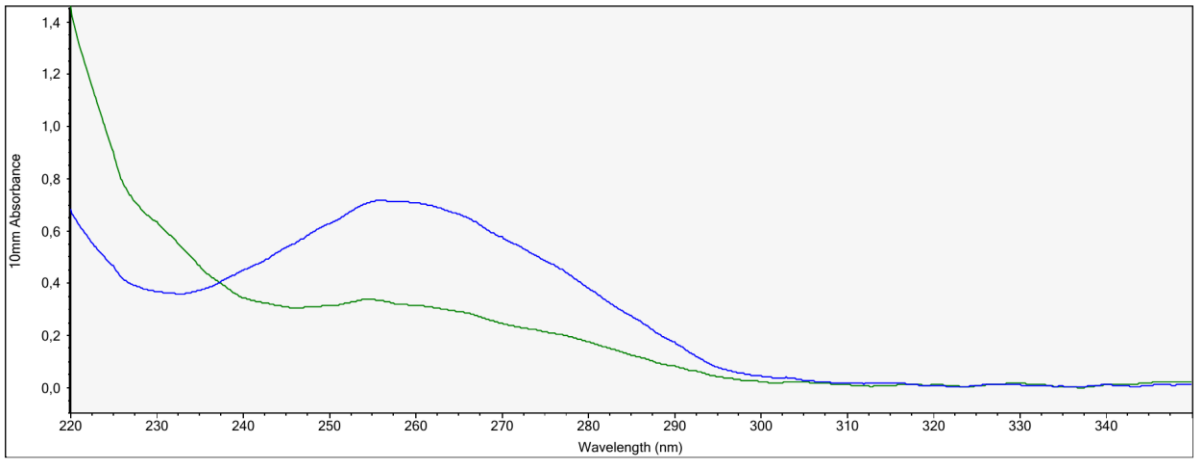


Figura 52: Comprobación de calidad del ADN del "sujeto 1".
Comprobación mediante un espectrofotómetro (NanoDrop), para detectar contaminación de proteínas y otros contaminantes

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

La muestra 1 tenía la concentración y pureza necesaria para realizar una librería de secuenciación ONT (incluso mejores que la del experimento anterior con el fago λ). Se procedió a prepararla mediante el protocolo (5.1.12.1 Preparación de la librería para secuenciación de nanoporos (protocolos SQK-LSK108 y 109)). Se utilizó una celda R9.4.1, secuenciando la muestra hasta "agotar" la celda (>72 horas), produciendo 646,953 lecturas. Las celdas van perdiendo capacidad de secuenciación cuando los poros se estropean, u obstruyen de forma irreversible de tal forma que incluso el cambio de corriente, y por lo tanto la inversión del flujo es incapaz de desatascarlo. Otro motivo por el que la celda va perdiendo capacidad de secuenciación es la disminución de electrolitos y ATP necesarios para el paso del ADN a través del poro. Algunos de estos problemas pueden solventarse tal como se indica en los apartados 6.1.1.9 Recarga de ATP y 6.1.1.10 rellenado.

Los resultados fueron muy esperanzadores de cara a cubrir las necesidades de obtener lecturas largas y en cantidad suficiente como para abarcar los diversos estudios planteados en nuestro grupo de investigación.

Varios meses después la empresa mejoró las celdas y la química, prácticamente doblando el rendimiento, esto unido a mejoras por nuestra parte en la preparación de las librerías y técnicas de relleno de librería y de ATP, nos permitió obtener un rendimiento por celda adecuado para nuestras necesidades (alrededor de un 5x del genoma humano).

Al igual que en el caso anterior, el procesado informático se realizó de forma manual, tal como se indica a continuación. A partir de los ficheros fastq, se alinearon contra el genoma hg19 con el minimap2, y se ordenaron e indexaron con Samtools.

Aquí nos encontramos con uno de los principales inconvenientes de las primeras versiones del MinKnow. Este programa generaba un fichero .fast5 por cada lectura. Por ello la secuenciación de una librería generaba cientos de miles de ficheros (o millones en cuánto fuimos capaces de mejorar el rendimiento obtenido en cada librería). Aunque el sistema operativo Linux es capaz de trabajar con ellos, manejarlos resulta muy engorroso y lento, además, en el CESGA (donde realizamos el procesado final y el almacenamiento de los datos) existe una limitación del número de ficheros por usuario. Fue por tanto necesario desarrollar una aplicación para gestionarlos, descrita en el apartado 6.2.1.1. MGD_agente. Afortunadamente en versiones posteriores del programa los ficheros .fast5 generados eran multilectura,

Metric	Value
Average percent identity	85.5
Fraction of bases aligned	0.9
Mean read length	8,767.2
Mean read quality	9.7
Median percent identity	87.0
Median read length	8,514.0
Median read quality	9.9
Number of reads	646,953.0
Read length N50	10,461.0
STDEV read length	4,754.5
Total bases	5,671,980,750.0
Total bases aligned	5,342,370,370.0
>Q5	646,953 (100.0%)
>Q7	646,422 (99.9%)
>Q10	287,242 (44.4%)
>Q12	1,971 (0.3%)
>Q15	0 (0.0%)

Tabla 5: Estadísticas de secuenciación del "sujeto 1".

conteniendo cada uno de ellos 4000 lecturas, haciendo innecesario el programa "MGD_agente".

Para que la información de la secuenciación de esta librería sea comparable a la información de las librerías de las muestras tumorales, se han vuelto a realinear los datos del "**sujeto 1**" contra el genoma de referencia Hg38 usando la pipeline estándar de secuencias ONT (5.2.3). Se obtuvo una cobertura de 1.6 x con el 90% de las lecturas alineadas.

Los resultados son mostrados en la tabla adjunta. Como se puede comprobar los valores de calidad y rendimiento son muy similares a los de la secuenciación del fago λ.

Las gráficas muestran igualmente unos patrones muy similares, con pequeñas diferencias, como el incremento del número de lecturas, (debido al mayor tiempo de secuenciación), la aparición de una lectura de más de 100,000 pares de bases, y que la cadena de control queda integrada con el resto de las lecturas de tamaño similar, debido al aumento del número de lecturas.

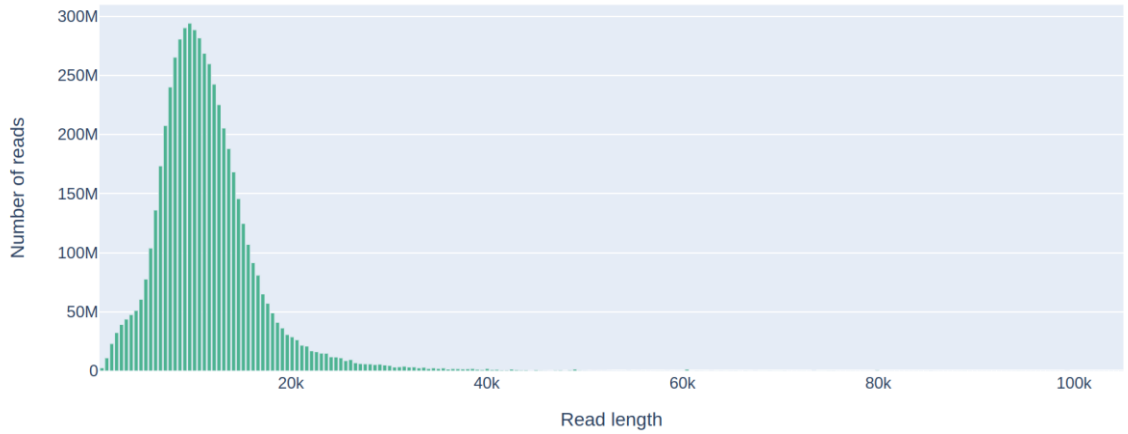


Figura 53: Salida de NanoPlot mostrando el número de lecturas por longitud de la secuencia leídas ("sujeto 1").

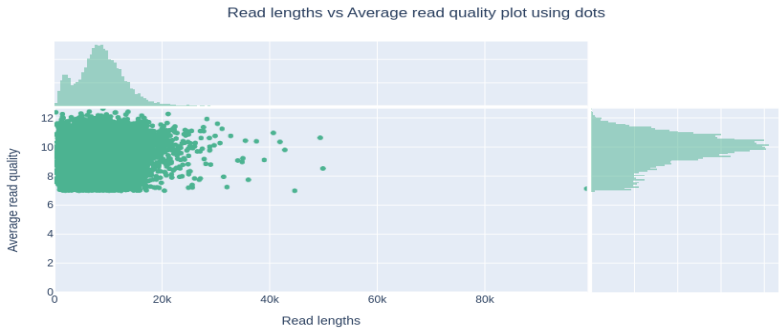


Figura 54: Salida de NanoPlot mostrando la calidad de las lecturas frente a su longitud ("sujeto 1")

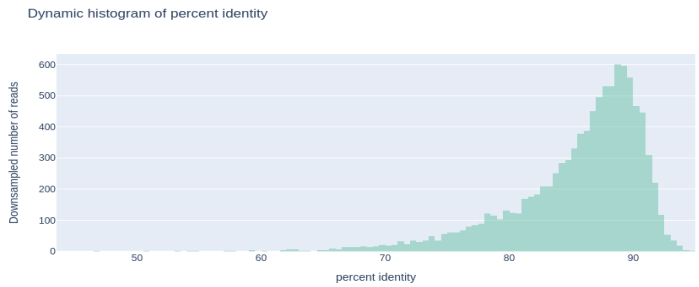


Figura 55: Salida de NanoPlot mostrando el porcentaje de identidad de las lecturas ("sujeto 1").

Nota 1: En las primeras librerías el basecaller se denominaba **Albacore**, el cual internamente usaba un sistema basado en cadenas de Márkov. Posteriormente, todas las librerías usadas en este trabajo se generaron empleando el basecaller **Bonito**, basado en una red neuronal convolucional. Este nuevo algoritmo mejoró la tasa de error en la asignación de nucleótidos, que pasó de cerca del 10 % a alrededor 2-3% aunque a costa de necesitar una o varias tarjetas gráficas de elevadas prestaciones (GPU).

Nota 2: Para la puesta a punto de los secuenciadores de nanoporos, se secuenciaron además ADN de levaduras (proporcionadas por Vanesa Hurtado del grupo de investigación de Miguel González Blanco), tumores caninos facilitados por el estudiante de doctorado Martin Santamarina, y de varios tipos de bivalvos, cada uno con sus particulares características y retos.

6.1.1.3 Secuenciación de amplicones

Como parte del proceso de puesta a punto de la unidad de secuenciación masiva ("*core facility*") se realizaron una serie de experimentos para secuenciar fragmentos de ADN amplificados mediante PCR, a partir de ahora *"amplicones"*. Esto permitió hacer comprobaciones cruciales para el artículo (Rodriguez-Martin et al. 2020) donde se usaron las líneas celulares (NCI-H2009 y NCI-H2087) para validar los algoritmos de detección de reordenamientos mediados por LINE1. Algunos de los amplicones tenían un tamaño superior al proporcionado por los secuenciadores de Illumina (figura 56). La secuenciación capilar Sanger tampoco resultaba práctica debido al gran número de estos amplicones. Por ello la secuenciación de nanoporos resultó ideal.

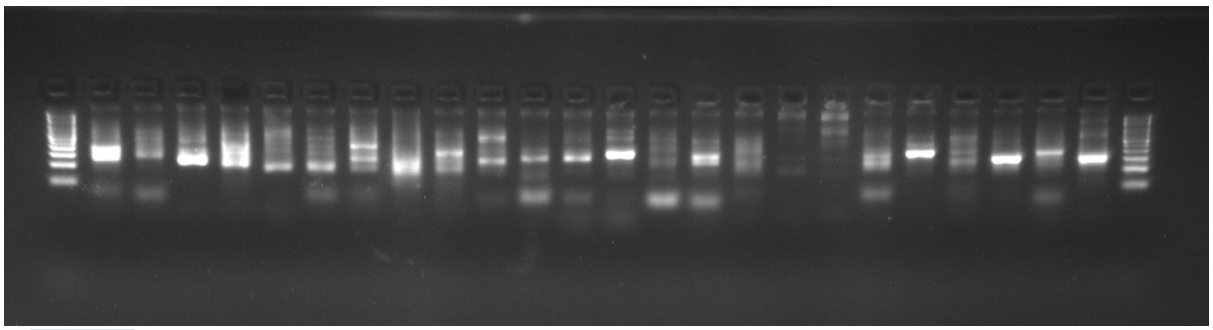


Figura 56: Fragmento de un gel de agarosa de amplicones de la muestra NCI-H2087.
TBE3% 120V 30m 10ul RS -TN-A1-H6, Escalera 1kb en los carriles externos (obtenida por Jorge Rodríguez Castro)

USC DE COMPOSTELA

Este protocolo está basado en el protocolo ACDE_9064_V109_REVE_23MAY2018 de ONT. Para amplificar los fragmentos de ADN que contenían los 20 reordenamientos somáticos mediados por L1, en su mayoría deleciones, se diseñaron cebadores ("*primers*") para realizar varias PCR. Los cebadores de PCR se diseñaron con Primer3 v0.4.0, para amplificar tanto los puntos de interrupción ("*breakpoints"*) de la inserción de LINE1 como su secuencia íntegra. El procedimiento realizado por los técnicos de laboratorio de nuestro grupo fue el siguiente (extraído y adaptado de (Rodriguez-Martin et al. 2020))

"Para el punto de ruptura 5', un cebador directo apuntó a la secuencia de ADN cadena arriba en relación con el punto de inserción mientras que el reverso apuntaba al extremo 5' de la inserción LINE1. Por otro lado, la amplificación del punto de corte 3' se basó en un cebador directo que era específico para el final de elementos específicos de LINE1.

(5'-GGGAGATATACCTAATGCTAGATGACAC-3')

Otro cebador inverso iba dirigido a la secuencia genómica inmediatamente adyacente al punto de corte de inserción 3' (R). En el caso de las transducciones huérfanas, se diseñaron cebadores inversos y directos para los extremos del inserto transducido no repetitivo. Se diseñaron un tercer par de cebadores que flanquean la inserción para amplificar la inserción completa y los alelos de referencia. Cada reacción de PCR contenía 10 ng de ADN, 5 pmol de cada cebador, 5U Taq ADN polimerasa (Sigma-Aldrich, número de catálogo D1806) con tampón 1x que contiene MgCl2, 0,2 mM de cada dNTPs y agua hasta un volumen final de 25 µl. Las condiciones de PCR fueron las siguientes: desnaturalización inicial a 95° C durante 7 minutos; luego 30-35 ciclos de 95° C durante 30 segundos, 60ºC 30 segundos, 72ºC 45 segundos y una extensión final de 72ºC 7 minutos. En algunos casos, cuando la amplificación falló, usamos Platinum Taq High-fidelity, con una desnaturalización de 94ºC y una extensión de 68ºC".

Tras cuantificar el resultado de cada amplificación preparé una librería, tratando de hacer que el número de moléculas de todos los amplicones fuese el mismo, por lo que se ajustaron las concentraciones de cada uno de ellos en función de su peso molecular, de tal forma que al final tuvimos una concentración de 4-50 femtomoles en 64 µl de disolución. A continuación, preparé una librería siguiendo el protocolo de preparación de una librería de amplicones para secuenciación por nanoporos (ver apartado 5.1.12.1.1). La librería se cargó en una celda R9.4 (FLO-MIN106, ONT), y se realizó la secuenciación siguiendo el procedimiento 5.1.12.3. Se obtuvo una cobertura de 8.2x para los amplicones de la muestra NCI-BL2087 y 9.17x para la NCI-H2087.

Debido a que todos los amplicones eran distintos y se alineaban en regiones distintas del genoma, no hizo falta ningún procedimiento informático especial, se usó por tanto la pipeline genérica de alineamiento (ver apartado 5.2.3).

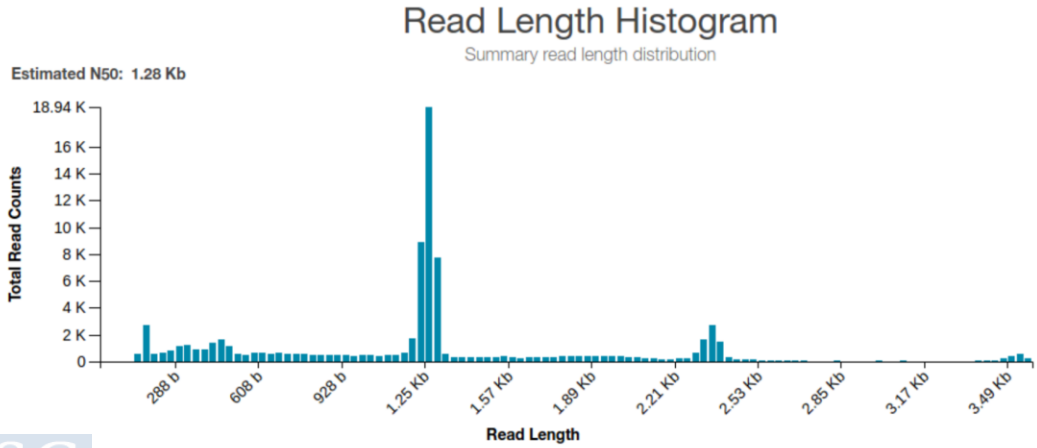


Figura 57: Gráfica producida por el programa MinKnow mostrando el número de lecturas por tamaño. Se pueden apreciar los picos correspondientes a los amplicones.

6.1.1.4 Secuenciación con marcadores secuenciales (*barcoding*)

Este tipo de secuenciación se realizó en varias ocasiones para secuenciar fragmentos muy similares de varias muestras en la misma celda, por lo que fue necesario marcar de forma inequívoca cada uno de los fragmentos de cada una de las muestras. Para ello se usan códigos de barras de ADN (*barcodes).* Estos consisten en pequeños fragmentos de ADN (12-18 pb) con una secuencia única distinta para cada muestra, y son concatenados a las cadenas de ADN a secuenciar.

Algunos de estos experimentos fueron:

- barcoding de amplicones de berberechos, realizado por Daniel García Souto.
- barcoding de amplicones del gen SCN5A implicado en el síndrome de Brugada, realizados por un estudiante del grupo del doctor Brugada, para el cual era necesario conocer la secuencia exacta y el número de copias de un fragmento repetitivo dentro de dicho gen. Tarea que no se podía hacer con las técnicas tradicionales de NGS debido a que los fragmentos cortos que producen no permiten diferenciar unos fragmentos de otros ni su número.
- barcoding del ADN nativo de 12 cepas de levadura a las que se les insertaron unos casetes artificiales. La librería fue llevada a cabo por Vanesa Hurtado de del grupo de Miguel González Blanco bajo mi supervisión, y es la base para el protocolo (5.1.12.1.2 Preparación de una librería con marcadores secuenciales (barcoding) para secuenciación por nanoporos) y es quizá el experimento más representativo de este grupo y se omite aquí al haberse descrito en el mencionado apartado. El procesado de datos se realizó de la siguiente manera. Los códigos de barras (*barcoding*) y adaptadores, se eliminaron de los ficheros .fastq mediante el programa Porechop **(https://github.com/rrwick/Porechop)** y se alineó contra la cepa de referencia S288C: sacCer3_S288C **(https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2/).** Para la detección de mutaciones puntuales (single nucleotide polymorphism, SNP) se usó el software Platypus (Rimmer et al., 2014), Freebayes (Garrison and Marth, 2012) y LongShot (Edge and Bansal, 2019). Tras la filtración de las variantes germinales, se utilizó el programa Variant Effect Predictor (VEP) del EnsembleFungi para la predicción de los efectos de las distintas variantes. (**https://fungi.ensembl.org/Saccharomyces_cerevisiae/Tools/VEP**)

6.1.1.5 Pruebas de fragmentación del ADN

Figura 58: Punta de jeringuilla 30G.

La integridad y longitud de las moléculas de ADN son cruciales para obtener buenas librerías y consecuentemente, buenos datos de secuenciación con secuencias lo más largas posibles, de ahí nuestro interés por conocer qué procedimientos del laboratorio podrían degradarlo. Por otro lado, para ciertos experimentos con ADN limitado y debido al funcionamiento de las celdas de secuenciación de nanoporos, es necesario fragmentar el ADN para obtener más moléculas a costa de su longitud. Por eso diseñamos un experimento para comprobar cómo afectaba a la longitud e integridad del ADN ciertos procedimientos experimentales como: mantenerlo en el frigorífico durante un par de días, el paso del ADN por una jeringuilla muy fina y la

fragmentación con un sistema comercial como el fragmentador G-Tube de Covaris (ver apartado 5.1.9 Fragmentación del ADN).

Para ello se partió de ADN de fago λ proporcionado por ONT (el mismo usado en el experimento 6.1.1.1 Secuenciación del fago λ). Se prepararon varias alícuotas de 10 ng de ADN del mencionado fago y se le aplicaron los procedimientos indicados en la tabla mostrada a continuación (tabla 5)

Muestra	Procedimiento
0	Escalera de alto peso molecular. (No se le realizó ningún procedimiento)
1	Fago λ, fragmentación con G-Tube (estándar), preparado dos días antes y mantenido en el frigorífico a 4º C (protocolo (5.1.9 Fragmentación del ADN))
2	Fago λ íntegro, descongelado dos días antes y mantenido en el frigorífico a 4º C
3	Fago λ, fragmentación con G-Tube (estándar), recién descongelado y procesado
4	Fago λ, íntegro, recién descongelado.
5	Fago λ, recién descongelado, fragmentación pasando 5µl muestra por una punta de jeringuilla 30G 10 veces (Punta similar a la de la figura superior)

Tabla 6 Pruebas de fragmentación/degradación del ADN

Las muestras se cargaron en un gel de agarosa al 0.8% y se corrieron a 75 voltios durante 15’ minutos, observándose lo siguiente

La permanencia del ADN en el frigorífico no lo degrada de forma apreciable, siendo muy similares entre ellas las bandas 2 y 4 (el fago λ íntegro) y las bandas 1 y 3 el fago λ fragmentado por el Covaris G-Tube.

1. La fragmentación por el Covaris G-Tube, al contrario de lo esperado no produjo una banda correspondiente al tamaño de fragmento de aproximadamente 10K pb, sino una sucesión de fragmentos de distintas longitudes que en el gel se observaba como mancha o frotis (“*smear*”), desde los fragmentos de alto peso molecular hasta fragmentos más pequeños del tamaño de corte deseado.
2. La fragmentación por la jeringuilla no producía una serie de bandas diferenciadas correspondientes a fragmentos de varios tamaños, sino no una sola banda correspondiente al fago λ integro con un poco menos de intensidad y un espectro continuo de fragmentos pequeños (“*smear*”).

Se decidió por tanto hacer un experimento específico para entender mejor el funcionamiento de los tubos de fragmentación Covaris G-Tube.

USC DE SANTIAGO DE COMPOSTELA

6.1.1.6 Pruebas de fragmentación Covaris G-Tube.

El objetivo de este experimento era doble, por un lado, entender mejor cómo la variación de la velocidad y tiempo de centrifugación y el número de pases afectaba al tamaño y "limpieza" del corte, y por otro lado adaptar el protocolo de fragmentación a los volúmenes de las librerías de MinION, donde se parte de entre 1-5 µg de ADN en un volumen de 46 µl. El protocolo del G-Tube está diseñado para alícuotas de hasta 30 µ de ADN, pero en 150 µl de disolución. Esto es un inconveniente, pues obliga a preparar librería en exceso (3 veces más de librería), o concentrar el ADN tras el corte.

Para este experimento, realizado con la colaboración de Daniel García Souto, se partió de alícuotas (10) de 2.3 µg de ADN del fago λ (medido en el Qbit) en 46 µl de agua libre de nucleasas.

Se procedió a cortar el ADN usando el protocolo (5.1.9 Fragmentación del ADN), con los tiempos, velocidades y número de pasos especificados en la tabla adjunta.

	Color	Velocidad (RPM)	Tiempo (seg)	Número de pases	Pico (pb)	DIN
1	Naranja	Ladder			varios	
1	Rojo oscuro	6,000	60	1	13,677	7.4
1	Rojo claro	6,000	60	2	14,491	7.5
1	Azul claro	7,200	60	2	14,665	7.7
1	Violeta claro	11,000	30	1	15,061	7.4
1	Violeta oscuro	11,000	30	2	15,081	7.5
1	Azul oscuro	7,200	60	1	15,506	7.5
2	Verde	Mezcla 7,200+4,200	60	1	16,129	7.7
1	Amarillo	4,200	60	1	16,481	7.5

Tabla 7: Pruebas de fragmentación con el Covaris G-Tube

Las muestras resultantes con el ADN cortado se analizaron utilizando una Tape Station en lugar de un gel de agarosa. Se esperaba que el tiempo de fragmentación, la velocidad de rotación de la centrífuga, y el número de pases afectase significativamente al tamaño de los fragmentos obtenidos tal como indica el fabricante del g-Tube (Covaris). Ver figura adjunta donde se fragmenta 150 µl de fago λ 1) 6k pb, 2) 8 Kpb, 3) 10 Kpb, 4) 20 Kpb.

USC DE SANTIAGO DE COMPOSTELA

Los resultados obtenidos por nosotros son que la fragmentación con el Covaris G-Tube produce un espectro continuo de fragmentos de diferentes tamaños, sin demasiada correlación con las condiciones del experimento para cada una de las alícuotas, tal como se puede observar en la figura de la Tape Station. Tal vez la falta de correlación entre los parámetros y el tamaño de corte tenga que ver con el bajo volumen empleado, 46 µl, en vez de los 150 µl indicados por el fabricante.

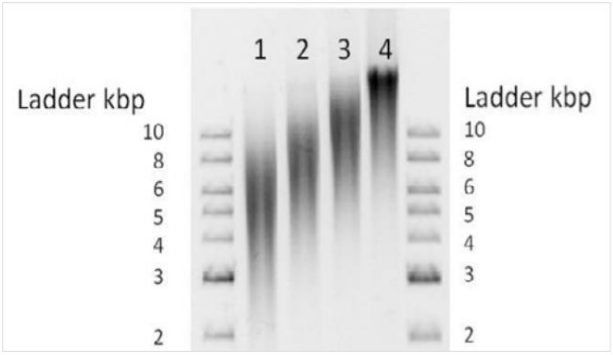


Figura 59: Fragmentación del ADN con el Covaris G-Tube. (Cortesía de Covaris)

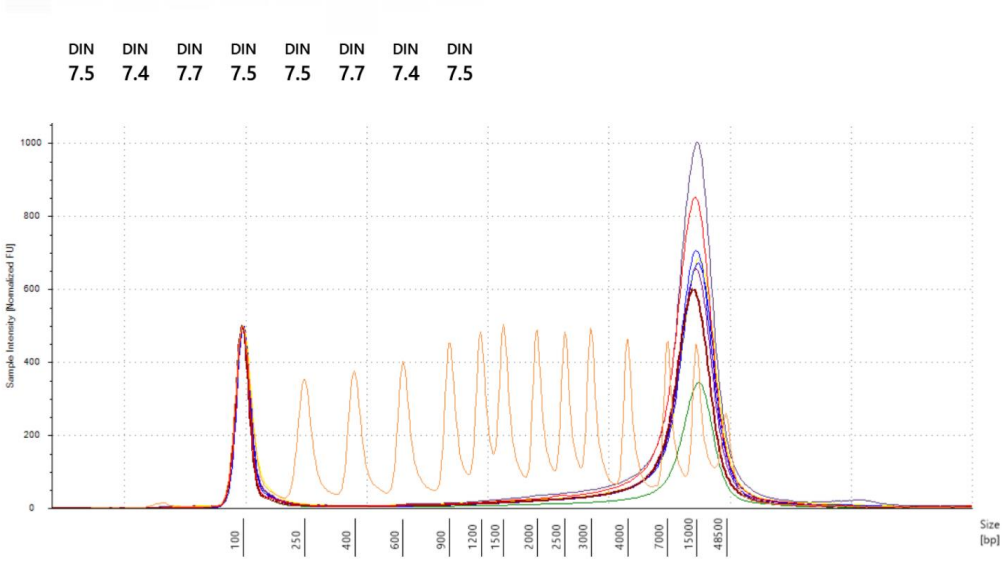


Figura 60: Salida de la Tape Station de las pruebas de fragmentación con el Covaris G-Tube.

6.1.1.7 Selección de ADN de cadenas largas

Como se ha comentado en varias ocasiones a lo largo del presente trabajo, el disponer de ADN íntegro y con cadenas largas (del orden de decenas de miles de pares de bases) era un factor clave para el éxito de este proyecto. Aunque actualmente en el laboratorio usamos un sistema altamente eficaz de selección de ADN de alto peso molecular (5.1.10 Selección de ADN de cadena larga (Circulomics SRE)), al comienzo de esta Tesis mismo no existía ninguna solución estandarizada para la obtención del mismo, por lo que fue necesario realizar una serie de experimentos para obtener ADN de longitud óptima. Estas pruebas y experimentos se describen a continuación.

6.1.1.7.1 Pruebas con membranas (o tripas) de diálisis.

La diálisis es un proceso físico, mediante el cual se separan moléculas de diferentes tamaños en una solución utilizando una membrana semipermeable. Las tripas de diálisis son tubos de membrana semipermeable utilizados para separar, en una solución, las moléculas

pequeñas de las grandes, pues las primeras son capaces de difundir a través de la membrana debido a su bajo peso molecular. Esta es la lógica subyacente para intentar separar y concentrar las cadenas de ADN de gran longitud, permitiendo que las cadenas de bajo peso molecular difundan a través de la tripa de diálisis (Krerowicz, Hernandez-Ortiz, y Schwartz 2019).

Como prueba de concepto y para comprobar la viabilidad de esta aproximación diseñamos un experimento donde usamos una tripa de diálisis SnakeSkin™, de Termo Fisher.

(https://www.thermofisher.com/order/catalog/product/88243 accedida 22.03.2023) con un MWCO (punto de corte de peso molecular) de 10KDa, lo que equivale a una molécula de DNA de 270 nucleótidos. **(https://www.genscript.com/conversion.html accedida 22.03.2023)**

Los fragmentos de este tamaño existen en gran cantidad incluso en ADN con muy buena integridad (figura 61) y en los primeros experimentos de secuenciación constituían una fracción importante del ADN secuenciado, por lo que su eliminación era fundamental para secuenciar los fragmentos de gran longitud que necesitábamos para este trabajo.

Figura 61: Fragmentos de ADN de pequeño tamaño que no interesa secuenciar.

Para llevar a cabo el experimento se disolvieron 250 ng de ADN de fago λ (5 µl) en un volumen aproximado de 9 ml de tampón de electroforesis TBE 1x y se introdujeron en la tipa de diálisis, previamente auto clavada en TBE 1x durante un par de horas (figura 62). Esta se dispuso en una cubeta de electroforesis con el mismo tampón tal como se muestran en las imágenes siguientes. Para facilitar la difusión se estableció una diferencia de potencial de 100V y se permitió que difundiera durante otro par de horas.

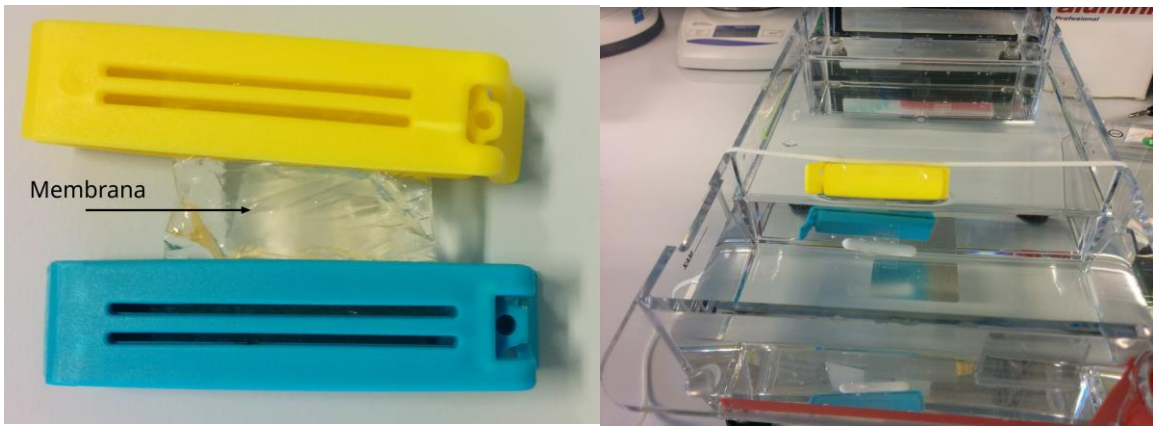


Figura 62: Pruebas de eliminación de fragmentos de ADN de reducida longitud mediante membranas de diálisis.

El experimento resultó un fracaso, puesto que no encontramos ADN en la disolución dentro de la tripa de diálisis, quedando este completamente adherido a la membrana, tal como pudimos comprobar al fotografiar la membrana en el transiluminador. Se hicieron pruebas para intentar despegarlo, poniendo la membrana en disoluciones con diferentes concentraciones de sales, pero resultó completamente imposible.

Probablemente el problema fue el haber usado membranas de celulosa en vez de membranas de policarbonato PDMS, como en los experimentos realizados por otros autores (Krerowicz, Hernandez-Ortiz, y Schwartz 2019), o tal vez el tiempo de autoclavado para eliminar el recubrimiento de conservación de las membranas no fue suficiente.

6.1.1.7.2 Recorte de bandas

Un procedimiento habitual para obtener fragmentos de ADN del tamaño deseado es el de migrar la muestra de ADN en un gel de agarosa, una vez separados los fragmentos por tamaños recortar la banda la banda de interés y extraer el ADN del gel de agarosa, normalmente usando un kit comercial.

Para ver si ese procedimiento era adecuado a nuestras necesidades, se llevaron a cabo una serie de experimentos descritos en los siguientes apartados. Los aspectos que nos interesaba comprobar eran los requerimientos de ADN (muchas veces un factor limitante debido a su escasez), el rango de tamaños obtenidos y la sencillez y velocidad del proceso (pues era necesario realizar cientos de librerías).

6.1.1.7.3 Gel de agarosa, migración en cubeta de electroforesis standard y extracción de ADN del gel

Para tratar de obtener ADN con fragmentos de gran tamaño se siguió el siguiente procedimiento: Se creó un gel de agarosa al 0.8% (TBE 1x), con pocillos de tamaño estándar. Se introdujo 1µg de ADN del fago λ (20µl) en uno de los pocillos. Tras correr 20 minutos a un voltaje de 120 Voltios se visualizó el gel en el transiluminador con el programa GelDoc (figura 63, izquierda). Las bandas de alto peso molecular se recortaron (figura 63, derecha) y se usó el kit de extracción de ADN de geles de agarosa Quiagen "QIAquick Gel Extraction Kit" siguiendo las recomendaciones del fabricante. Se usó 1.1 µl de buffer, incubándolo 10 minutos a 50º C. **(https://www.qiagen.com/us/products/discovery-and-translational-research/dna-rna-purification/dna-purification/dna-clean-up/qiaquick-gel-extraction-kit?catno=28704 , accedido 27.03.2023)**

Tras cuantificar (Qbit) el ADN obtenido, descubrimos que no habíamos recuperado nada de ADN

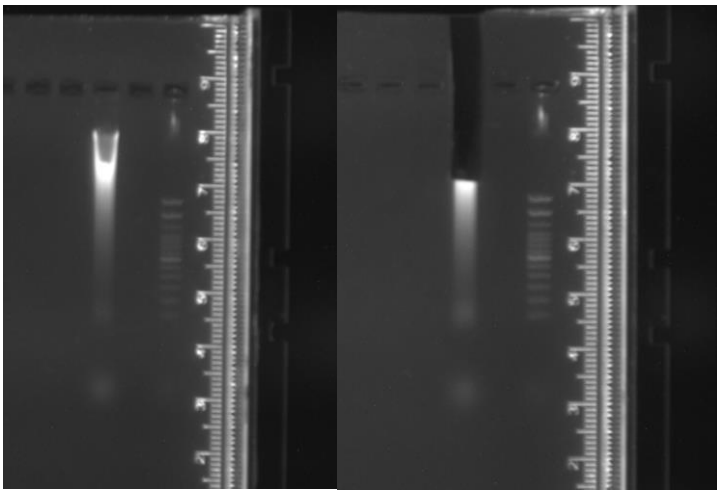


Figura 63: Recorte de bandas I.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Se repitió el experimento, doblando la cantidad de ADN (2µg en 40 µl), haciendo los pocillos tres veces más grandes (el resto de los parámetros se mantuvieron igual). Tras correr el gel (Figura 64, izquierda) las bandas eran mucho más anchas. Tras recortarla (figura 64, derecha) y extraer el ADN con el kit anteriormente mencionado. Los resultados fueron igualmente nulos.

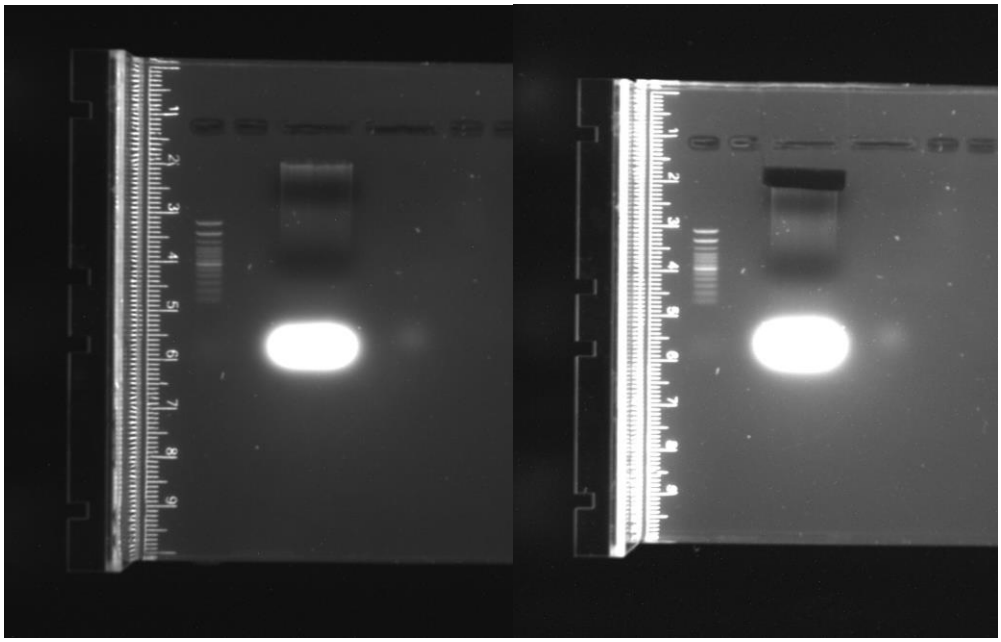
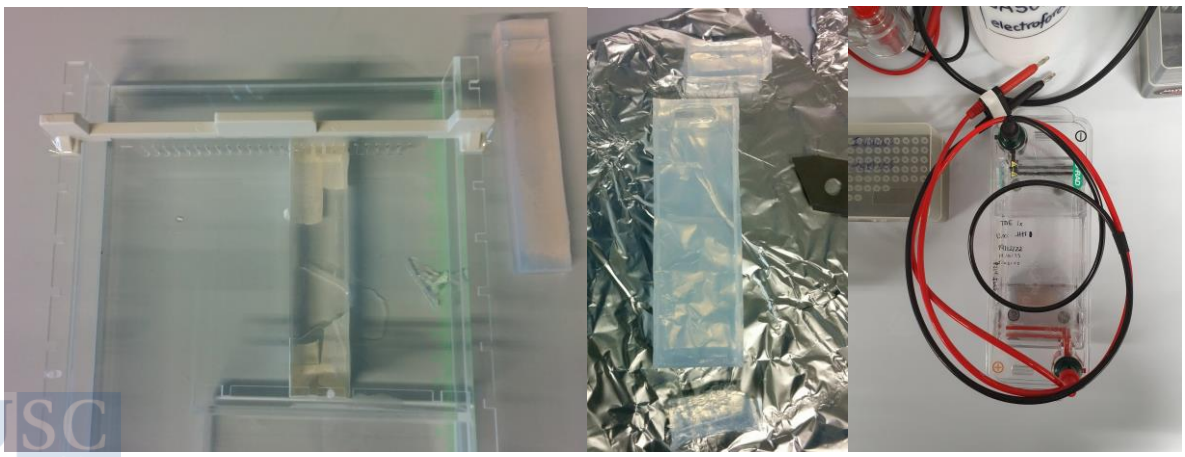


Figura 64: Recorte de bandas II.

6.1.1.7.4 Gel de agarosa de grandes dimensiones y cubeta de electroforesis pequeña

Queriendo aumentar más la cantidad de ADN a introducir en los pocillos, decidí hacer un gel de grandes dimensiones y con pocillos gigantes. Para ello se adaptó la cubeta tradicional de creación de geles para producir un gel de agarosa de 8 x 2 x 1 cm (figura 65 izquierda y centro)



USC UNIVERSIDADE DE SANTIAGO

Figura 65: Creación de un gel de agarosa de grandes dimensiones.

Se introdujo en el pocillo 5 µg (100 µl) de ADN del fago λ. El gel se migró en una cubeta de electroforesis Mini-Sub Cell GT Cell de Bio-rad (figura 65, derecha), durante una hora a un voltaje de 60 Voltios.

Tras su visualización con UV se comprueba que, a pesar de la alta cantidad de ADN de partida, queda muy poca cantidad dentro del gel, y desde luego no forma las esperadas bandas características cuando el ADN se separa por sus diferentes pesos moleculares. El ADN quedó disuelto en el buffer de la cubeta tras difundirse por los laterales del gel.

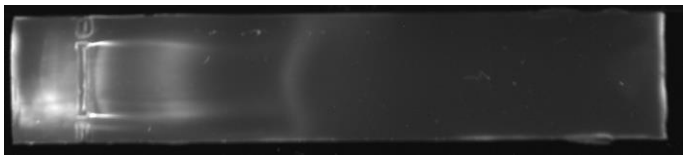


Figura 66: Gel de agarosa de grandes dimensiones en el transiluminador, mostrando la ausencia de ADN.

Decidí entonces hacer una cubeta de electroforesis con las dimensiones exactas del gel de electroforesis (8x2x1 cm) para evitar que el DNA difundiera hacia los lados.

6.1.1.7.5 Gel de agarosa de grandes dimensiones y cubeta de electroforesis de construcción propia

Con el objetivo de migrar un gel de grandes dimensiones sin que el ADN difundiese por los lados construí una cubeta de electroforesis, tal como se describe en (6.1.3 Diseño y construcción de cubeta de electroforesis de reducidas dimensiones).

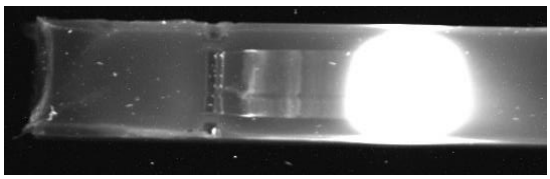


Figura 67: Gel de agarosa de grandes dimensiones en el transiluminador, con ADN migrado en la cubeta de lectroforesis de construcción propia.

Tras preparar otro gel de agarosa 0,8% + TAE 1X de grandes dimensiones (8x2x1 cm) procedimos a cargar otros 5 µg (100 µl) de ADN del fago λ. Debido a sus pequeñas dimensiones esta cubeta, corría con 40 Voltios.
Tras media hora se puede ver que el ADN sigue dentro del gel y forma bandas, con el ADN separado por su peso molecular. Se recortó el fragmento de gel correspondiente al ADN de alto peso molecular y se procedió a extraer el ADN con el anteriormente mencionado kit de extracción de ADN de geles de agarosa Quiagen "QIAquick Gel Extraction Kit", siguiendo las instrucciones del fabricante. Una vez más el resultado fue nulo.

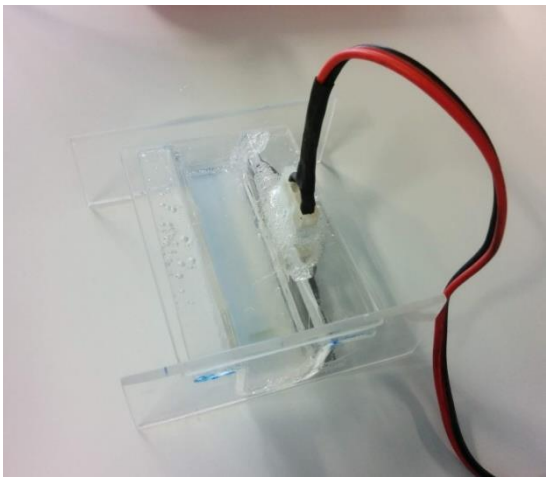


Figura 68: Cubeta de electroforesis de construcción propia.

A posteriori se me ocurrió que hubiésemos podido probar a migrar el gel, recortarlo descartando los fragmentos de bajo peso molecular y migrarlo de nuevo con la corriente invertida de tal forma que solo el ADN

de alto peso molecular volviese al pocillo. El problema una vez más hubiese sido el conseguir que el ADN quedase en él y no migrase pasado el pocillo tal como se describe en el apartado 6.1.1.7.7 Gel de doble densidad y cubetas estándar con inversión de la corriente (GD-Piping) o tal vez al ser más grande el pocillo hubiésemos conseguido extraerlo.

6.1.1.7.6 Conclusiones del recorte de bandas y la extracción del ADN de los geles de agarosa

Analizando detenidamente cuál pudo ser el problema de que a pesar de tener grandes cantidades de ADN en el fragmento de gel no consiguiéramos extraerlo del mismo, nos dimos cuenta de que el kit está diseñado para extraer ADN de geles de agarosa en el rango de 70 - 10K pb. La banda que estábamos recortando era de un alto peso molecular (fago λ), con una longitud mucho mayor que los límites de extracción del kit. El sesgo cognitivo provino de que este kit se usaba en nuestro laboratorio de forma habitual, pero con fragmentos de ADN con un tamaño considerablemente menor a 10K pb, del orden de cientos de pares de base a mil pares de bases utilizados para secuenciar con Illumina, Sanger capilar.

Más tarde descubrimos que existían kits comerciales de extracción de ADN de con límites de tamaño mayores, pero ya estábamos inmersos en la búsqueda de una solución que no implicase la extracción del gel de agarosa. Esos esfuerzos se muestran en los siguientes apartados.

6.1.1.7.7 Gel de agarosa de doble densidad y cubetas estándar con inversión de la corriente (GD-Piping)

Siguiendo con el planteamiento de no tener que extraer el ADN del gel de agarosa con el kit "Agarose Gel DNA Extraction Kit", diseñamos un procedimiento en el que se migraba ADN en un gel de agarosa hasta que el ADN se separaba por peso molecular, momento en el cual se cortaba del gel el ADN de bajo peso molecular, se invertía la corriente de tal forma que el ADN de alto peso molecular regresase a los pocillos. Para evitar que el ADN se pasase de los pocillos se diseñó un gel con dos densidades tal como se muestra en el esquema. Uno con agarosa al 0.8% adecuado para la separación el ADN y otro con agarosa al 5 % para frenar el ADN cuando este migraba de nuevo hacia los pocillos.

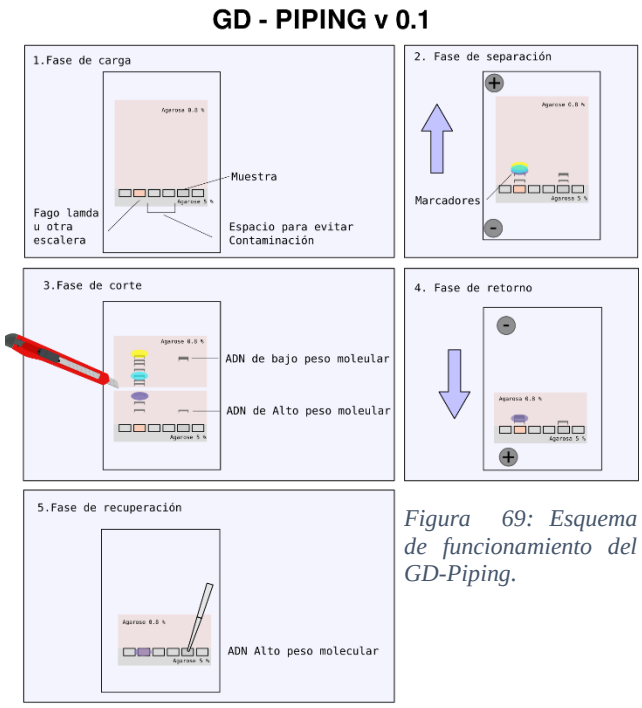


Figura 69: Esquema de funcionamiento del GD-Piping.

El funcionamiento del sistema se describe detalladamente a continuación (figura 69).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1. Fase de carga: en un gel de grandes dimensiones se crean pocillos de grandes dimensiones para contener volúmenes grandes de ADN. Estos pocillos están en el borde entre los geles de distintas densidades, de tal forma que cuando la corriente es directa el ADN pueda migrar y separarse, y cuando la corriente es inversa el ADN quede frenado en los pocillos.
2. Fase de separación: El gel se migra de tal forma que el ADN se separa en función de su peso molecular. El frente de bajo peso molecular al migrar más rápido es el que más se separa.
3. Fase de corte: En vez de recortar una banda y extraer el ADN de interés, en este sistema se recorta todo el gel que contiene el ADN de bajo peso molecular y que no nos interesa. Se descarta ese fragmento del gel.
4. Fase de retorno: Se invierte la polaridad en la cubeta de electroforesis de tal forma que el ADN vuelva a los pocillos. El gel de 5% de agarosa dificulta (en teoría) que una vez el ADN ha regresado a los pocillos, este se vuelva a introducir en el gel
5. Fase de recuperación: Se recupera de los pocillos el ADN de alto peso molecular, que debería corresponder al ADN de cadenas de gran longitud.

Se hicieron pruebas de funcionamiento, como por ejemplo el mostrado en la figura 70 donde se introdujo en uno de los pocillos el marcador de peso molecular y en otro azul de bromofenol para marcar el avance. El gel se corrió a 120 V 20 minutos, se paró, procedió a cortar el gel, y se invirtió la corriente con el mismo voltaje durante 10'.

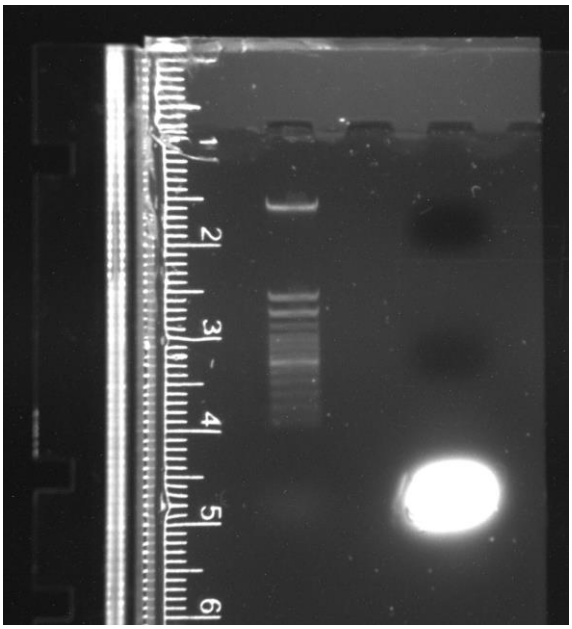


Figura 70: Gel de agarosa del sistema GD-Piping.

Tras recoger la disolución y cuantificarla con el Qbit, no encontramos ADN, que se había ya introducido en el gel de agarosa de 5%, que a pesar de su alta densidad no impidió su penetración.

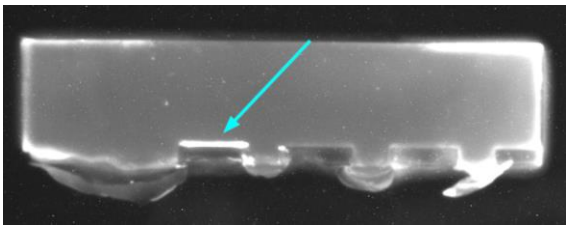


Figura 71: Gel de agarosa al 5% con ADN. (señalado por la flecha)

Debido a la adquisición de un sistema comercial el BluePiping, decidí abandonar este tipo de experimentos y centrarme en él, tal como se describe en el siguiente apartado.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.1.1.7.8 BluePiping.

Nuestro grupo de investigación adquirió la máquina BluePiping para intentar seleccionar fragmentos de ADN de gran tamaño. Este dispositivo de la empresa Sage Science separa los fragmentos de ADN por medio de un campo pulsátil. Su funcionamiento se basa en migrar una muestra de ADN en un gel de agarosa mediante electroforesis. Cuando el dispositivo detecta que el ADN tiene el tamaño deseado (gracias a un marcador que migra con las muestras) lo redirige a una cámara donde se puede recoger (electroelución).

Figura 72: BluePiping.

Alguna de sus características destacadas es que puede crear campos de 25V, 100V, 150V pulsados o constantes. Para la detección óptica del ADN usa luz de 470 nm (excitación), 525 nm (emisión).

Tras varias pruebas, (no mostradas por concisión) se desestimó su uso posterior debido a que se necesitaban cantidades de ADN imposibles de obtener de un tejido primario (sin replicación del ADN). Según nuestras estimaciones del orden de 16 a 32 µg de ADN para obtener 1µ de ADN de alto peso molecular.

6.1.1.7.9 Gel de agarosa con pocillos intermedios y cubeta de electroforesis standard

Para evitar el problema de tener que recortar la banda y extraer el ADN del gel de agarosa, se diseñó otro gel con una serie de pocillos intermedios. El funcionamiento esperado era el siguiente: al migrar el ADN en el gel, el ADN de bajo peso molecular pasaría por el segundo grupo de pocillos, entraría en ellos por un lado y saldría por el otro entrando de nuevo en el gel. Mientras el ADN de alto peso molecular seguiría migrando por el gel, hasta llegar al segundo grupo de pocillos. Una vez dentro se pararía de migrar el gel, pudiéndose recuperar el ADN en ellos. El problema con esta aproximación es que era extremadamente difícil el saber cuándo el ADN con el peso molecular de interés se encontraba en los pocillos. Como puede comprobarse en la figura 73 todo el ADN ha pasado ya el segundo grupo de pocillos y se encuentra **dentro** del gel, por lo que ya no se pudo recuperar. Curiosamente este es el principio de funcionamiento del dispositivo comercial, E-Gel de la casa Invitrogen, la cual consiste en un dispositivo a la que se le pueden acoplar unos cartuchos que contienen geles de agarosa prefabricados (para hacer electroforesis) y un transiluminador, todo en uno. Existen diferentes tipos de cartuchos, un

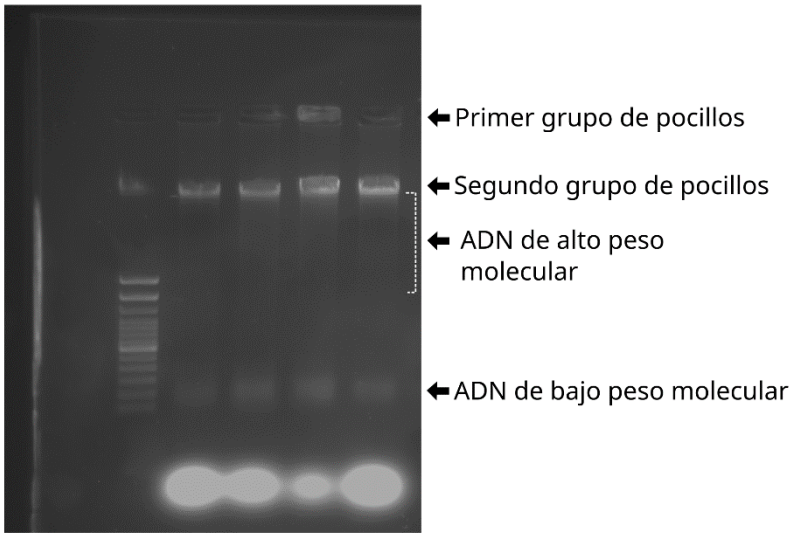


Figura 73: Gel de agarosa con dos grupos pocillos.

tipo de ellos permite la selección de ADN de ciertos pesos moleculares y consisten en dos juegos de pocillos similares a los de nuestro diseño. Al poderse activar el transiluminador al mismo tiempo que se realiza la electroforesis, se puede observar en tiempo real el paso del ADN a través del segundo grupo de pocillos los pocillos, lo que en teoría permitiría seleccionar el ADN del peso molecular de interés. Nuestro laboratorio dispuso de esa máquina para su evaluación, que describimos en el apartado siguiente.

6.1.1.7.10 Invitrogen E-gel y cartuchos de selección de tamaño

Figura 74: Invitrogen e-gel.

El E-Gel de la casa Invitrogen es un dispositivo de electroforesis que utiliza geles de agarosa sin tampón dentro de unos prácticos cartuchos intercambiables. Estos cartuchos llevan incorporados los electrodos, un sistema de generación de iones, un par de matrices de intercambio de iones, un tampón de pH, y tinción SYBR Safe. El flujo de iones a través de la matriz de intercambio genera un campo eléctrico constante que produce la migración del ADN. Dispone además de un transiluminador de luz ultravioleta, que permite la visualización en tiempo real de la migración del ADN. Es fácil de usar, seguro, y tiene un alto rendimiento. Existen varios tipos de cartuchos con geles de agarosa de distinta densidad.

La empresa también fabrica unos cartuchos con un doble juego de pocillos, uno de ellos a mitad del gel (CloneWell 0.8%), lo que permite la recuperación del ADN. Por ello decidimos usarlo para seleccionar ADN de alto peso molecular adecuado para la preparación de librerías de MinION.

Para el primer experimento de selección de ADN por tamaños utilizamos ADN de berberecho, del cual disponíamos en abundancia, puesto que estábamos secuenciándolo para generar el primer genoma de referencia de la especie. Partimos de una muestra con una concentración de 76.8 ng/µl.

Siguiendo las recomendaciones del fabricante

(https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2FMAN0017342_egel_clonewell2_QR.pdf) (accedida 15.06.2018) se introdujo en todos los pocillos 10 µl de ADN (768 ng) de muestra (procesada según la tablas a continuación) + 9,5 µl de agua libre de nucleasas + 2.5 µl tampón de carga 10X

Muestras	
1,2,3	768 ng ADN cortado con el Covaris G-Tube a 14.000 rpm, una (1) pasada
4	Escalera 1Kb plus DNA (25 µl)
5,6,7	768 ng ADN cortado con el Covaris G-Tube a 14.000 rpm, dos (2) pasadas

Tabla 8: Pruebas de selección de ADN con Invitrogen e-gel

Migramos el gel durante 27 minutos. A partir de aquí introducíamos la cantidad de agua indicada en la 1ª columna, migrábamos el gel el tiempo indicado en la 2ª columna y

recuperábamos el ADN de los 6 pocillos (un volumen aproximado de 10 µl por cada uno), obteniéndose el total indicado en la última columna (medido con Qbit)

agua libre de nucleasas (µl)	Tiempo (min.) migrando el gel	comentario	ADN (ng) recuperado
40	1.5	Sale muy poca cantidad, rellenamos con 80 µl y extraemos el ADN	519.6
20	2		39.96
0	2	Recogemos el ADN sin rellenar	346.8
40	2		192
40	2		123.6
40	2		69.6
Total			**1291.56**

Tabla 9: ADN recuperado con el dispositivo Invitrogen e-gel

Lo que constituye una recuperación el 28.02 % del ADN introducido.

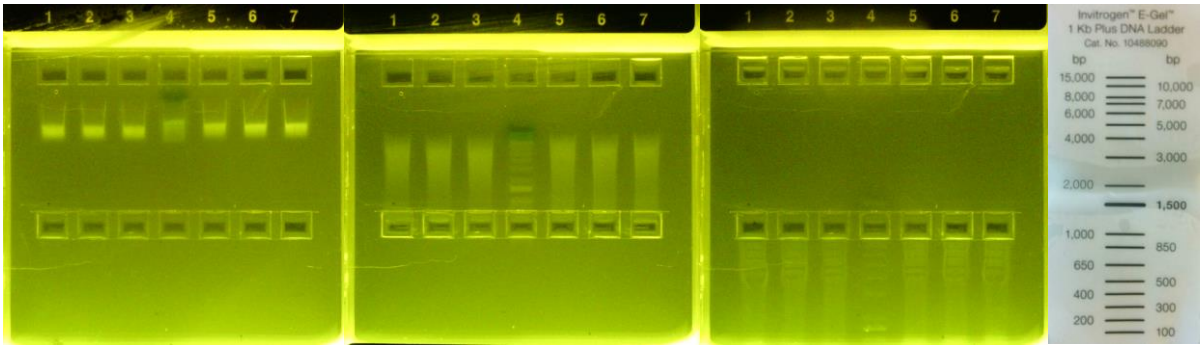


Figura 75: Migración del ADN en el dispositivo invitrogen e-gel, junto con la escalera de ADN para referencia.

Para intentar mejorar el porcentaje de recuperación, decidimos introducir más ADN del berberecho usado en el experimento anterior (20µl a 76.8 ng/µl) a pesar de estar por encima de las recomendaciones del fabricante.

Muestras	
1,2,3,5,6	1536 ng (20µl) ADN cortado con el G-Tube a 14.000 rpm, una (1) pasada
4	Escalera eXact Gene 24 Kb (25 µl)
7	1000 ng (20µl) ADN de fago λ (íntegro sin cortar)

Tabla 10: Recuperación de ADN en el e-gel

Migramos el gel durante 27.5 minutos. A partir de este momento, migramos el gel 30 segundos, pero esperamos un minuto y medio (1.5) para recoger la muestra del pocillo, tras lo

cual rellenamos con 40 µl de agua libre de nucleasas y volvemos a migrar el gel otros 30 segundos y así sucesivamente durante 14 minutos

Tras esos 14 minutos, seguimos corriendo el gel otros 10 minutos para ver si éramos capaces de extraer el ADN del fago λ, que al ser de gran tamaño > 24K pb tardaba más en migrar hasta el pocillo de recuperación.

Agua libre de nucleasas (µl)	Tiempo (seg) migran-do el gel	Tiempo (seg) espera para recoger muestra	Comentario	Volumen de muestra recuperado 5 pocillos (µl)	ADN (ng) recuperado
40	30	90		550	270.6
0	30	90	En el paso anterior me olvido rellenar, por lo que en este paso no obtenemos muestra		0
40	30	90		500	279
40	30	90		500	162
40	30	90		750	150
40	30	90		600	92.4
40	30	90		750	114
Total					**1068**

Tabla 11: Segunda prueba de recuperación de ADN con el Invitrogen e-gel

Se recuperó el 13.90 % del ADN introducido. El porcentaje de recuperación es menor que en el experimento anterior quizás porque habíamos superado las recomendaciones del fabricante respecto a la cantidad de ADN de partida en cada pocillo, o porque al migrar el gel 7 veces 30 segundos, en vez de 6 veces 2 minutos (experimento anterior) recogíamos una franja más estrecha de ADN.

Se puede observar en la figura siguiente como el ADN del fago λ es de mayor integridad y peso molecular que el del berberecho.

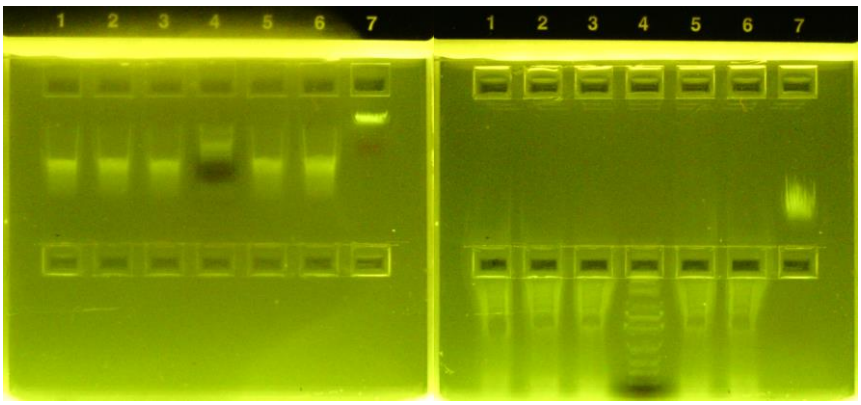


Figura 76: Segunda prueba de migración de ADN en el invitrogen e-gel.

Para comprobar el tamaño e integridad de los fragmentos de ADN obtenidos en los distintos momentos del tiempo, se migró el ADN recuperado de los pocillos intermedios del Invitrogen en

un gel de agarosa del 0.8%. Para ello se introdujo en cada pocillo del gel de agarosa 50 µl de muestra y 5µl de tampón de carga. Se migró a 100V 15, 35, 72 minutos. A los 72 minutos las bandas eran demasiado tenues que no se distinguían por lo que se muestra la imagen a los 35 minutos.

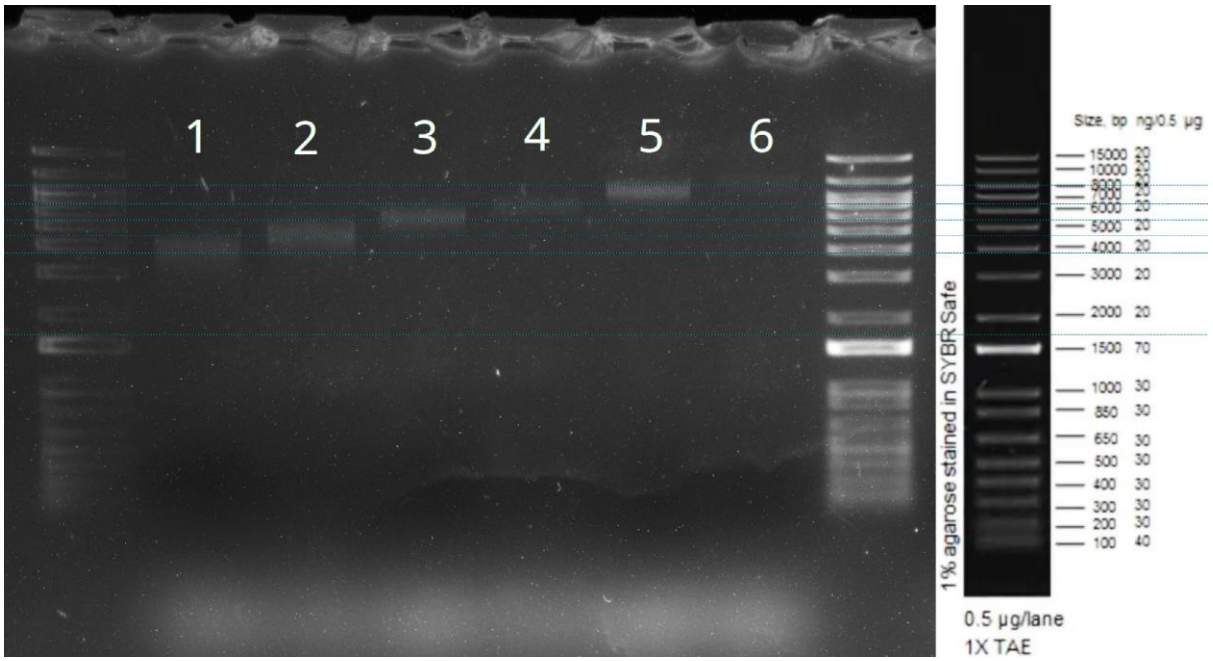


Figura 77: Gel de agarosa con el ADN recuperado del invitrogen e-gel.

Como se puede comprobar las bandas contienen fragmentos con rangos de tamaños muy estrechos, que van desde los 4kb a 8kb muy acordes a la integridad del ADN del berberecho. Por lo que este sistema en principio es adecuado para la recuperación de ADN de pesos moleculares medios y la limpieza de los fragmentos pequeños. Tiene las desventajas de que es un proceso largo, tedioso y propenso a errores, lo que hace su aplicación no sea adecuado para la generación de las grandes librerías requeridas para los experimentos realizados por nuestro grupo de investigación en general y las de este trabajo de Tesis en particular.

6.1.1.7.11 Limpieza ADN con magnetic beads y buffer de limpieza de ONT

La purificación basada en bolitas magnéticas (magnetic beads, xbeads) es un método ampliamente utilizado en biología molecular para aislar ADN de alta calidad de diversas fuentes, incluidos tejidos animales, materiales vegetales y cultivos microbianos (Schalamun et al. 2019). Este método es muy eficiente y puede proporcionar ADN altamente purificado con mínima contaminación, lo que lo hace ideal para su uso en secuenciación, PCR y otras técnicas de biología molecular.

De hecho, el protocolo de preparación de las librerías de ONT (5.1.12.1) hace uso extensivo de ellas. Por ello diseñamos el siguiente experimento para entender mejor su funcionamiento, incluyendo la proporción óptima de bolitas magnéticas, si eran igual de efectivas las bolitas magnéticas de un recipiente recién estrenado, que las de un recipiente

abierto varias semanas. También cómo afectaba el proceso a ADN que ya tenía gran integridad.

Se seleccionó ADN de berberecho, cuya integridad no era demasiado buena (DIN 6-7) y se limpió usando bolitas magnéticas de un recipiente recién abierto (N), de otro de varias semanas abierto (V) y usando proporciones de 1x o 0.4x. El procedimiento es el mismo que se usa en la preparación de las librerías MinION, y básicamente consiste en los siguiente:

- Volver a suspender las bolitas magnéticas AMPure XP mediante agitación vorticial.
- Agregar **1x o 0.4x** de bolitas magnéticas AMPure XP resuspendidas a la reacción y mezclar sacudiendo el tubo.
- Incubar en un mezclador Hula (mezclador rotatorio) durante 5 minutos a temperatura ambiente.
- Hacer descender la muestra al fondo del eppendorf y volver a colocar el tubo en el imán, para que se forme un pellet. Mantenga el tubo en el imán y pipetear el sobrenadante.
- Lave las bolitas magnéticas agregando 250 µl de ABB (Adapter binding buffer). Mover las bolitas magnéticas para volver a suspenderlas, y hacerlas descender en el recipiente, luego devuelva el tubo a la gradilla magnética y deje que las bolitas magnéticas sedimenten formando un pellet. Retirar el sobrenadante con una pipeta y desechar.
- Repita el paso anterior.
- Hacer descender la muestra al fondo del eppendorf y volver a colocar el tubo en el imán, para que se forme un pellet. Pipetear cualquier sobrenadante residual. Deje secar ~30 segundos, pero sin permitir que seque el pellet hasta el punto de agrietarse.
- Retirar el tubo de la gradilla magnética y volver a suspender el sedimento en 61 µl de agua libre de nucleasas.
- Incubar durante 2 minutos a temperatura ambiente.
- Precipitar las bolitas magnéticas en un imán, para formar un pellet hasta que el eluido sea transparente e incoloro, (al menos 1 minuto).
- Retirar y conservar 61 µl del eluido en un tubo Eppendorf DNA LoBind limpio de 1,5 ml.
- Cuantificar 1µ l de muestra eluida utilizando un fluorímetro Qubit.

Nota: El ABB (Adapter binding buffer) formaba parte de los reactivos de las primeras versiones del kit de preparación de librerías de ONT. Posteriormente fue sustituido por dos tampones para la preparación de librarías enriquecidas en fragmentos largos (LFB) o también incluyendo fragmentos cortos (SFB). La composición exacta es desconocida al ser la fórmula propietaria de ONT, pero diversas fuentes le atribuyen composiciones similares a "10mM tris-HCL, 1mM EDTA pH 8.0, 1.6 M NaCl, 11% PEG 8000". o 10% PEG 8000, 500 mM NaCl, 10 mM Tris-HCl pH 8.0". En todo caso diversas concentraciones de PEG Tris-HCL y NaCl. **https://international.neb.com/protocols/2021/10/18/protocol-for-uhmw-dna-cleanup-in-the-oxford-nanopore-technologies-ul-library-prep-workflow** (accedido en abril 2023)

Las muestras comparadas fueron, ADN del fago λ (de 48k pb) fragmentado usando un tubo G-Tube de Covaris, a 14.400, durante 60 segundos. En otra muestra se fragmentó el ADN del fago λ agitándolo violentamente en un vortex durante 30 segundos (eppendorf 0.2 ml). También se limpió el ADN del fago λ íntegro con bolas magnéticas al 0.4x. y se dejó una muestra de ADN del fago λ íntegro.

Tras los tratamientos se preparó un gel de agarosa al 0.8% al TAE 1x con las muestras según la tabla a continuación.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Muestra	Tratamiento
1	Escalera 1K plus DNA
2	m1: ADN berberecho, limpieza con xbeads 1x (recién abiertas)
3	m2: ADN berberecho, limpieza con xbeads 0.4x (recién abiertas)
4	m3: ADN berberecho, limpieza con xbeads 1x (no recientes)
5	m4: ADN berberecho, limpieza con xbeads 0.4x (no recientes)
6	Fago λ, fragmentado en g-Tube 14.400 rpm, 60 seg
7	Fago λ, vortex 30 seg. (en un eppendorf de 0.2 ml)
8	Fago λ, limpieza con xbeads 0.4x
9	Fago λ íntegro
10	Librería MinION sin procesar (sacada del spot on)

Tabla 12: Muestras del experimento de limpieza de ADN con bolitas magnéticas

El gel se migró a 100V y se tomaron fotos los 39 y 60 minutos mostradas en la figura 78.

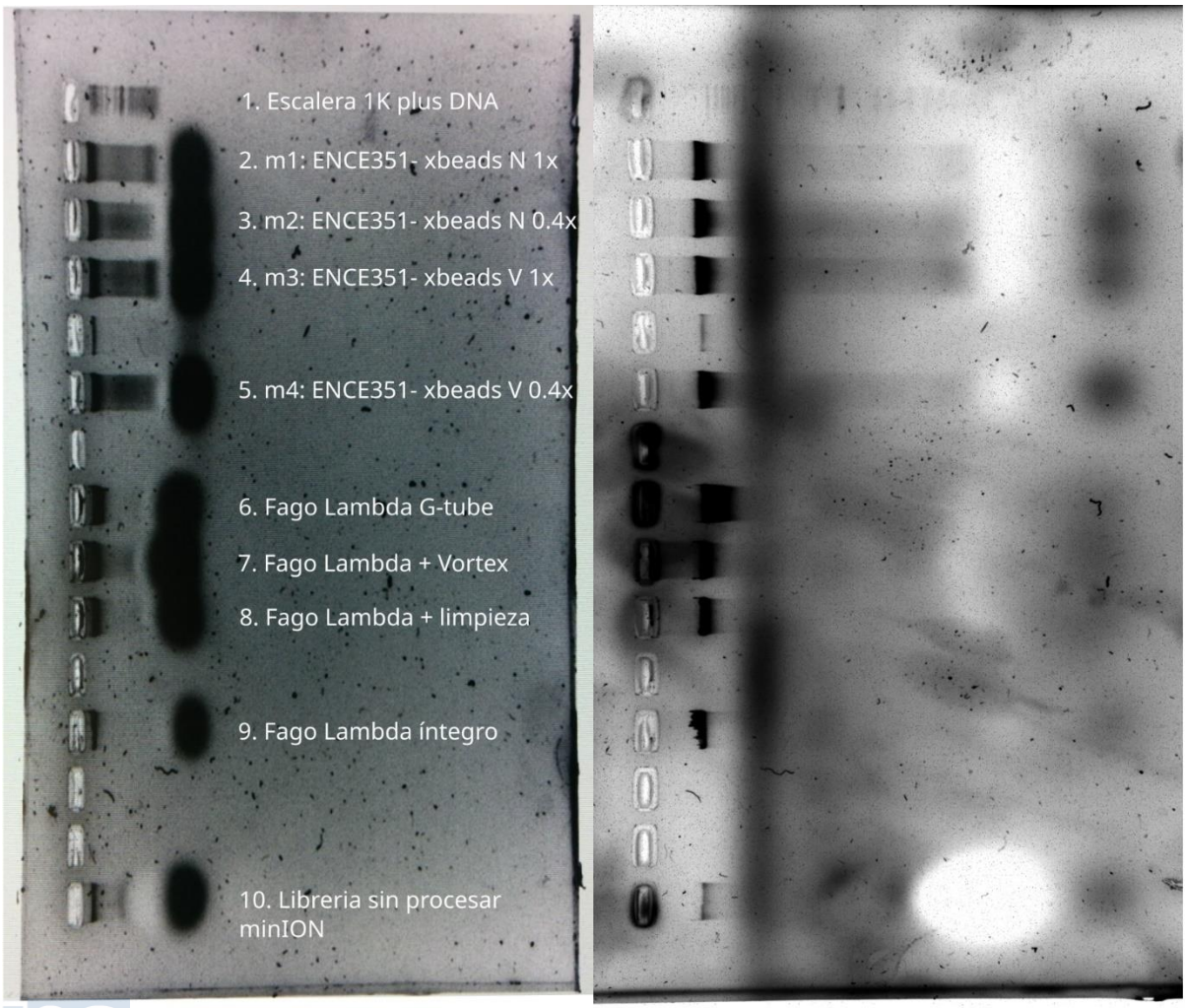


Figura 78: Gel de agarosa con el ADN tras limpiarse con distintos procedimientos con bolitas magnéticas..

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Los resultados son bastante claros, las bolitas magnéticas funcionan de forma similar independientemente si provienen de un bote recién abierto o de uno que lleva varias semanas. En cambio, sí que hay diferencias entre usar 1x (2 y 4) o 0.4x (3 y 5) de bolitas magnéticas, siendo estas últimas las que permiten remover los fragmentos más pequeños del ADN.

Vemos cómo el ADN del fago λ forma una banda muy nítida (9) y estrecha que corresponde a sus 48K pb y vemos que tras romperlo con el G-Tube (6) y el vórtex (7), la banda se hace mucho más ancha debido a la existencia de fragmentos de menor tamaño, siendo la rotura con el vórtex la que produce un frotis (smear) mayor. Se puede comprobar como el hecho de limpiar el ADN produce una pequeña degradación (8) del ADN.

No relacionado con el experimento de limpieza, la muestra 10, contiene la librería no procesada por el secuenciador tras secuenciar el mismo ADN del berberecho que se extrajo de la celda de secuenciación (Spot on). Se ve que es una banda bastante nítida, debido a que el secuenciador fue procesando preferentemente los fragmentos más pequeños, dejando los de mayor peso molecular.

Nota **:** Revisando la bibliografía, descubrimos una Tesis y un artículo que trataba en parte los mismos experimentos que realizamos nosotros. (Schalamun et al. 2019)

6.1.1.8 Problemas a la hora de cargar la librería

En ciertas ocasiones, en algunas celdas que llevaban mucho tiempo almacenadas o ya se habían utilizado (y limpiado con nucleasas) el puerto de carga (spot on) podía estar atascado debido a que al líquido de almacenamiento se había secado o a restos de una carga anterior, en esos casos se recurrió a humedecerlo con agua mili-Q, antes de proceder a la carga de la librería

En algún otro caso la librería empezaba a introducirse en el puerto de carga (spot on), pero se atascaba y dejaba de admitir más librería. En esos casos se procedía a inclinar la celda unos grados, de tal forma que la librería empezara a fluir por los conductos del puerto de carga. La inclinación debe ser ligera y por poco tiempo para que no se vacíe todo el líquido del puerto de carga, y por lo tanto entrase una burbuja en la cámara de secuenciación con la consiguiente destrucción de los poros debajo de la burbuja.

Figura 79: Inclinando la flow cell de un MinION para ayudar a la introducción de la librería.

6.1.1.9 Recarga de ATP (refuelling)

Esta técnica consiste en la introducción de entre 250-500 ml de FB + FLT. Cuando la velocidad de secuenciación de las lecturas cae por debajo de 300-350 nucleótidos por segundo (lo óptimo son 450 nucleótidos por segundo). Esto se suele producir cuando el consumo de ATP está siendo mayor del esperado o existen muchos poros sin ocupar y el ATP se pierde, tras una secuenciación muy larga.

Un ejemplo es el del caso mostrado en la figura 80, donde se puede observar que la velocidad con la que las cadenas de ADN están pasando por los poros ha descendido bruscamente. En este caso como no había operarios en el laboratorio se debería haber parado de forma remota la secuenciación a la espera de poder producir el refuelling.

Se observan los aumentos de la velocidad del paso de las secuencias por el poro tras el refuelling (a las 18 horas y a las 25 horas).

Nota: FB + FLT contienen entre otros productos ATP, necesario para el funcionamiento de las helicasas que se encargan de desenrollar el ADN e irlo introduciendo por el poro.

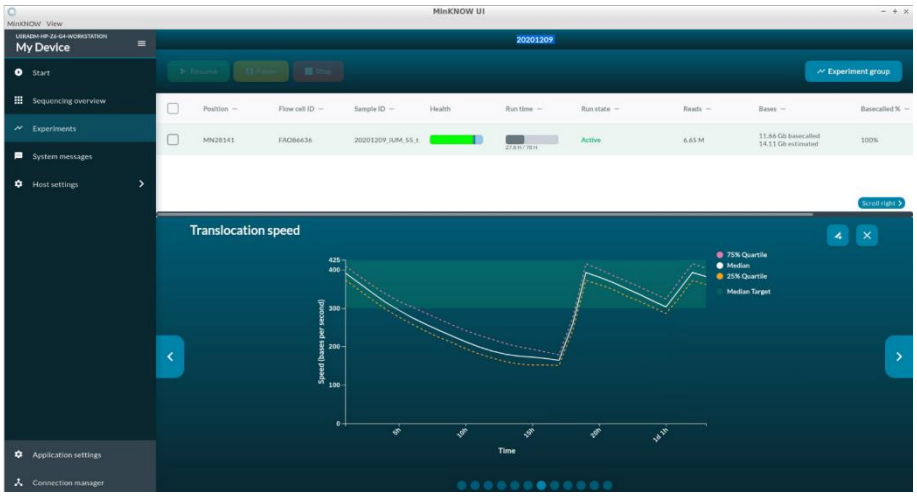
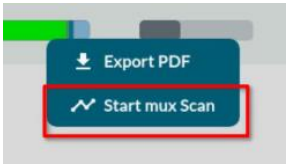


Figura 80: Salida del programa MinKnow mostrando la velocidad de secuenciación de una librería tras el refuelling.

Es recomendable que tras el "refuelling" se realice un reescaneo (mux Scan) de los poros para recuperar aquellos que la aplicación ha marcado como inactivos tras el descenso del número de cadenas de ADN que entran en los poros.



6.1.1.10 Rellenado (refiling)

Esta técnica consiste en la limpieza de la celda (ver procedimiento 5.1.12.4), y la recarga de más librería siguiendo el procedimiento estándar. Es conveniente usar esta técnica cuando la librería atasca mucho los poros de tal forma que la inversión de la corriente no es capaz de desatascarlos. La lógica subyacentes es que los agentes de limpieza tienen DNAsas, estas suelen desbloquear el poro, permitiendo que esté listo para su utilización tras la recarga de la librería.

6.1.2 Diseño y construcción de un mezclador de baja velocidad

La preparación de librerías de MinION de ONT, requieren de una agitación muy suave para no dañar el ADN. Por ejemplo con un dispositivo similar al mostrado en la figura 81. Inicialmente nuestro laboratorio no disponía de este dispositivo, por lo que se decidió hacer uno que fuese, rápido y fácil de construir por lo que se optó hacerlo con el juego de construcción LEGO©. Las características necesarias eran una tasa de giro de 1 vuelta

Figura 81: Ejemplo de un mezclador comercial.

por segundo y capacidad para preparar 4 o 5 librerías al mismo tiempo.

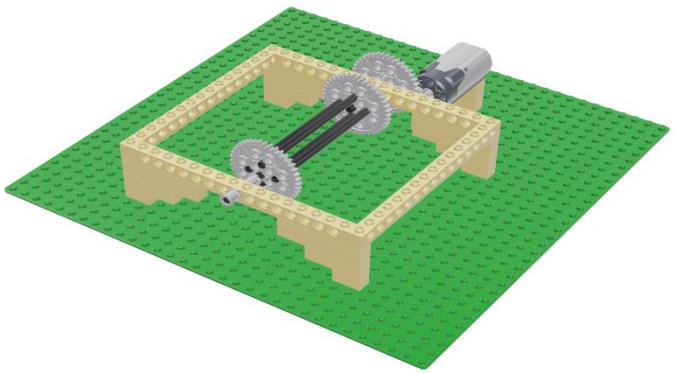


Figura 82: Diseño de un mezclador de baja velocidad con Lego

Se diseñó un prototipo con un programa CAD (computer aided design) específico

https://www.mecabricks.com/

El diseño consiste básicamente en un soporte para los tubos eppendorf, enganchado a un eje movido por un pequeño motor a través de un sistema de engranajes para reducir la velocidad a la tasa de giro necesaria. Se muestran a continuación unas imágenes del dispositivo construido. El dispositivo funcionó admirablemente bien y se utilizó para las realización de las primeras decenas de librerías.

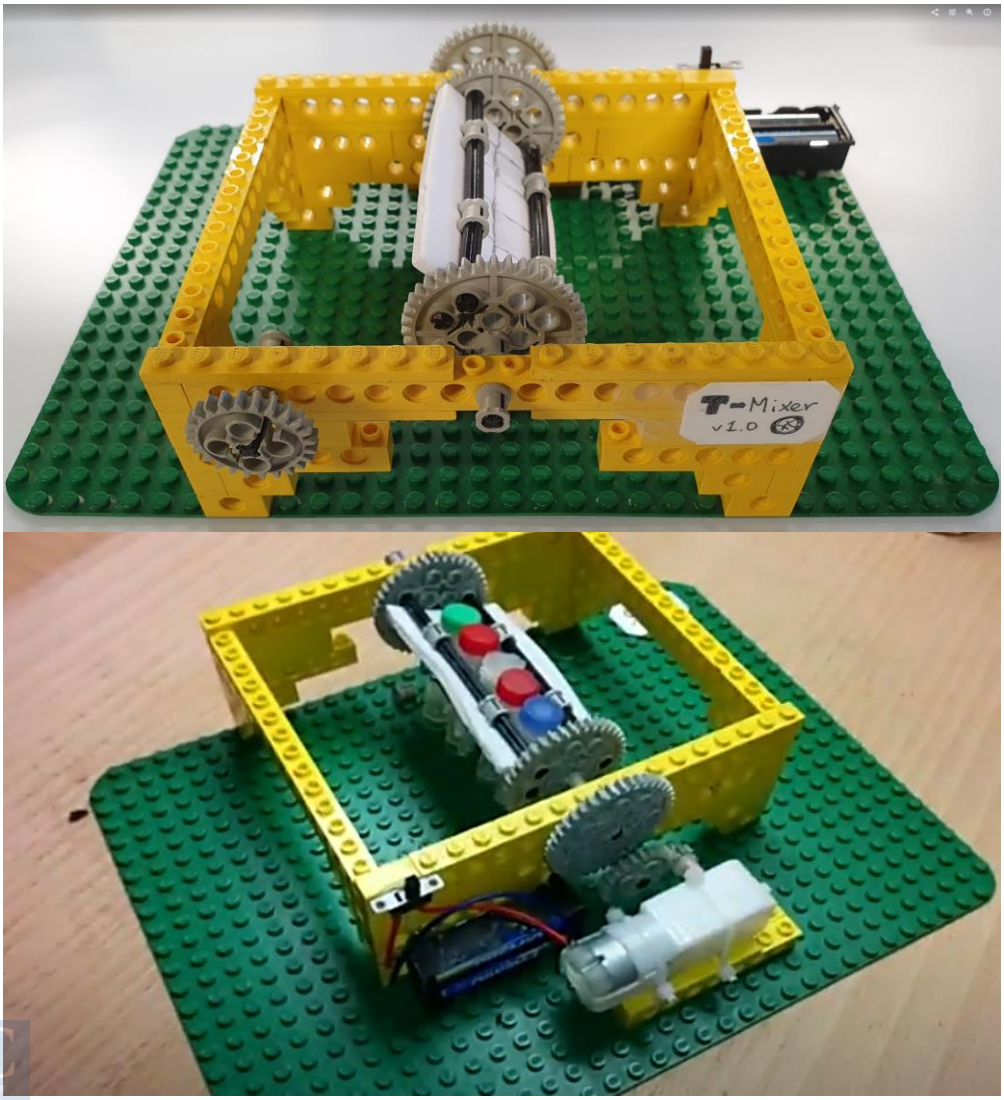


Figura 83: Mezclador de baja velocidad de construcción propia.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.1.3 Diseño y construcción de cubeta de electroforesis de reducidas dimensiones

Con el objetivo de poder migrar un gel de grandes dimensiones (8x2x1 cm) sin que este difundiese por los bordes del gel, cosa que sucedía en las cubetas de electroforesis standard, decidí crear una cubeta de electroforesis exactamente del tamaño del gel (ver apartado 6.1.1.7.2 recorte de bandas).

El despiece y el esquema de montaje se muestra a continuación. El material utilizado fue el polimetil metacrilato (PMMA), comúnmente conocido como plexiglás, vidrio acrílico, o simplemente metacrilato. Es un material transparente, químicamente inerte, que no reacciona con el tampón o las muestras que se separan durante la electroforesis. Esto es importante porque cualquier interacción química entre el material y las muestras podría interferir con el proceso de separación o alterar los resultados. Por último, es un material resistente y duradero que puede soportar las altas tensiones y temperaturas utilizadas durante la electroforesis sin agrietarse o romperse.

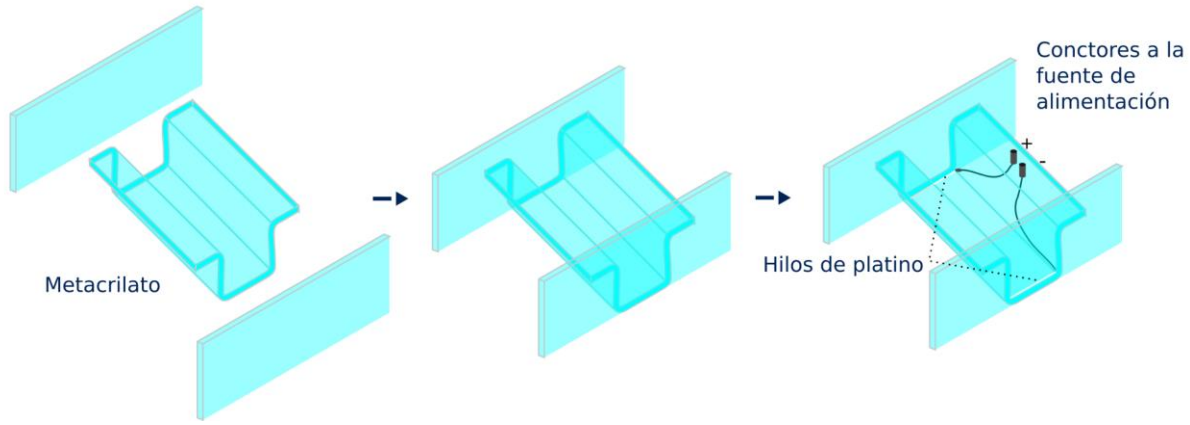


Figura 84: Diseño de cubeta de electroforesis para recorte de gel.

Para el soldado de las piezas se utilizó el propio metacrilato disuelto en acetona. Este adquiere la consistencia de una melaza bastante densa, que permite la unión y sellado de piezas de metacrilato. Cuando la acetona se evapora las piezas quedan firmemente unidas.

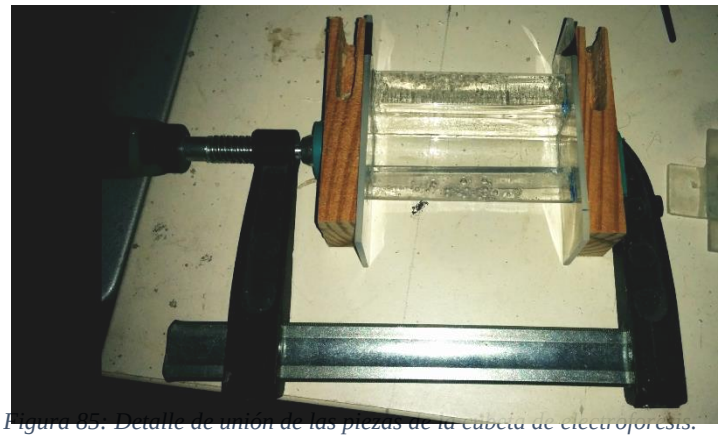


Figura 85: Detalle de unión de las piezas de la cubeta de electroforesis.

Inicialmente se utilizaron electrodos de zinc y cobre que se extrajeron de un cartucho de electroforesis usado de la máquina E-Gel power - Snap descrita en (6.1.1.7.10 Invitrogen E-gel y cartuchos de selección de tamaño). El problema que encontramos fue que el electrodo de cobre se deshacía tiñendo el gel de azul y verde como se muestra en la figura 86 (derecha).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

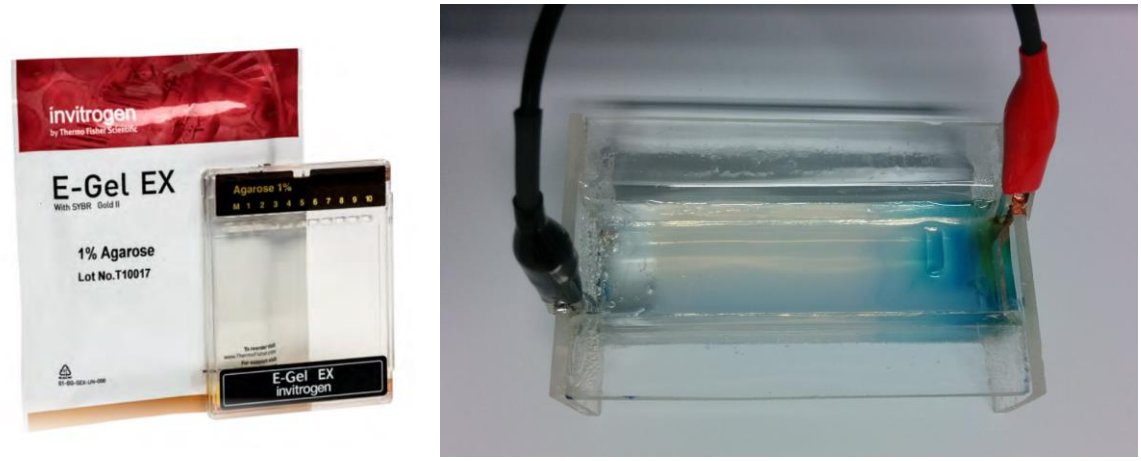


Figura 86: Cartuchos de los cuales se extrajeron los electrodos (izquierda) y prueba de funcionamiento con uno de los electrodos deshaciéndose.

Curiosamente la corriente sólo circulaba cuando se aplicaba el polo positivo al electrodo de cobre y el negativo al de zinc y no a la inversa. Funcionando de alguna forma como un diodo.

Para que funcionase correctamente fue necesario usar electrodos de platino, de un precio exorbitante, que por suerte pudimos extraer de una cubeta de electroforesis rota. Estos electrodos no se deshacen y pueden servir tanto para el cátodo como para el ánodo. Se muestra a continuación la cubeta montada figura 87 (izquierda), la cubeta desmontada junto con otra cubeta para la preparación del gel (centro) y una figura (izquierda) de la cubeta con un gel de agarosa dentro del transiluminador lista para añadir el ADN y hacer una migración tal como se indica en (6.1.1.7.5 Gel de grandes dimensiones y cubeta de electroforesis de construcción propia)

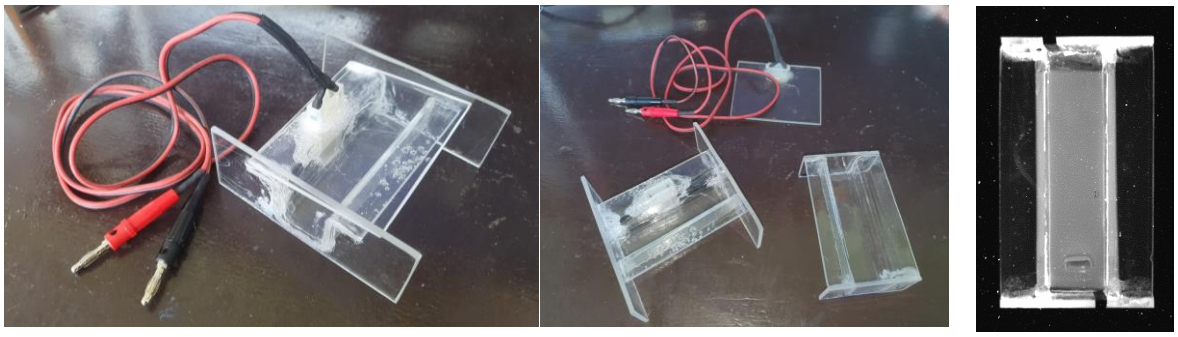


Figura 87: Cubeta de electroforesis de construcción propia (izquierda y centro) y detalle de la cubeta dentro del transiluminador con un gel de grandes dimensiones (derecha).

6.1.4 Diseño y construcción de una cámara refrigeradora

Para que la secuenciación de nanoporos funcione correctamente, la temperatura del dispositivo tiene que mantenerse en un rango bastante estrecho de temperaturas alrededor de 34º C, por ello los MinION disponen de un sistema para regularla consistente en una serie de resistencias para aumentar la temperatura y un pequeño ventilador para disminuirla intercambiando aire con el ambiente.

Pero cuando la temperatura del laboratorio pasa de 28 -30 grados los MinION no son capaces de mantener la temperatura de trabajo, como se observa en la figura 88. Esto afecta a la velocidad de las reacciones químicas necesarias para el desenrollamiento de la doble cadena de ADN y su paso a través de los nanoporos, lo que afecta negativamente a la precisión de la predicción de los nucleótidos que están siendo secuenciados (*basecalling*).

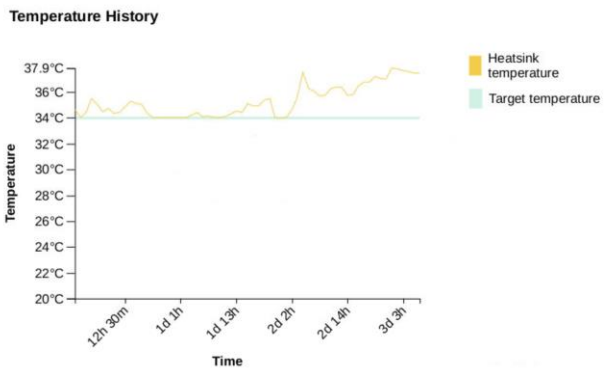


Figura 88: Salida del MinKnow mostrando las desviaciones de temperatura durante la secuenciación de una librería.

Tras investigar varias alternativas comerciales se optó por el diseño y construcción de una cámara refrigeradora específica para los MinION similar a la de (Sánchez et al. 2020) o algunos otros diseños de constructores amateurs.

6.1.4.1 Características

- Espacio suficiente para 5 - 8 MinION
- Volumen interior = 6 x 15 x 58.5 cm = 5.26 litros
- Capacidad de bajar la temperatura de la cámara de 5 a 10 grados con respecto al exterior.
- Controlado por Microprocesador: Arduino Nano.
- Sencillez de manejo.
- Construcción sencilla, robusta y económica.

El dispositivo está formado por una cámara aislada donde alojar los secuenciadores, un sistema para extraer el calor, y un sistema para controlar la temperatura. Existen varios sistemas para extraer calor de un sistema aislado, uno basado en la expansión y compresión de gases, usado en los frigoríficos de uso cotidiano, no obstante, para una cámara de tan reducido tamaño, implica una complejidad innecesaria, por lo que se ha optado por un sistema basado en el efecto Peltier ayudado por una serie de ventiladores para extraer el calor de los radiadores.

6.1.4.2 Construcción de la cámara

Para la construcción de la cámara, que es muy sencilla, se ha optado por una base y una bóveda de acero inoxidable de 2mm, rodeada de aislante de Poliestireno Extruido XPS de 40 mm (figura 89, panel A, B, G, I-8) y recubierta por una caja de DM de 3mm (figura 89, panel A, B). En la bóveda se encuentra insertado el circuito térmico para extraer el calor.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.1.4.3 Circuito térmico

Está basado en una unidad de intercambio de calor semiconductora basada en el efecto Peltier. El efecto Peltier es una propiedad termoeléctrica que produce una diferencia de temperatura cuando se aplica un voltaje eléctrico a dos metales o semiconductores conectados por una "unión de Peltier". Al pasar una corriente eléctrica por el circuito, se produce una transferencia de calor entre las uniones, enfriando una y calentando la otra. Este efecto se debe a la expansión de los electrones al fluir de una región de alta densidad a una de baja densidad, lo que resulta en la disminución de temperatura en la región de baja densidad.

El circuito térmico (figura 89 panel I) insertado en el aislante (I-8) consiste en un ventilador externo (I-1) que intercambia el calor del radiador externo (I-2) con el ambiente (panel A, figura 89), un cubo de 4x4x4 cm de aluminio (Panel D, E, F, I-3 figura 89) que evacúa el calor desde la unidad Peltier (I-5), e internamente otro pequeño radiador (I-5) y ventilador (I-6) para extraer el calor de la cámara (I-9) (Panel B, figura 89). Todas las piezas están unidas mediante una serie de gomas conductoras del calor (I-7). El objetivo de estas gomas es el contacto sin espacios de aire (un excepcional aislante) entre los distintos elementos del circuito térmico ya que todas las superficies a pesar de estar pulidas (Panel E, figura 89) tienen en su superficie irregularidades microscópicas.

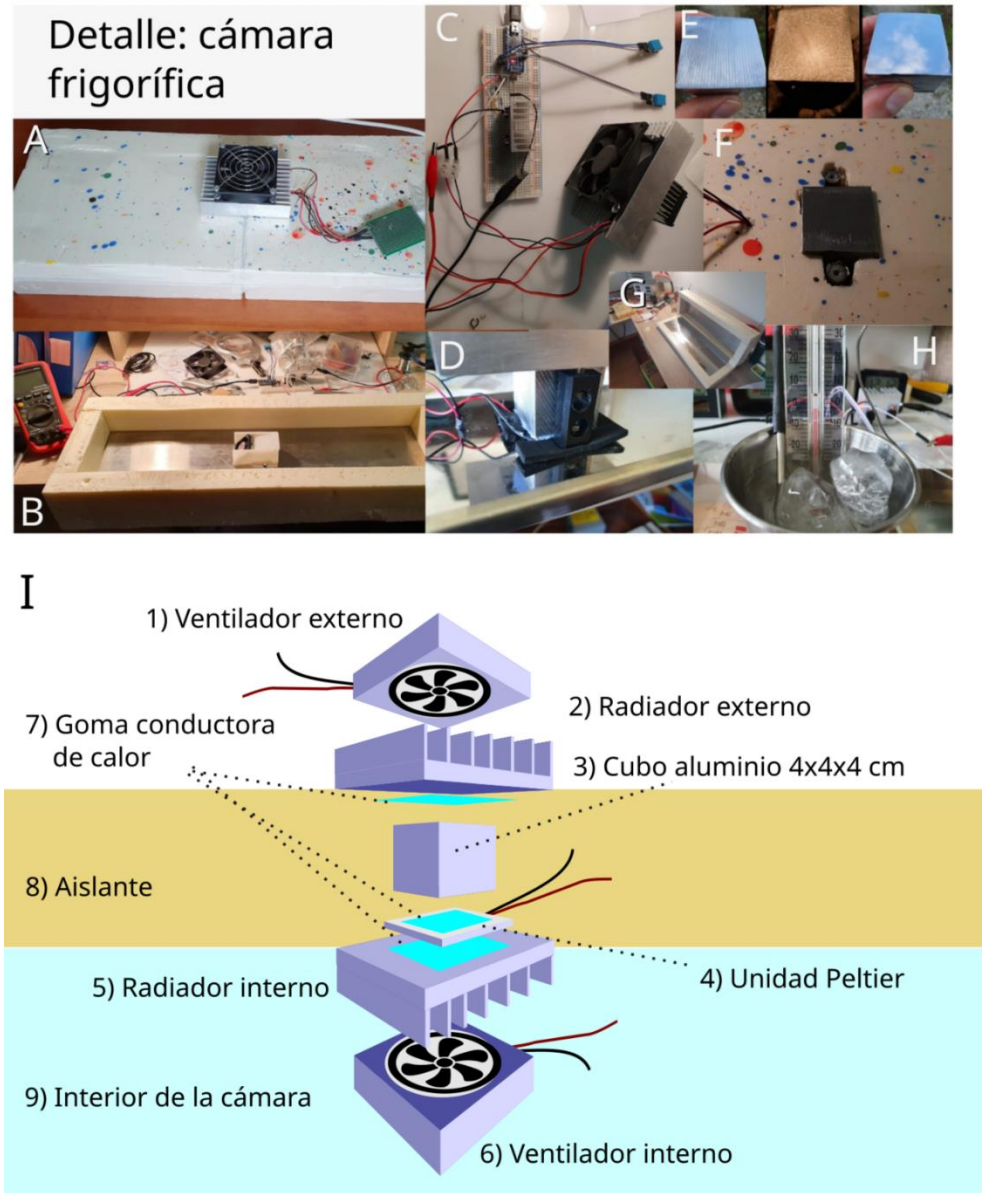


Figura 89: Construcción de la cámara refrigeradora.

6.1.4.4 Circuito eléctrico

Consta de tres partes:

- Sistema Peltier y dos ventiladores para dispersar el calor.
- Circuito eléctrico de control
- Fuente de alimentación AT, común a ambos sistemas.

La cámara frigorífica puede funcionar en modo manual, donde el encendido y apagado se realiza mediante un simple interruptor. También puede funcionar automáticamente mediante un sistema controlado por microprocesador (Arduino), y es descrito a continuación. Para controlar la temperatura y el encendido de la unidad Peltier TEC1-127, se ha recurrido a un microcontrolador Arduino Nano Rev3.0 y al sensor de temperatura DHT11. Básicamente el circuito controla la temperatura del interior de la cámara y enciende o apaga la unidad Peltier y los ventiladores para ajustar la temperatura interior de la cámara a la temperatura deseada. Aunque el microcontrolador es autónomo se conecta a un PC para monitorizar los datos y ajustar la temperatura.

El diseño del circuito eléctrico se muestra en la figura 91 y fue diseñado usando el programa Fritzing 0.9.3- http://fritzing.org.

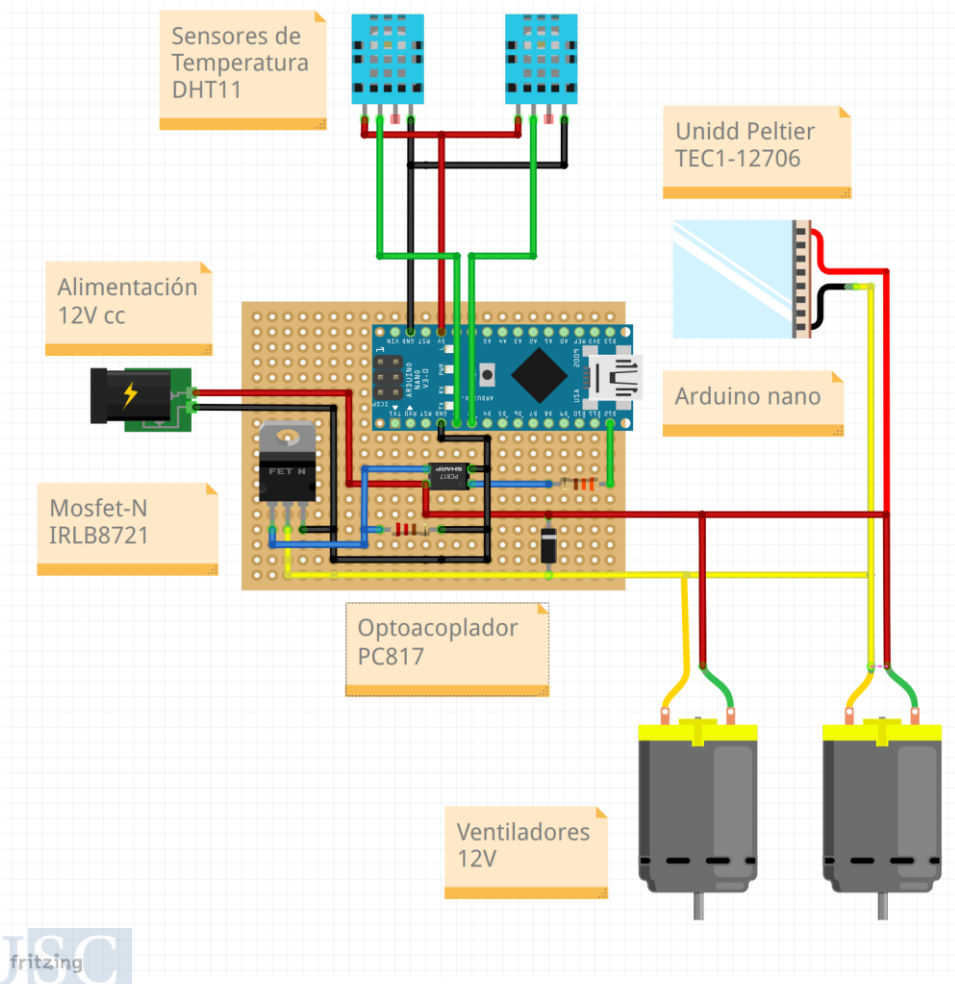


Figura 90: Esquema de los componentes electrónicos y su interconexión (realizada con el programa Fritzing).

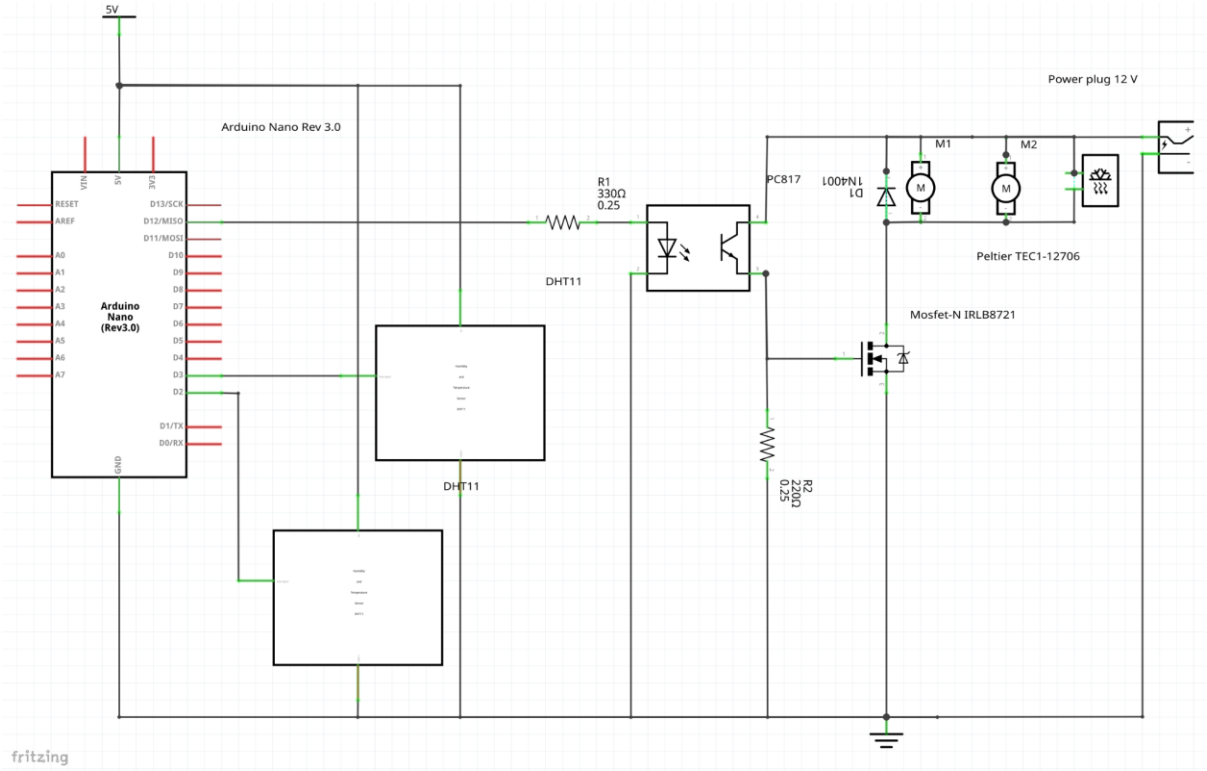


Figura 91: Esquema eléctrico del circuito eléctrico de la cámara refrigeradora.

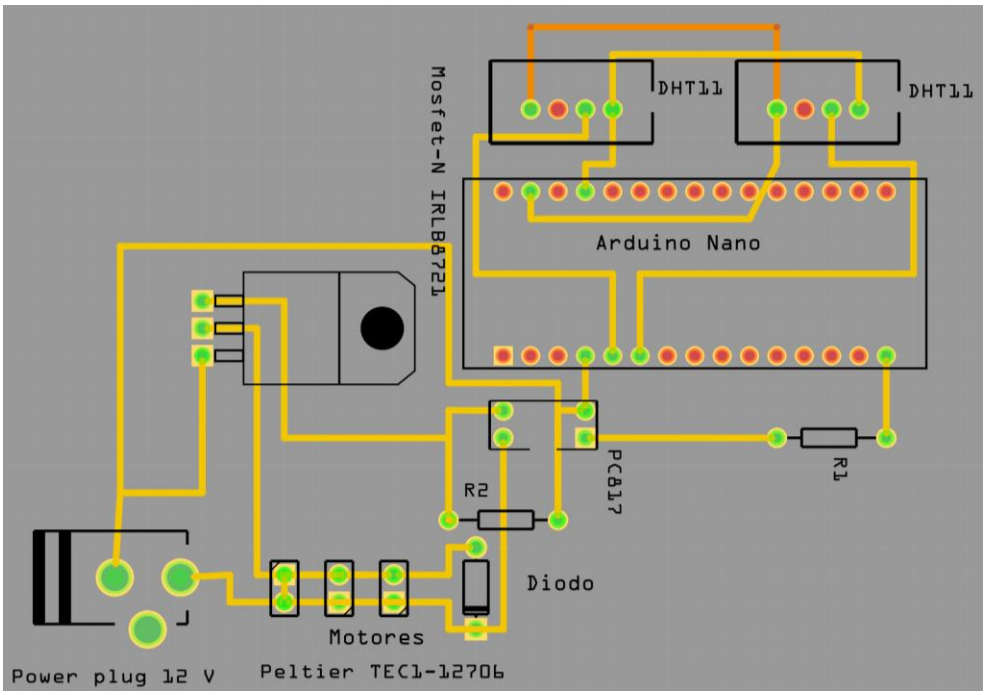


Figura 92: Esquema del PCB (printed circuit board).

Lista de materiales

Label	Part Type	Properties
Arduino Nano Rev 3.0	Arduino Nano (Rev3.0)	type Arduino Nano (3.0)
D1	Rectifier Diode	type Rectifier; package 300 mil [THT]; part number 1N4001
D2 x HT11	Humidity and Temperature Sensor DHT11	DHT11; u1 20-95%RH; u2 0-50°C; u3 ±5%RH ±2°C; output signal Digital Signal; variant 1; dc 3.3-5.5V DC
M1,M2	Ventilador 12V	
Mosfet-N IRLB8721	Basic FET N-Channel-flat	type n-channel; package TO220 [THT]; variant 1
PC817	Sharp PC817	Optocoupler, variant 2; package THT
Peltier TEC1-127	Peltier Element	
Power plug 12 V	Power plug	
R1	330Ω Resistor	bands 4; resistance 330Ω; power 0.25; package THT; tolerance ±5%; pin spacing 400 mil
R2	220Ω Resistor	bands 4; resistance 220Ω; power 0.25; package THT; tolerance ±5%; pin spacing 400 mil

Tabla 13: Lista de materiales de la cámara refrigeradora

6.1.4.5 Pruebas de la cámara refrigeradora

Una vez construida la cámara (figura 93) se realizaron varias pruebas para comprobar su capacidad de refrigeración. Estas pruebas consistieron en ver las curvas de temperatura (figura 94) con la cámara vacía (lineas 1-6) y con una masa de agua de medio litro de a 90ºC. Se observa en las curvas que el sistema es capaz de extraer el calor de la cámara y bajar la temperatura 6-7 grados con respecto a la temperatura exterior, siendo esta el factor limitante. En las pruebas con la masa de agua a 90º C se observa que la cámara, con el sistema de refrigeración apagado va perdiendo calor y tarda cerca de 10 horas en igualar su temperatura con el exterior (línea naranja), en cambio con el sistema de refrigeración encendido alcanza la temperatura ambiente en algo más de hora y media y consigue bajarla temperatura interior 7-8 grados con respecto al exterior.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

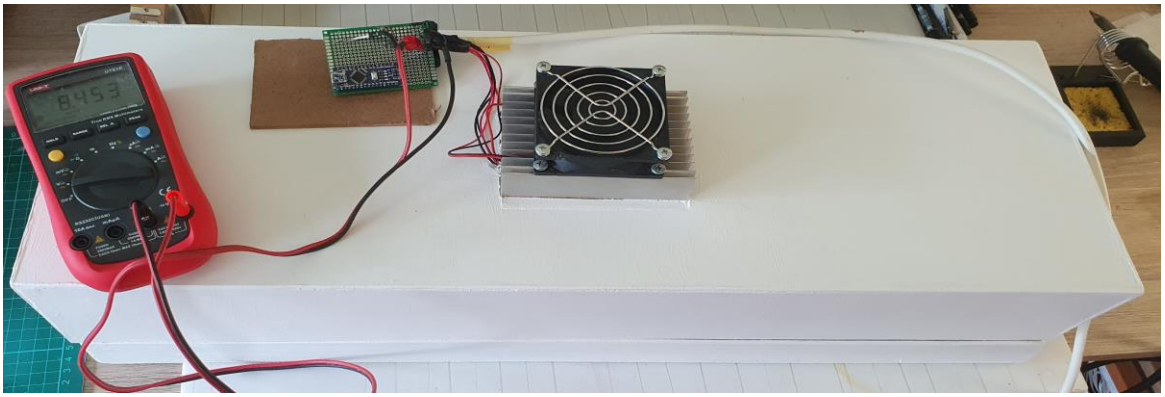


Figura 93: Pruebas de funcionamiento de la cámara refrigeradora.

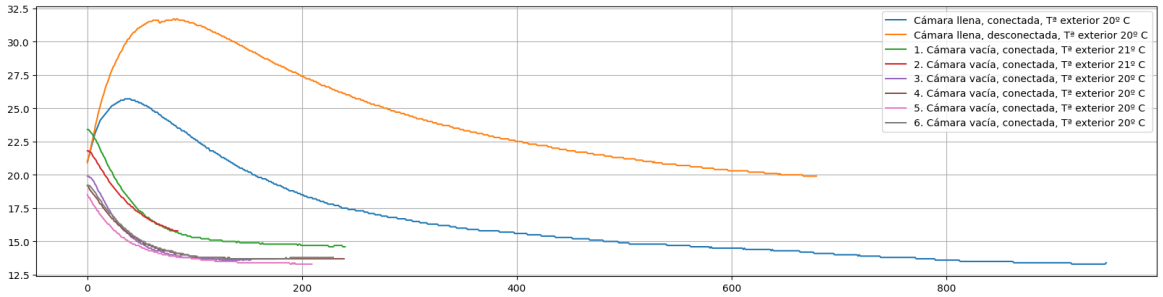


Figura 94: Gráficas de temperatura. (prueba de la cámara refrigeradora)

También se realizaron pruebas del circuito de control en un protoboard (figura 89 panel C), permitiendo controlar la temperatura en el interior y exterior de la cámara y encendiendo y apagando el sistema de refrigeración según las necesidades programadas.

Desafortunadamente en las pruebas con el circuito final, una sobre corriente destruyó el microprocesador Arduino, por lo que no se pudo probar el funcionamiento de la cámara completamente montada y con el sistema automático. Al mismo tiempo mejoras en el sistema de refrigeración ambiental en el laboratorio hicieron innecesaria la cámara frigorífica, por lo que el desarrollo se discontinuó.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.2 DESARROLLO DE PROGRAMAS INFORMÁTICOS.

Una parte importante del trabajo de esta Tesis fue el desarrollo de programas y aplicaciones para, por un lado gestionar la ingente cantidad de datos producidos, y por otros su análisis principalmente la metilación de los LINE1. Algunos de ellos son descritos en los siguientes apartados.

6.2.1 Programas auxiliares

6.2.1.1 MGD_agente: el problema de los millones de ficheros

Inicialmente el programa que generaba los datos de secuenciación, Albacore y posteriormente Guppy (hasta la versión 18/12/2018) generaban un fichero con extensión *.fast5* por cada una de las moléculas secuenciadas por el MinION, estos ficheros están codificados un formato de datos jerárquico (Hierarchical Data Format 5 (HDF5)). Esto hacía que por cada muestra secuenciada se generasen, literalmente, cientos de miles de ficheros, imposibles de gestionar incluso a pesar de trabajar en el computador de altas prestaciones (high performance computer HPC) del Centro de Supercomputación de Galicia (CESGA) (el número máximo de ficheros permitido para las cuentas genéricas es de doscientos mil ficheros). Posteriormente, a pesar de obtener una cuenta sin esta limitación trabajar con tantos ficheros era inasumible, llegando en alguna ocasión a trabajar con más de diez millones de ficheros. Para solucionar esta situación, se diseñó el programa MGD_agente, que gestionaba los cientos de miles de ficheros comprimiéndolos y generando una estructura más manejable por el sistema operativo. (ver esquema de la figura 95)

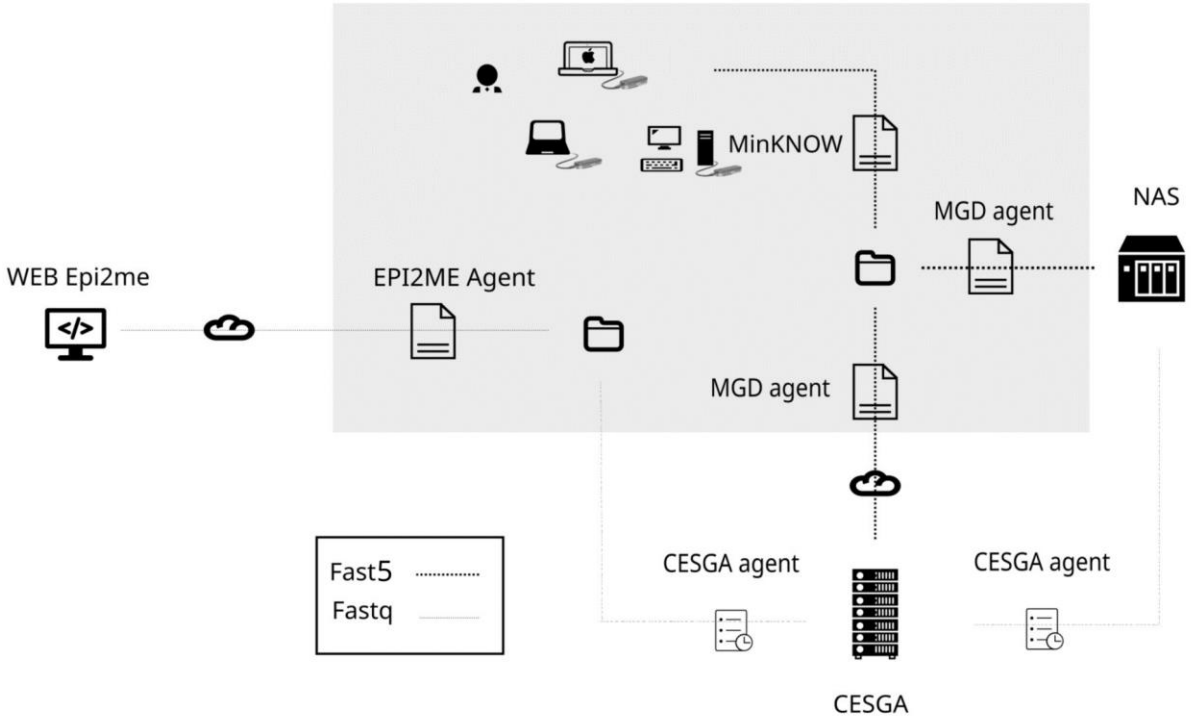


Figura 95: Esquema del sistema MGD-Agente.

Afortunadamente la empresa Oxford Nanopore Technologies, solucionó este problema, haciendo que el programa de *basecalling* Guppy, a partir de su versión 18.12, generase un solo fichero *.fast5* con cuatro mil lecturas (por defecto), haciendo la salida del programa mucho más manejable. Debido a ello se discontinuó el uso del programa. No obstante hay una copia el repositorio adjunto a este manuscrito en el directorio.

< ../6_programas/mgd_minion_agent-master.zip/ >

6.2.1.2 SVI. Integrador de variantes estructurales (Structural variants integrator)

Las inserciones de LINE1, así como del resto de transposones, son un tipo de variación estructural. Para el análisis genérico de los LINE1, su comparación e integración con otro tipo de variantes estructurales generados por diversos algoritmos (SV *basecallers*) existen varios programas como SURVIVOR, **(https://github.com/fritzsedlazeck/SURVIVOR)**, SVtools **(https://github.com/hall-lab/svtools)**, etc. Pero en general para integrar la salida de cada uno de los *basecallers* se recurre a programas y *scripts* realizados *ad hoc.* Para remediar esta situación se desarrolló un programa genérico para realizar estas funciones denominado **SVI**, capaz de integrar ficheros de datos generados por otros programas, posibilitando la selección, intersección, unión, y diferencia de variantes estructurales, así como, de realizar su análisis y la generación de gráficas y tablas. El programa funciona desde la línea de comandos o desde una consola de texto integrada en este. La idea central del programa es tener un formato interno genérico para todos los tipos de variantes estructurales y convertir la salida de cada uno de los *basecallers* particulares a este formato "universal" para luego hacer los análisis en ese formato. En caso de que aparecieran nuevos *basecallers*, solo haría falta hacer *de novo* el adaptador de su salida, sin tener que modificar el resto del programa. Una de sus características es su capacidad de integrar los datos de los *basecallers* más comunes de lecturas cortas y largas, Svaba, Smoove, Svim, Delly, Sniffles.

Nota: El programa, desarrollado en Python y Pandas se encuentra en el repositorio adjunto

<../06_programas/svi-master.zip>

6.2.2 Methyl-Meiga

Methyl-Meiga es un programa informático, desarrollado en Python, para el estudio de la metilación de los LINE1 en tumores usando lecturas largas de ONT. Es el componente principal, y la base de nuestro análisis informático de los patrones de metilación de los LINE1 en tumores. Surge como respuesta a la falta de programas específicos para tal cometido, ya que la mayoría de los programas y pipelines existentes, o están diseñados para utilizar lecturas cortas (Illumina), o no eran específicos de retrotransposones (variantes estructurales). La mayor parte de los análisis realizados hasta el momento con lecturas largas de los patrones de metilación de los retrotransposones se realizaron mediante *scripts* y análisis *ad hoc*.

Otro de los objetivos es permitir una mejor visualización de los datos de metilación, pues si bien el IGV en el modo bisulfito, presenta la información de una forma muy precisa y clara, los nuevos formatos con las etiquetas, **mn** y **ml** en los ficheros bam, no dan tanta

información como se puede comprobar en la siguiente figura, donde las citosinas sin metilar no parecen reflejadas.

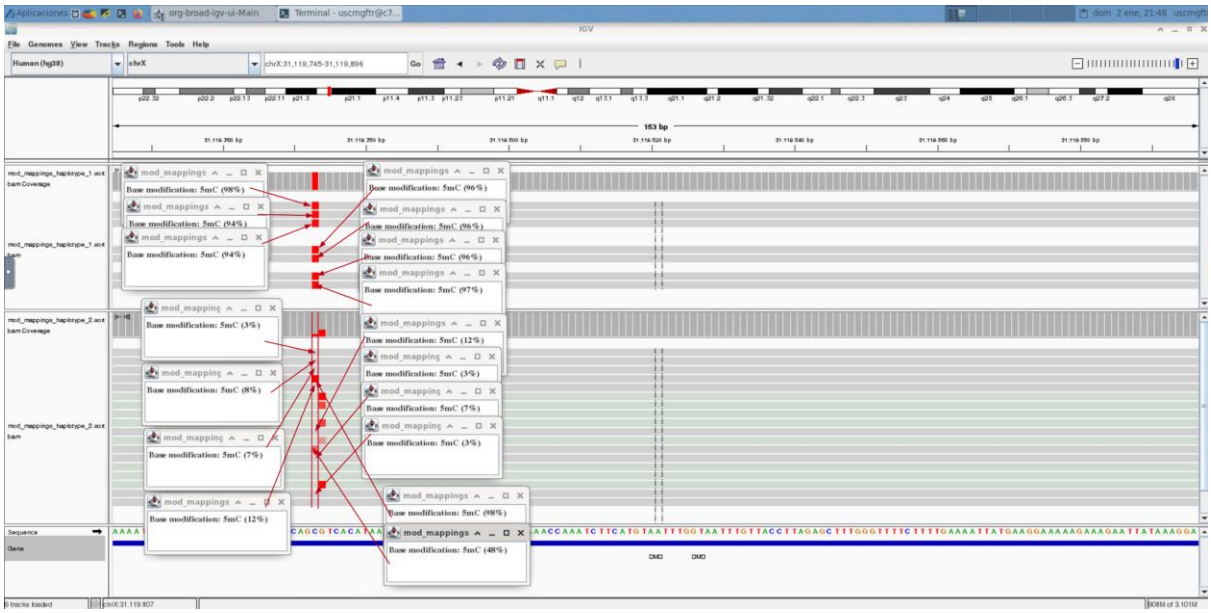


Figura 96: Imagen del IGV previo a marzo 2023, donde las citosinas no metiladas no se distinguen.

Nota: En su constante mejora las nuevas versiones de IGV (a partir de marzo 2023) han incorporado la posibilidad de visualizar el estado de metilación de las CpG en secuenciaciones de lecturas largas con las etiquetas **mm** y **ml**, marcándolas en rojo si están metiladas y en azul si están desmetiladas, tal y como funcionaba con las muestras tratadas con bisulfito.

Methyl-Meiga hace uso de los datos generados por Megalodon, un programa desarrollado por la misma empresa que produce los secuenciadores (ONT). Este programa es un *basecaller* que obtiene la información de las secuencias de ADN o ARN incluyendo las modificaciones en las bases (en concreto el estado de metilación) a partir de los datos de la señal eléctrica (*squiggle*) generada por los secuenciadores de nanoporos de ONT (en formato .fast5). Se decidió usar Megalodon, debido al alto grado de correlación de los resultados que produce en comparación con el "*Gold standard"* (bisulfito + Secuenciación por Illumina) tal como aparece reflejado en la literatura (Yuen et al. 2021; Q. Liu et al. 2019; Y. Liu et al. 2021), no obstante también he comprobado personalmente la validez de esa correlación (ver apartado 6.3). Otro de los motivos para su uso es su simplicidad de manejo y el que permite usar los datos de ficheros multilectura, así como la posibilidad de hacer el basecalling de metilación sobre lecturas específicas y no solo de todo el genoma. Otros programas como Nanopolish, con muy buena correlación con el *"Gold standard"*, obligaba a convertir los ficheros de datos multi lectura en ficheros con una sola lectura cada uno, haciendo que los análisis resultasen muy engorrosos. Otros programas como Tombo, signalAlign, Nanomod, mCaller, DeepSignal, DeepMod, o Guppy, no obtenían buenas puntuaciones al comparar sus resultados con el *"Gold standard"* (Yuen et al. 2021; Q. Liu et al. 2019; Y. Liu et al. 2021). Otros programas como methBERT, deepMP o pipelines como las de METEORE, aparecieron con posteridad.

USC DE SANTIAGO DE COMPOSTELA

El programa tiene dos modos de funcionamiento, para elementos germinales y/o regiones en el genoma de referencia y para inserciones somáticas. El fichero de entrada es un sencillo fichero de texto plano con las coordenadas de los loci que queremos estudiar. Para los elementos germinales hemos seleccionado 693 LINE1 íntegros, la mayor parte de ellos L1HS, pues son los que presentan actividad en los seres humanos (ver apartado 7.2.3.3 Base de datos definitiva de LINE1 germinales, íntegros). Para el estudio de las inserciones somáticas se han usado los datos obtenidos de ejecutar Meiga sobre el conjunto de datos de secuenciación de las muestras tumorales y normales objeto de este estudio (ver apartado 7.2.6 Inserciones somáticas de LINE1). Methyl-Meiga también puede utilizar ficheros genéricos o de la salida del programa xTea.

Los datos de entrada pueden corresponder a loci genéricos en el genoma de referencia, a transposones o a inserciones víricas. El programa crea un par de "cromosomas ficticios" por cada elemento a analizar (ya sea un loci, transposón o virus). Dependiendo de si el elemento se encuentra en el genoma de referencia o no estos cromosomas ficticios se construyen de forma diferente. Si el elemento se encuentra en el genoma de referencia, el alelo lleno consiste en la secuencia de interés más las regiones corriente arriba y corriente abajo, el alelo vacío consiste en las regiones corriente arriba y corriente abajo, borrándose artificialmente el elemento. Por el contrario en caso de que el elemento no esté en el genoma de referencia, se crean también dos cromosomas pero de la siguiente manera, uno con el genoma de referencia tal cual (correspondiente a alelo vacío) y otro con el genoma de referencia en el que se ha insertado la secuencia del transposón o virus (correspondiente al alelo lleno). Como se comentaba anteriormente, se incluyen también en estos "cromosomas artificiales" parte del genoma de referencia alrededor de la inserción, es decir, las regiones corriente arriba y corriente abajo del punto de inserción (en nuestro caso 20.000 pares de bases, no obstante este dato está parametrizado y se puede modificar). La idea de crear un genoma artificial es proporcionar la información de interés al programa Megalodon, el cual alinea las lecturas indicadas (o todas las lecturas si se corre en modo "todo el genoma") contra el genoma de referencia que le proporcionamos (en este caso dos cromosomas ficticios por cada elemento). Esto nos va a permitir comprobar si el genoma es homocigoto o heterocigoto para cada uno de los elementos o regiones de estudio. La información resultante es **fuertemente filtrada** (figuras 97 y 98) y procesada para evitar los problemas inherentes de trabajar con transposones e inserciones víricas con miles o cientos de miles secuencias similares repartidas por todo el genoma, lo que puede producir alineamientos espurios por parte del alineador que usan los *basecallers* (minimap2 en el caso de Megalodon). Para poder detectar correctamente los alineamientos a ambos cromosomas artificiales, hace falta indicarle a Megalodon el parámetro *"--allow-*

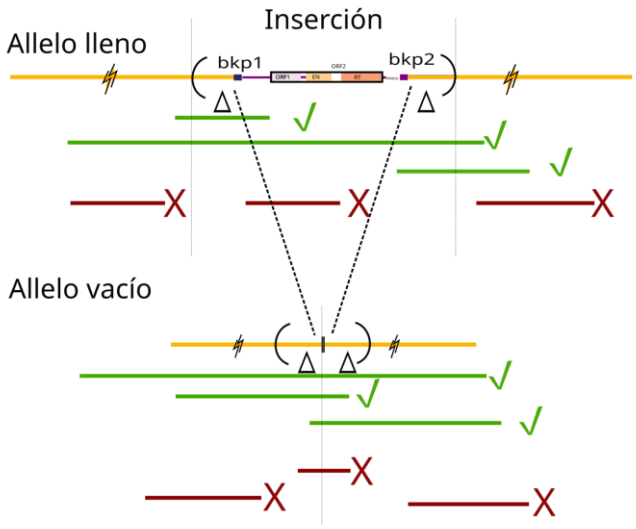


Figura 97: Esquema del filtrado de alineamientos realizado por Methyl-Meiga.

En amarillo los cromosomas artificiales generados. Solo son usados los alineamientos en verde, mientras que los que se muestran en rojo son descartados.

supplementary", lo que produce múltiples alineamientos, que como comentábamos anteriormente hace falta filtrar. Este filtrado se produce a varios niveles: Se eliminan lecturas que no se solapan con alguno de los puntos de ruptura (*breakpoints, bkp*) donde se produce la inserción o se encuentra el elemento (+/- Δ pares de bases), o éstas son muy pequeñas (líneas rojas en la figura 97, recuadros rosas en la figura 98) o están completamente contenidas dentro de la inserción (por lo que al no poder alinearlo con la parte unívoca del genoma no se puede garantizar que pertenezcan a esa inserción o elemento). También se filtran las lecturas que alinean mejor en el otro alelo (recuadro rojo en la figura 98)

Figura 98: Visualización en el IGV del filtrado de alineamientos espurios.

Una vez los datos han sido filtrados se generan los correspondientes ficheros de alineamiento ya filtrados. Estos ficheros de alineamiento son los que contienen la información de metilación que es extraída para su posterior procesado.

Debido a los fuertes requerimientos informáticos del programa Megalodon, Methyl-Meiga se puede correr de forma modular, pudiendo ejecutar la sección correspondiente Megalodon en procesadores con aceleradores gráficos (GPU) y el resto de las secciones en procesadores solo con CPU.

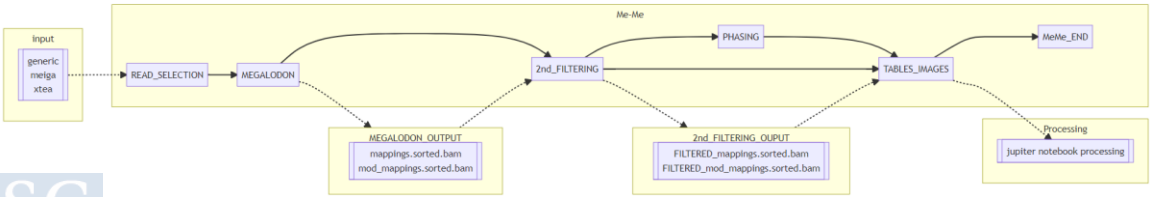


Figura 99: Diagrama de bloques del funcionamiento de Methyl-Meiga.

Methyl-Meiga hace uso de los siguientes ficheros:

1. Archivo bed (o similar) con la posición (chr start end) y datos adicionales de los LINE1 o loci a estudiar (como su orientación)

2. Genoma de referencia (principalmente hg38 en este trabajo)

3. Datos de secuenciación sin procesar (raw) generados por el MinION (.fast5)

4. Archivo de alineación (.bam) para seleccionar las lecturas de la región de interés. Este fichero se obtiene a partir de la alineación de ficheros .fastq según la pipeline (5.2.3 Alineación de datos de secuenciación MinION (ONT lecturas largas))

5. En caso de disponer datos de haplotipo estos se incorporan a los resultados.

El flujo de trabajo de Methyl-Meiga es el siguiente.

Paso 1) 1º filtrado: Para cada elemento o loci se obtienen del fichero de alineamiento (.bam) aquellas lecturas que alinean en las regiones de interés. Para cada una se comprueba que cumple con los criterios de calidad, *"mapeabilidad"*, y son lo suficientemente grandes como para contener un delta $\boldsymbol{\Delta}$ (1,000) de pares de bases aguas arriba o aguas abajo (dependiendo de la orientación del elemento o loci) que nos **garanticen que el correspondiente alineamiento sea correcto** (esto se debe a que un segmento repetitivo de una lectura puede alinear en varios sitios por lo que podría estar alineando en un elemento o loci que no corresponde).

Methyl-Meiga genera 5 archivos intermedios.

1. Archivo con el nombre de las lecturas que alinean con la característica de interés (transposón, L1, loci genérico) y que pasan los filtros.

2. Fichero con un genoma artificial constituido por varios cromosomas, dos por cada elemento o loci de estudio.

3. Un archivo gff con información del elemento, o loci para mostrar en el IGV.

4. Archivo de referencia y depuración con todos los alineamientos de los elementos o locus de interés (incluido si no pasan los filtros del paso 1)

5. Fichero de log

Paso 2) Methyl-Meiga internamente lanza Megalodon, que se puede ejecutar en modo emulación de bisulfito (para hacer controles visuales en IGV), o en modo "etiquetas" que genera un fichero .bam con la información de metilación embebida (**mm, ml**). Este paso genera un par de ficheros .bam sin filtrar (uno con la información de los alineamientos y otro con la información de la metilación)

Paso 3) Tras finalizar Megalodon, Methyl-Meiga hace un 2º filtrado (figura 98) para eliminar aquellos alineamientos espurios (cientos de ellos por elemento) que este ha generado. Para ello se comprueba que las lecturas solo alinean en el elemento en el cual alineaban en el primer paso (lo que garantiza que cada alineamiento es único) y que las lecturas cumplen una serie de criterios de calidad de alineamiento y solapamiento de las regiones únicas del

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

genoma. Este paso genera un par de ficheros .bam filtrados (uno con la información de los alineamientos y otro con la información de la metilación)

Paso 4) Generación de listados e imágenes. El programa genera imágenes para cada uno de los elementos analizados, donde se pueden observar los patrones de metilación e hidroximetilación de cada una de las CpG para cada una de las lecturas, tanto en el alelo que contiene la inserción (alelo lleno) como el que no tiene inserción (alelo vacío). En caso de disponer de información de *phasing* de las lecturas esta es mostrada en el margen izquierdo.

El programa muestra el alineamiento del LINE1 en la posición 20,000 puesto que desde la posición 0 hasta la 20,000 es la región corriente arriba del elemento. (lo mismo corriente abajo desde la 26,000 a la 46,000). Independientemente de la orientación en la que el LINE1 esté insertado, siempre se muestra con la región 5' UTR del elemento a la izquierda, de forma que sea más fácil reconocer los patrones.

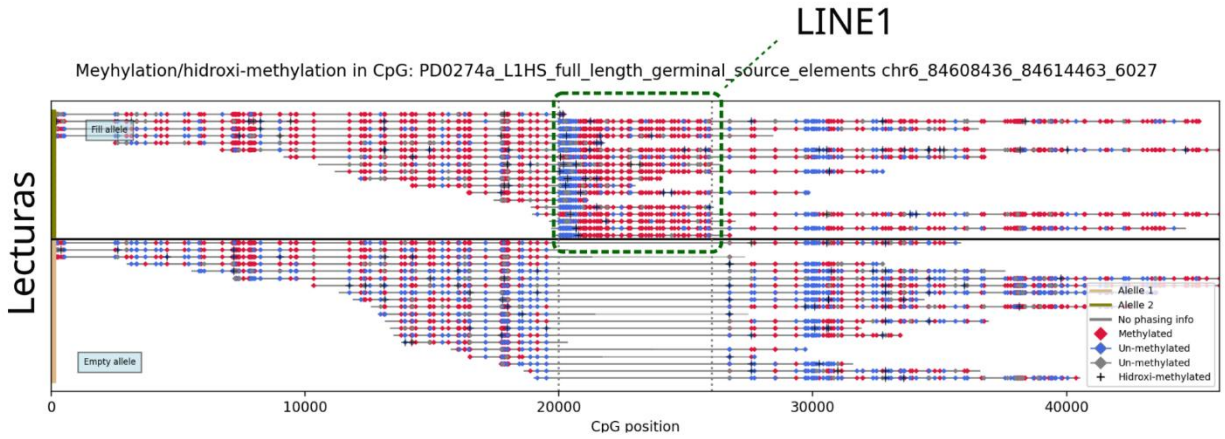


Figura 100: Salida de Methyl-Meiga mostrando un LINE1 germinal.
El elemento de interés se encuentra en el recuadro verde, en este caso un LINE1. Tal como se construyen los cromosomas artificiales la posición del elemento de interés empieza en la coordenada 20,000. Se pueden observar los alineamientos al alelo lleno (arriba) y al alelo vacío (abajo). Cada línea horizontal corresponde a una lectura obtenida por el MinION y los puntos azules y rojos corresponden a CpG sin metilar y metiladas respectivamente. Las + corresponden a CpG hidroximetiladas. Las barras verde y crema de la izquierda corresponden a los haplotipos de cada una de las lectura. que en este caso se corresponden a la perfección con los metil haplotipos.

La figura 101 (página siguiente), es una composición, a modo de ejemplo, de los distintos niveles de zoom que Methyl-Meiga produce de una inserción somática y elemento fuente del que proviene.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

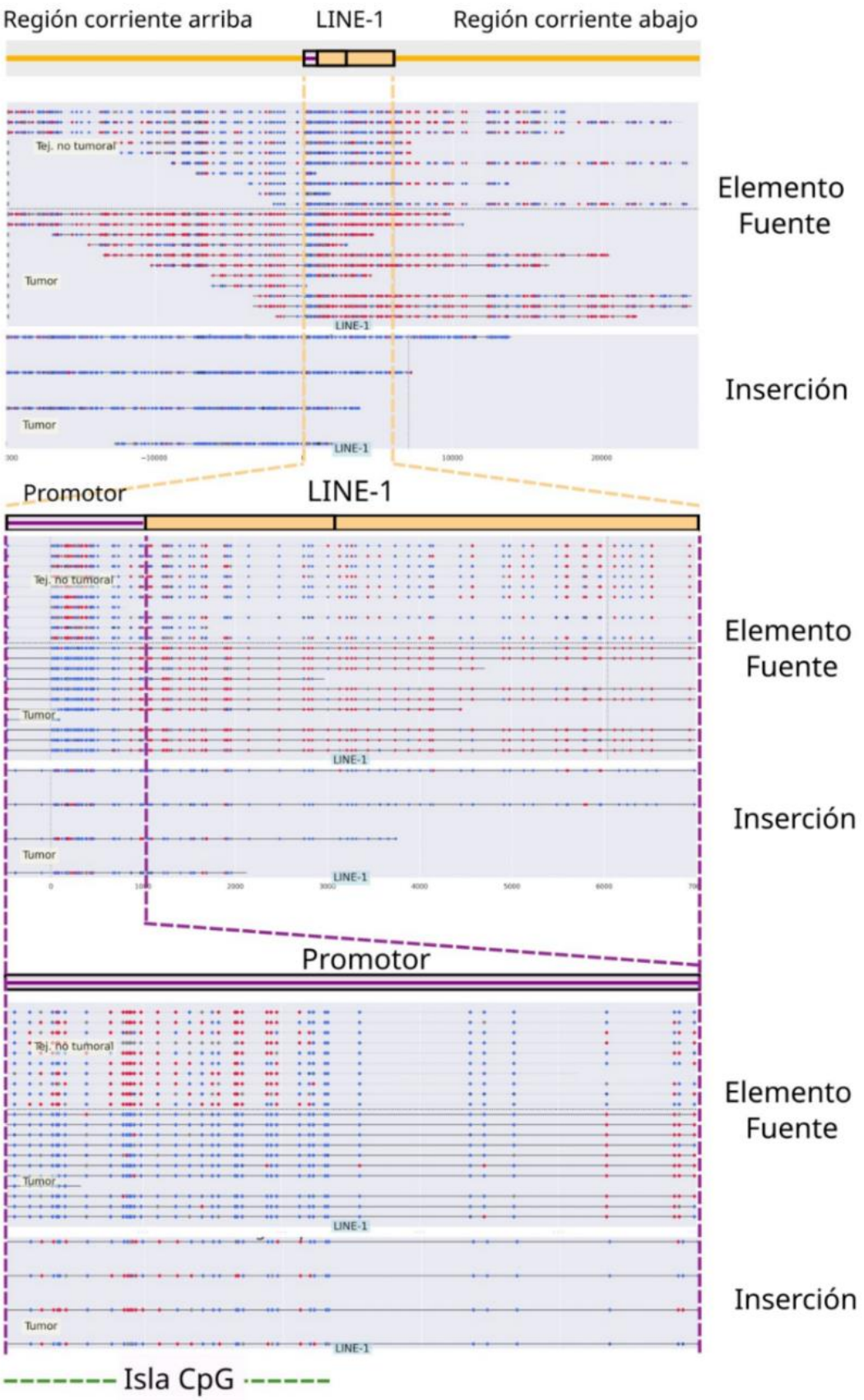


Figura 101: Composición de figuras producidas por Methyl-Meiga con distintos niveles de zoom. Inserción somática y el elemento fuente del cual proviene. Visión general (parte superior), solo el LINE1 (medio), o solo el promotor (abajo)

Lista de los ficheros producidos por Methyl-Meiga

Fichero	**Descripción**
Megalodon_reads.txt	Lista con la identificación de las lecturas facilitadas a Megalodon
Megalodon_artificial_ref_genome.fa	Genoma de referencia artificial, contra el que se alinean las lecturas de los LINE1 (u otros elementos o loci)
annotations.gff3	Fichero para facilitar la visualización de los ficheros de alineamiento generados por Methyl-Meiga en el IGV
1st_step_alignmen.tsv.gz	Información de las lecturas usadas para el análisis
FILTERED_mappings.sorted.bam	Fichero de alineamiento de las lecturas contra el genoma de referencia artificial
filtered_reads.tsv.gz	Información del filtrado de los alineamientos producidos por Megalodon
FILTERED_mod_mappings.sorted.bam	Fichero de alineamiento de las lecturas modificadas (información de metilación) contra el genoma de referencia artificial
mod_info_by_read_and_CpG.tsv.gz	**Fichero de salida de Methyl-Meiga con la información filtrada de la metilación e hidroximetilación por CpG y lectura. Es la base para los posteriores análisis.**
log.log	Fichero de información de la ejecución del programa
stats_allele.tsv.gz	**Genotipado de los elementos estudiados en la muestra.**

Tabla 14: Lista de los ficheros producidos por Methyl-Meiga

Paso 5) Postproceso, análisis de la información, generación de estadísticas y gráficas resumen con libros (*notebooks*) de Python.

Nota: El programa Methyl-Meiga se lanzó 2 veces para todas las muestras de esta cohorte, tanto para los tumores como para el tejido adyacente. En la primera ejecución se analizaron los elementos germinales, íntegros (7.2.3.3 Base de datos definitiva de LINE1 germinales, íntegros). En la segunda ejecución se analizaron las inserciones somáticas de LINE1 detectadas con MEIGA-LR (7.2.6 Inserciones somáticas de LINE1). En el tejido adyacente no existen los inserciones somáticas de los transposones, no obstante se ejecutó Methyl-Meiga con los mismos ficheros que en los tumores, para comprobar el estado original de metilación de las regiones donde se produjo la inserción.

6.2.3 Desarrollo de *scripts* de análisis estadístico y creación de gráficas (*notebooks* de Python).

Una vez procesadas las muestras con Methyl-Meiga, fue necesario analizar los datos obtenidos, por lo que se programaron varias centenas de *scripts* (más de 250), muchos de ellos de prueba, para intentar entender los datos, extraer conclusiones y generar tablas y gráficas.

Los *notebooks* y *scripts* definitivos se describen con más detalle en el apartado donde se usan y se incluyen en el repositorio adjunto a este manuscrito.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.3 VALIDACIÓN DE LA TECNOLOGÍA DE SECUENCIACIÓN DE DATOS DE METILACIÓN

En su concepción original esta investigación se nutría de estudios previos que analizan un número limitado de elementos fuente de los que se comprobaba el estado de metilación de las primeras CpG accesibles a la tecnología de secuenciación utilizada. Ejemplos de estos estudios son el trabajo de Tubio et al. en 2014 publicado en Science, los cuales se fundamentan en la secuenciación de ADN previamente transformado con bisulfito, ya sea de todo el genoma mediante secuenciación masiva (WGBS) o parte de él, acotando los loci y regiones de interés con cebadores específicos. Esta tecnología era (y es), en el momento de desarrollo de esta Tesis de doctorado, no necesariamente más precisa que la determinación de la metilación directa a partir de lecturas de Nanopore, pero si más contrastada que esta última. De ahí la necesidad de evaluar simultáneamente ambos métodos para contrastar resultados y ponderar la validez de los datos de ONT en la detección de la metilación.

En el momento de elaboración de este texto hay ya disponibles algunas herramientas informáticas para detectar metilación a partir de datos de lecturas largas de ONT (i.e., Bonito, Megalodon, Meteore, Remora, Nanopolish, Deep signal, …), así como varios estudios comparativos (Yuen et al. 2021; Q. Liu et al. 2019; Y. Liu et al. 2021) donde se analiza con detalle la validez de la tecnología de ONT y de estos *basecallers* para obtener la información de metilación de las citosinas en dinucleótidos CpG. En todo caso, en el momento de iniciar nuestra investigación las opciones eran más limitadas y la corroboración de su validez y precisión se limitaba, principalmente, a un *white paper* distribuido por Oxford Nanopore Technologies. En los estudios comparativos mencionados, se constata el alto grado de correlación de esta tecnología comparado con el ya mencionado "G*old standard*" de datos de WGBS generados con Illumina y analizados con Bismark, así como la precisión y eficiencia de los *basecallers* de metilación existentes, siendo el más prometedor por su precisión y consumo de recursos informáticos Megalodon. Pese a ello, decidimos hacer las comprobaciones pertinentes y nuestro propia comparativa (*benchmark*) entre WGBS-Illumina-Bismark y Nanopore-Megalodon, por dos motivos: corroborar los estudios anteriores y confirmar que estábamos haciendo un uso adecuado de esta nueva tecnología.

En este sentido, durante el curso de la investigación sobre retrotransposición y cáncer en nuestro grupo de investigación, se analizaron exhaustivamente dos líneas celulares derivadas de adenocarcinoma de pulmón, la NCI-H2087 y la NCI-H2009, así como sus equivalentes normales derivados de linfocitos circulantes inmortalizados (descritas en el apartado 7.1.2). Las líneas celulares tumorales se secuenciaron, entre otras tecnologías, por WGBS, lo que propició un *"Gold standard"* de datos bisulfitados de los que se dedujo el estado promedio de metilación de las CpGs utilizando Bismark (ver apartado 5.2.4). Además, se obtuvieron datos de secuenciación masiva de lecturas largas con Nanopore, de los que se pudo extrapolar el estado promedio de metilación en los loci CpG y por lectura empleando Megalodon (ver apartado 5.2.5). Disponer de ambos conjuntos de datos de metilación global para el conjunto de estas muestras, esto es, generados a partir de lecturas cortas con WGBS y a partir de lecturas largas con Nanopore, posibilita evaluar la eficiencia y precisión en la determinación de la metilación mediante Nanopore en todas las posiciones CpG del genoma en regiones no repetitivas. El procedimiento y el procesado de los datos se describen en los apartados a continuación.

6.3.1 Datos de metilación de las líneas celulares obtenidos mediante pretratamiento de bisulfito + Illumina

Como parte de este estudio se secuenciaron después del tratamiento con bisulfito (WGBS) las líneas celulares NCI-H2009 y NCI-H2087. Para ello tras la extracción del ADN con el protocolo (5.1.4) de extracción de ADN a partir de líneas celulares, se enviaron a secuenciar 5.58 ng y 2.32 ng de ADN (NCI-H2087 y NCI-H2009 en un volumen de 50 µl respectivamente) a la empresa Macrogen (Seúl, Corea sur). El tipo de librería solicitado fue WGBS "*Accel Methyl-Seq DNA"*. Como resultado se obtuvieron los ficheros:

```
-rwxrwxr-- 1 uscmgadm uscmg  37G feb 19  2020 DNAH2009_1.fastq.gz
-rwxrwxr-- 1 uscmgadm uscmg  38G feb 19  2020 DNAH2009_2.fastq.gz
-rwxrwxr-- 1 uscmgadm uscmg  38G feb 16  2020 DNAH2087-bis_1.fastq.gz
-rwxrwxr-- 1 uscmgadm uscmg  41G feb 16  2020 DNAH2087-bis_2.fastq.gz
```

Tabla 15: Ficheros de WGBS de las líneas celulares.

Para procesar la información se usó el programa Bismark (ver apartado 5.2.4) Inicialmente se lanzaron estos procesos en uno de los nodos "Thin" del CESGA (2 procesadores Haswell 2680v3, 24 / cores / 128 GB memoria / 1 disco HDD de 1TB / 2 conexiones 1GbE / 1 conexión Infiniband FDR@56Gbps). El proceso no dio terminado en el tiempo máximo de los procesos "estándar" (100 horas), tampoco incluso tras el personal del CESGA aumentar el tiempo en un 50%.

Viendo los problemas que estaba provocando la ejecución en el CESGA se descargaron datos de metilación similares a los nuestros: **SRR1035882** (**ncbi.nlm.nih.gov/sra/?term=SRR1035882**) para comprobar si el problema era debido a nuestros datos o al flujo de ejecución. Estos datos son de tejido normal de pulmón, perteneciente a un estudio de patrones de metilación en distintos tipos tumorales (Vidal et al. 2017) . Los datos son pair-end Illumina HiSeq 2000-Bisulfite-Seq, 21G bases. En ellos se probaron los procedimientos y programas antes mencionados terminando correctamente, de lo cual deduje que el problema estaba en nuestros datos.

Tras hacer pruebas de nuevo con solo algunos de los cromosomas de nuestros datos, comprobamos que terminaban correctamente pero cerca de los límites de tiempo máximo permitido a los trabajos en el CESGA (100 h).

Debido a la ingente cantidad de recursos que requeriría procesar todo el genoma se decidió probar otra estrategia, trasladando las tareas a un servidor del grupo, (HP Z6-G4 64 procesadores 128 GB, con 3 discos **SSD de 2TB** y 2 HDD 8TB) donde tras 30 horas por muestra, terminó correctamente. Sospecho que esta diferencia de tiempos se debió al acceso constante al disco por parte de Bismark, por lo cual el uso de los discos de estado sólido (SSD) fue crucial. Probablemente nuestro conjunto de datos aunaba una alta cobertura, y la complejidad inherente de los tumores, donde variaciones en el número de copias (*copy number*), y la alta clonalidad podrían dificultar el alineamiento.

Como veremos más adelante el programa de basecalling de datos de metilación Megalodon, tiene requerimientos similares en cuanto al uso del almacenamiento en disco. Siendo el uso de discos magnéticos estándar (HDD) poco adecuados para la tarea.

6.3.2 Datos de metilación obtenidos con ONT + Megalodon

Una de las grandes ventajas de la tecnología de nanoporos de Oxford Nanopore Technology es que a partir de los mismos datos en bruto (raw) de la secuenciación de las muestras se puede obtener la información la secuencia de nucleótidos y modificaciones epigenéticas como la metilación. Tal como se describe en el apartado 5.2.5 se procedió a lanzar el programa Megalodon sobre los datos en bruto (.fast5) de las líneas celulares tumorales (NCI-H2009 y NCI- H2087). De todos los ficheros generados por Megalodon para este este análisis se usó el fichero con la información de metilación por CpG y lectura

per_read_modified_base_calls.txt

6.3.3 Superioridad de la tecnología de ONT para analizar y estudiar la metilación de los LINE1

En primer lugar, y para comprobar la idoneidad de la secuenciación y el mapeo de las lecturas de WGBS, se procedió a revisar manualmente parte de los elementos germinales LINE1. La figura 102 contiene una captura de pantalla del *Integrative Genome Viewer* (IGV), mostrando el alineamiento de lecturas cortas de WGBS (arriba) sobre una posición genómica en el cromosoma chr22 (Chr22:28663284-28669316), así como el equivalente con alineación de lecturas de Nanopore en la misma posición (abajo). Las bases resaltadas en rojo indican citosinas metiladas, en tanto que aquellas en azul se corresponden como citosinas no metiladas. Esta posición alberga uno de los LINE1 más activos de forma recurrente en tumores, tal y como describen (Rodriguez-Martin et al. 2020). Se pueden apreciar amplias diferencias de cobertura y, como era previsible, las lecturas solo abarcan de forma inequívoca el comienzo del transposón, coincidiendo con el 5’ UTR del retrotransposón, y su término en 3’. Esto es de esperar, ya que dada su repetitividad el resto de las lecturas mapeadas sobre el cuerpo del LINE1 pueden provenir de cualquiera de los LINE1 de todo el genoma, lo que impide establecer conclusiones definitivas acerca de los patrones de metilación de LINE1 individuales más allá de sus extremos. Tal limitación se ve solventada con lecturas de Nanopore, que permiten resolver en toda su longitud estas regiones conflictivas repetitivas en el genoma, así como en otras posiciones de baja *“mapabilidad”*. Esta limitación es una de las principales motivaciones para el desarrollo del presente trabajo haciendo uso de lecturas de secuenciación largas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

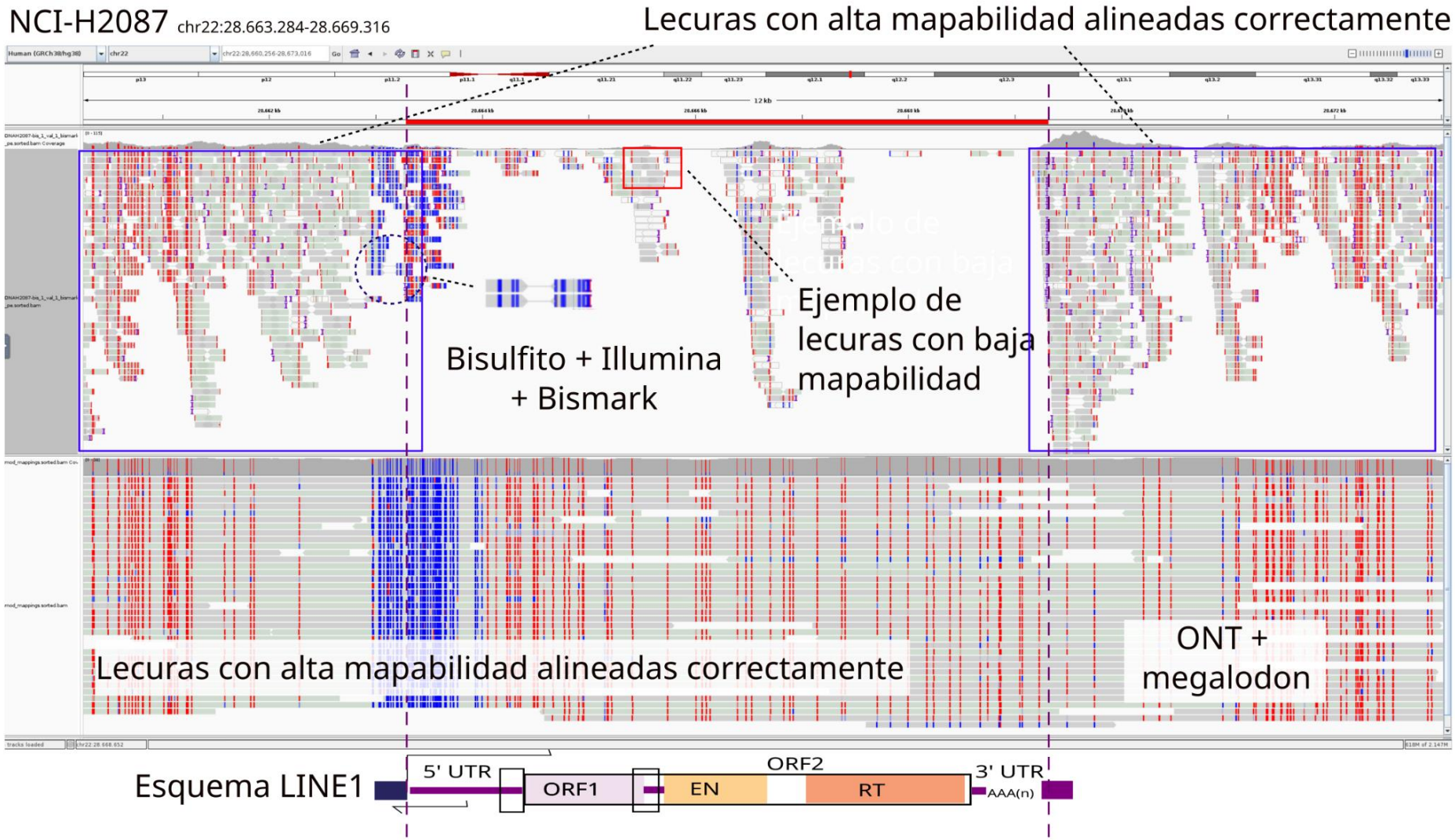


Figura 102: Diferencias de mapabilidad sobre un LINE1 entre lecturas cortas de Illumina y largas de ONT .

6.3.4 Comparativa de los datos de metilación de Megalodon con los datos bisulfito + Illumina ("Gold standard")

A pasar de las grandes diferencias en la mapabilidad entre los dos conjuntos de datos en las regiones repetitivas (como los transposones), los datos globales de metilación son muy similares entre sí, siendo la metilación media global obtenida por Nanopore-Megalodon ligeramente menor que la que deduce WGBS-Illumina-Bismark, respectivamente un 8.66% y un 6.05% menores para la NCI-H2009 y la NCI-H2087 (tabla 16). Si bien no disponemos de una explicación de esta diferencia, hay varias posibles explicaciones. Es posible que la desviación a la baja en ambas muestras en Nanopore-Megalodon con respecto al *"Gold standard"* sea debida a diferencias en la eficiencia de la transformación con bisulfito a lo largo del ADN genómico o, en otras palabras, a la presencia de regiones poco permeables a este tratamiento que comprometan la eficiencia de la transformación en esas regiones. También es posible que haya diferencias en la sensibilidad de las tecnologías de secuenciación utilizadas para detectar la metilación, lo que podría llevar a una subestimación de la metilación global en Nanopore-Megalodon en comparación con respecto el *"Gold standard"*. Además, la elección de las muestras y la variabilidad biológica entre ellas, en concreto, que fueron secuenciadas en pases distintos, también podrían influir en las diferencias observadas en la metilación global. En todo caso, tales diferencias, son bajas, en línea con lo descrito en estudio previos (Yuen et al. 2021; Q. Liu et al. 2019; Y. Liu et al. 2021) para datos de ONT.

	a) WGBS-Illumina-Bismark	b) Nanopore-Megalodon	Porcentaje de diferencia (a-b)/b *100
NCI-H2009	65.483 %	60.264 %	8.66%
NCI-H2087	70.399 %	66.381 %	6.05%

Tabla 16: Comparación de la metilación de las líneas celulares.
a) "Gold standard", bisulfito + Illumina + Bismark
b) Metilación obtenida por secuenciación de nanoporos + Megalodon

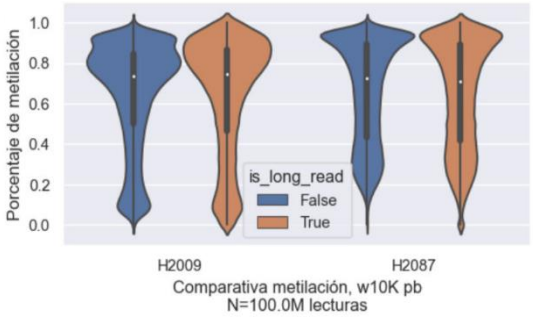


Figura 103: Comparación de la metilación de las líneas celulares (NCI-H2009, NCI-H2087) con el "Gold estándar".
Bisulfito + Illumina + Bismark, (Gold standard) en azul.
Metilación obtenida por secuenciación de nanoporos + Megalodon, en naranja

La regresión lineal entre los valores de metilación por posición CpG existentes entre ambos conjuntos de datos, ya sea por posición individual o por ventanas de 10Kb a lo largo de todo el genoma, indica una muy fuerte correlación directa entre los resultados generados por las dos tecnologías que, además, es en ambos casos estadísticamente significativa (figuras 104 y 105)

Los valores obtenidos fueron: r = 0.70 y r = 0.82. para la comparación de la metilación de CpG individuales (NCI-H2009 y la NCI-H2087 respectivamente) y r = 0.9 y r = 0.89 para la media de metilación en ventanas de 10,000 pares de bases (NCI-H2009 y la NCI-H2087 respectivamente). En todos los casos el *"p-value"* era prácticamente cero (0)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En conclusión, tras la comparación y validación de la tecnología de secuenciación de datos de metilación mediante Nanopore-Megalodon, podemos afirmar que esta tecnología es válida, precisa y fiable para la detección de la metilación en dinucleótidos CpG. La alta correlación directa y significativa entre los resultados obtenidos con Nanopore-Megalodon y el "*Gold standard*" de datos de WGBS generados con Illumina y analizados con Bismark, así como la precisión y eficiencia de los *basecallers* de metilación existentes, demuestran la capacidad de esta tecnología para la determinación del estado de metilación de las citosinas en dinucleótidos CpG. Además, la utilización de lecturas de Nanopore permite superar las limitaciones de *"mapabilidad"* de regiones repetitivas del genoma, lo que permite obtener información más precisa y detallada sobre los patrones de metilación en estas zonas conflictivas e inabarcables por lecturas cortas, entre otras, los LINE1. De esta forma, se confirma que la tecnología de secuenciación seleccionada no solo es válida, precisa y fiable, sino que también permite obtener información más completa y detallada sobre la metilación del genoma, lo que resulta fundamental para avanzar en la comprensión de los procesos moleculares involucrados en la regulación epigenética y activación de LINE1.

Nota: El análisis se realizó con el notebook

<../notebooks/0_vision_general/20230324_Correlation_bismark_megalodon_02.ipynb>

Este *script* importa los datos de metilación de ambos programas, obtenidos tras aplicar los flujos (5.2.4 Alineación de datos de metilación. Pretratamiento con bisulfito + Illumina + Bismark, 5.2.5 Alineación de datos de metilación. Lecturas largas obtenidas con ONT + Megalodon) sobre los datos de las líneas celulares. A continuación, los filtra, agrupa y realiza las gráficas y las correlaciones pertinentes que han sido mostradas en esta sección (correlación lineal de Pearson).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

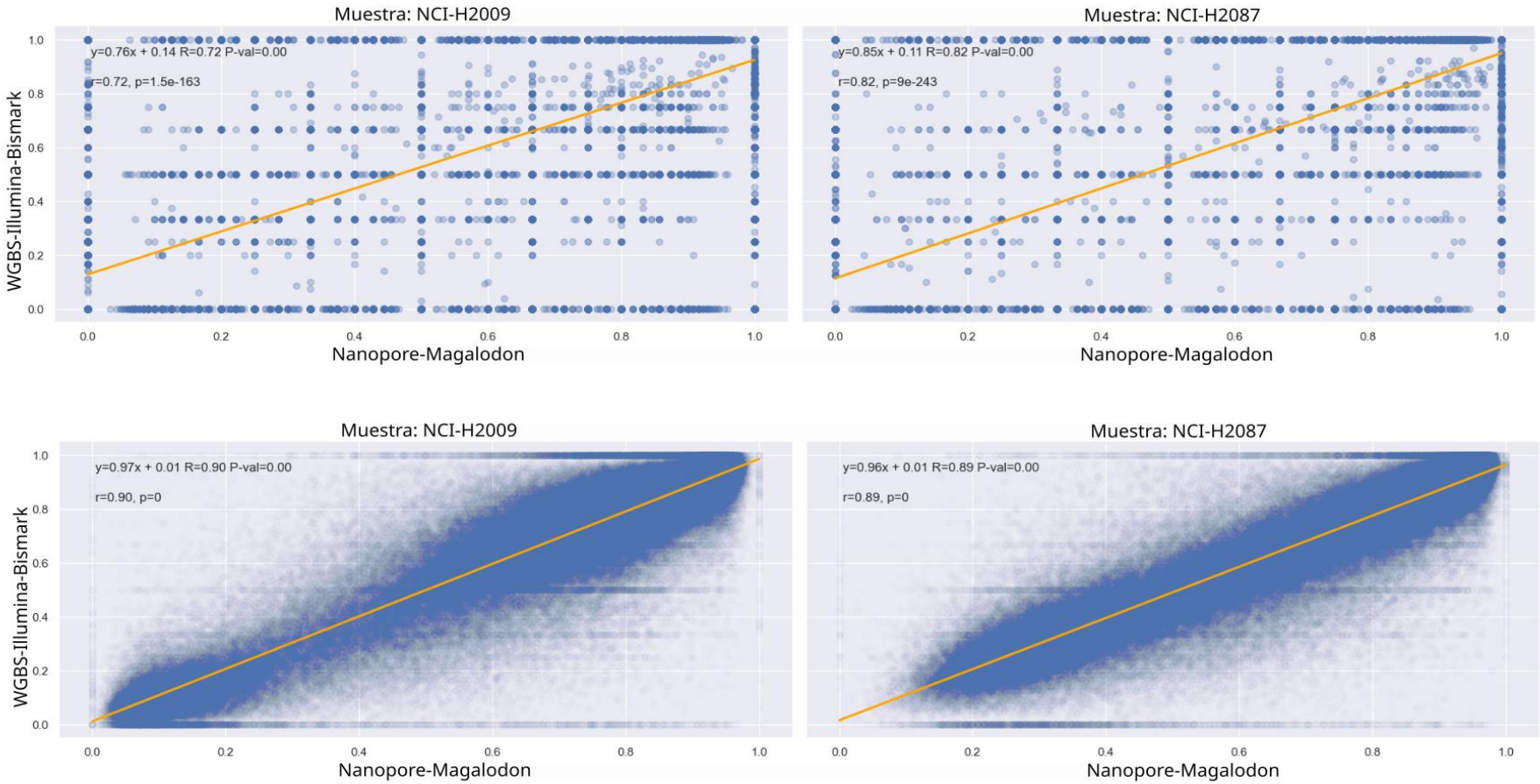


Figura 104: Correlación de la metilación de las líneas celulares.

Metilación Gold standard (bisulfito + Illumina + Bismark) comparado con la metilación obtenida por secuenciación de nanoporos + Megalodon. Panel superior: comparación de los valores medios de metilación en aquellas posiciones donde existe en ambos conjuntos de datos. Panel inferior: Metilación media en ventanas de 10,000 pares de bases.

6.4 COMPARATIVA DE MÉTODOS GRÁFICOS Y ESTADÍSTICOS DE METILACIÓN GENERAL

La visualización gráfica de datos de metilación es una herramienta fundamental para la comprensión y análisis de los patrones de metilación de los LINE1 en muestras tumorales y su tejido adyacente. En este sentido, resulta esencial la verificación de la fiabilidad de los datos obtenidos por el secuenciador MinION y el programa Megalodon, ya que esto nos permitirá abordar los siguientes análisis con mayor confianza y situar los resultados en un contexto general. Durante el desarrollo de la investigación, se utilizaron diversas herramientas de software disponibles para la comunidad de investigadores en el campo de la metilación. Sin embargo, pronto se evidenció la necesidad de desarrollar herramientas específicas para la tarea, que permitieran una visualización más precisa y detallada de los patrones de metilación de los LINE1. En esta sección se resumirá el uso que le hemos dado a las distintas herramientas gráficas disponibles, cómo las hemos adaptado para las necesidades específicas de este proyecto y por último la descripción de los programas para generar nuestras propias gráficas para satisfacer las necesidades específicas de este trabajo de investigación.

6.4.1 Methylartist

El programa Methylartist **(https://github.com/adamewing/methylartist)** es una herramienta específica para generar gráficas de metilación a partir de lecturas largas de Nanopore, lo que lo convierte en idóneo para nuestros datos. Para su funcionamiento se requiere tanto la información generada por Megalodon (u otros *basecallers*) como de un archivo de anotaciones genómicas de interés en las cuales queremos evaluar su estado de metilación en un juego de muestras determinado. En el estudio presente se emplearon dos ficheros con las anotaciones genómicas en formato *"bed"* sobre la versión GRCh38.p13 del genoma humano de referencia (hg38). El primero incluye un genoma segmentado en bloques de 6k y todos los LINE1 conocidos en el genoma humano clasificados en distintas categorías en base a criterios funcionales y estructurales de los mismos. Así, este archivo incluye entre otras, una clasificación funcional en L1_FLI (LINE1 completos íntegros), L1_FLNI (LINE1 completos no íntegros) L1_FLO (LINE1 con los dos marcos de lectura abierta) según la base de datos de L1base, la asignación de elementos con actividad conocida según el estudio de Pancancer (PCAWG) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium) (HOT) y las distintas categorías de LINE1 (L1HS, L1P1, L1P2, L1P3, L1P3b, L1P4, L1P4a, L1P4b, L1P4c, L1P4d, L1P4e, L1P5, L1PA2, L1PA3, L1PA4, L1PA5, L1PA6, L1PA7, L1PA8 y L1PA8A) consideradas por RepeatMasker y obtenidas del track del genome browser **(https://genome.ucsc.edu/)**. Esta última clasificación aplica a los LINE1 y se basa en su antigüedad y la presencia o ausencia de ciertos fragmentos de secuencias en el elemento LINE1 y en la estructura de sus ORFs (open reading frames) internos. L1HS es la categoría principal en nuestro estudio y se refiere a los elementos LINE1 completos y funcionales en seres humanos. Por contra, los L1PX (con X >1) son LINE1 surgidos en la evolución de primates y que han acumulado mutaciones que eliminan su funcionalidad en el genoma humano, dentro de la cual se distinguen varias subfamilias (L1P2 a L1P5 y L1PA2 a L1PA8A). La metodología seguida para obtener estos ficheros e integrarlos en un único archivo se describe en (7.2.3.2 Análisis inicial con datos obtenidos de la web de L1base y RepeatMasker).

El segundo archivo contiene coordenadas de diversas características funcionales del genoma humano, como regiones promotoras, genes, intrones, exones, regiones intergénicas y UTRs. Estos datos fueron obtenidos por un lado de la Eukaryotic Promoter Database **(https://epd.epfl.ch//index.php, última vez accedida en diciembre de 2022),** considerando una extensión de 500 pb corriente arriba del inicio del promotor y 1000 pb corriente abajo de dicha posición. Además, incluye información de la *"Release 42 de Gencode"* para GRCh38.p13.

(https://www.gencodegenes.org/human/, accedido por última vez en 26/10/2022)

Para generar este archivo, inicialmente se utilizó un *script* de Bash denominado "regions.sh" (del cual se muestra la versión para el *Gencode release* 41).

```
#!/bin/bash
#20211227 adaptaded by fjt from https://davetang.org/muse/2013/01/18/defining-
genomic-regions/
#20220825 updated version 41
v=41
#wget
"http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_""$v""/gencode.v
"$v".annotation.gtf.gz"

#get merged exons
cat gencode.v$v.annotation.gtf | awk 'BEGIN{OFS="\t";} $3=="exon" {print $1,$4-
1,$5,"exon"}' | sortBed | mergeBed -i - -c 4  -o distinct >
gencode_v${v}_exon_merged.bed

#intronic regions
cat gencode.v$v.annotation.gtf | awk 'BEGIN{OFS="\t";} $3=="gene" {print $1,$4-
1,$5,"intron"}' | sortBed | subtractBed -a stdin -b gencode_v${v}_exon_merged.bed
> gencode_v${v}_intron.bed

#let's intersect the two files
#this shouldn't produce any output
intersectBed -a gencode_v${v}_exon_merged.bed -b gencode_v${v}_intron.bed

#And finally to define intergenic regions, we use complement Bed to find regions
not covered by genes.

mysql --user=genome --host=genome-mysql.cse.ucsc.edu -A -e \
"select chrom, size, 'intergenic' from hg38.chromInfo order by chrom"  >
hg38.genome

cat gencode.v$v.annotation.gtf |awk 'BEGIN{OFS="\t";} $3=="gene" {print $1,$4-
1,$5,"intergenic"}' |sortBed |complementBed -i stdin -g hg38.genome >
gencode_v${v}_intergenic.bed

cat gencode_v${v}_intergenic.bed |awk 'BEGIN{OFS="\t";} {print $0,"intergenic"}' >
gencode_v${v}_intergenic_complete.bed

cat gencode.v"$v".annotation.gtf | awk '($3=="UTR")
{printf("%s\t%s\t%s\t%s\n",$1,int($4)-1,$5,"UTR");}' |sort -T . -k1,1 -k2,2n |
bedtools merge -c 4  -o distinct > gencode_v"$v"_UTR.bed
cat gencode.v"$v".annotation.gtf | awk '($3=="gene")
{printf("%s\t%s\t%s\t%s\n",$1,int($4)-1,$5,"gene");}' |sort -T . -k1,1 -k2,2n |
bedtools merge -c 4  -o distinct> gencode_v"$v"_genes.bed

cat gencode_v"$v"_exon_merged.bed  gencode_v"$v"_genes.bed
gencode_v"$v"_intergenic_complete.bed  gencode_v"$v"_intron.bed
gencode_v"$v"_UTR.bed | sort -T . -k1,1 -k2,2n > gencode_v"$v"_all_features.bed
```

Código 7: Ejemplo de código para la obtención de regiones genómicas (regions.sh)

En iteraciones posteriores, se empleó un *script* de Python incluido en el notebook de Python:

<../notebooks/00_tools/get_regions_hg38_00.ipynb>

el cual se encuentra en el repositorio adjunto a esta Tesis. Este último *script* descarga directamente las anotaciones funcionales del proyecto Gencode en su versión GRCh38.p13 (https://www.gencodegenes.org/human/release_42.html, obtenida en diciembre de 2022). A partir de estos datos, el *script* genera un archivo que contiene las regiones funcionales del genoma, cuyos patrones de metilación resultan de interés para caracterizar en las muestras de nuestra cohorte. El *script* de Python proporciona simplicidad, rapidez en el desarrollo y la capacidad de visualizar la información de forma directa y de procesamiento de datos. Además el programa es capaz de agrupar segmentos y anotaciones solapantes, de tal forma que una misma región no esté representada varias veces. Por ejemplo, en la figura 105 se muestra cómo se compactan los segmentos parcial o totalmente solapantes en una región de 120,000 pares de bases, donde se agrupan más de 50 segmentos solapantes en 6 grupos distintos.

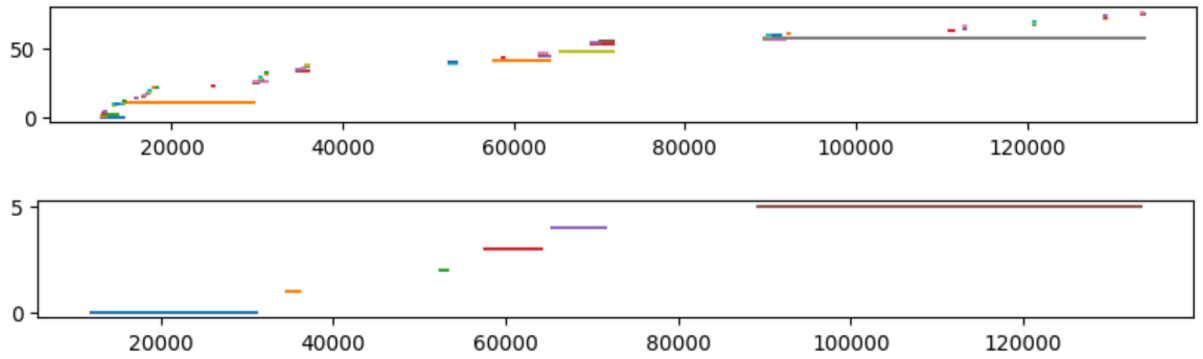


Figura 105: Ejemplo de compactación de segmentos genómicos.

En sus versiones iniciales Methylartist se ejecutaba en dos pasos sucesivos. En el primero se creaba una base de datos del programa a partir de la información de Megalodon *(methylartist segmeth).* En el segundo paso generaban los gráficos tipo violín *(methylartist segplot …. -v).* En sucesivas versiones del programa es posible generar los gráficos sin la necesidad de recrear la base de datos intermedia, si bien ello limita la versatilidad del programa para realizar gráficos compuestos por varias muestras y requiere de la ejecución de ambos pasos para reconstruir todas las figuras con la consiguiente pérdida de tiempo. A fin de ejecutarlo en el CESGA se creó el siguiente *script* (launch_methylartist_segmeth.sh).

```
#!/bin/bash
#-----------------------------------------------------------------
#Launch methylartist
#
# FT-III system.
#
# This script requests one dedicated core on one shared node
#-----------------------------------------------------------------
#SBATCH -J S_Me-Me           # Job name
#SBATCH -c 32                     # Total number of tasks
#SBATCH -o 09_logs/m_artists.%j.out # Specify stdout output file (%j expands to
jobId)
#SBATCH -t 10:00:00              # Run time (hh:mm:ss) - 1.5 hours
#SBATCH --mem=64G                  # Total memory demanded - 20 GB RAM
```

```
# Load any necessary modules
# Loading modules in the script ensures a consistent environment.
module load cesga/2020

source /mnt/netapp2/uscmg_aplic/3_environments/python/01_ONT/bin/activate

# check the hostname
hostname
sample=$1

base_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_javi/2
_meme/04_Megalodon_WG/01_methylartist/

mkdir -p "$base_dir""$sample"/
cd "$base_dir""$sample"/

echo processing "$base_dir""$sample"/

time  methylartist segmeth -i
"$base_dir"/hg38_features_6kb_bins_L1base_no_tab_chr_notation_L1_recent.bed -b
$base_dir/../"$sample"/"$sample"_Megalodon_WG/mod_mappings.sorted.bam  --ref
/mnt/netapp2/genomas_enfermedad/1/0_reference/0_hg38/GCA_000001405.15_GRCh38_no_al
t_analysis_set.fna --motif CG -p 32

time  methylartist segmeth -i "$base_dir"/gencode_v41_all_features_plus_FL-L1.bed
-b $base_dir/../"$sample"/"$sample"_Megalodon_WG/mod_mappings.sorted.bam  --ref
/mnt/netapp2/genomas_enfermedad/1/0_reference/0_hg38/GCA_000001405.15_GRCh38_no_al
t_analysis_set.fna --motif CG -p 32

time methylartist segplot -s
hg38_features_6kb_bins_L1base_no_tab_chr_notation_L1_recent.mod_mappings.sorted.se
gmeth.tsv -v
                           time methylartist segplot -s
hg38_features_6kb_bins_L1base_no_tab_chr_notation_L1_recent.mod_mappings.sorted.se
gmeth.tsv -v --svg

time methylartist segplot -s gencode_v41_all_features_plus_FL-
L1.mod_mappings.sorted.segmeth.tsv -v
time methylartist segplot -s gencode_v41_all_features_plus_FL-
L1.mod_mappings.sorted.segmeth.tsv -v --svg

##hg38_features_100kb_bins_L1base.bed
#hg38_features_6kb_bins_L1base_no_tab_chr_notation.bed
#hg38_features_6kb_bins_L1base_no_tab_chr_notation_L1_recent.bed
#hg38_features_6kb_bins_L1base_no_tab_chr_notation_L1_recent_small.bed
#L1_recent.bed
```

Código 8: Script para el lanzamiento de Methylartist.

La herramienta utilizada para la recreación de gráficos de metilación por característica genómica muestra su versatilidad en la generación de visualizaciones precisas y detalladas de los patrones de metilación en diferentes tipos de muestras biológicas. Para ilustrarlo, se presentan a continuación dos ejemplos de gráficas de violín correspondientes a una muestra de tumor de cabeza y cuello PD0277a y a su tejido adyacente PD0277c. La figura 106 muestra los patrones de metilación en regiones funcionales específicas del genoma, lo que permite evaluar la metilación en áreas de interés particular, como las regiones promotoras de los genes o los exones. Por otro lado, la figura 107 muestra la metilación en las distintas categorías funcionales de LINE1, lo que permite analizar el estado de metilación de estos elementos transponibles en diferentes contextos genómicos. Cabe destacar que estas gráficas

son solo dos ejemplos de la gran cantidad de información visual que se puede obtener a partir de esta herramienta. El repositorio adjunto contiene una variedad de gráficos adicionales que proporcionan una representación más completa y detallada de los patrones de metilación en diferentes tipos de muestras biológicas y contextos genómicos. En resumen, Methylartist es una herramienta versátil y muy útil para analizar, visualizar y comparar patrones de metilación en diversas muestras biológicas.

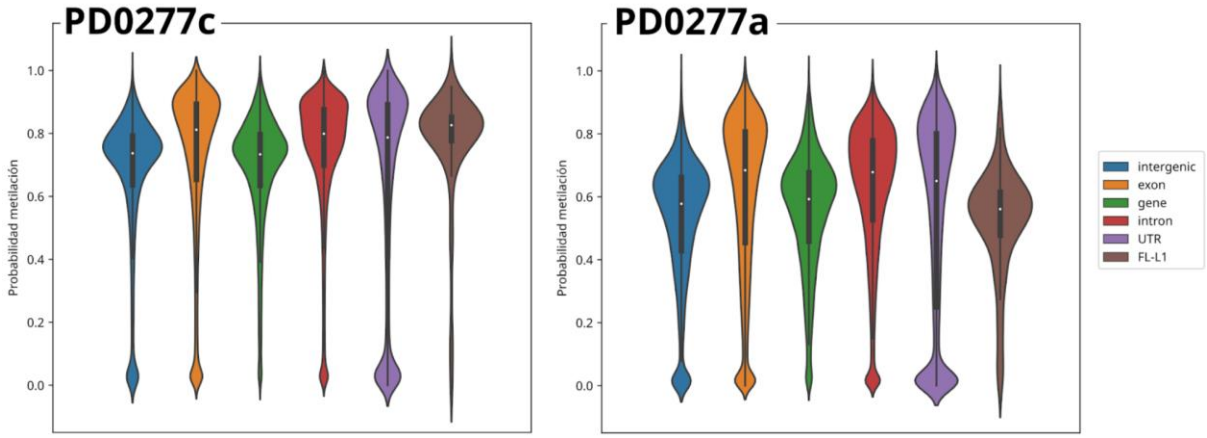


Figura 106: Gráfica de violín producida por Methylartist de la Metilación en regiones funcionales. (PD0277c tejido adyacente y PD0277a tumor)

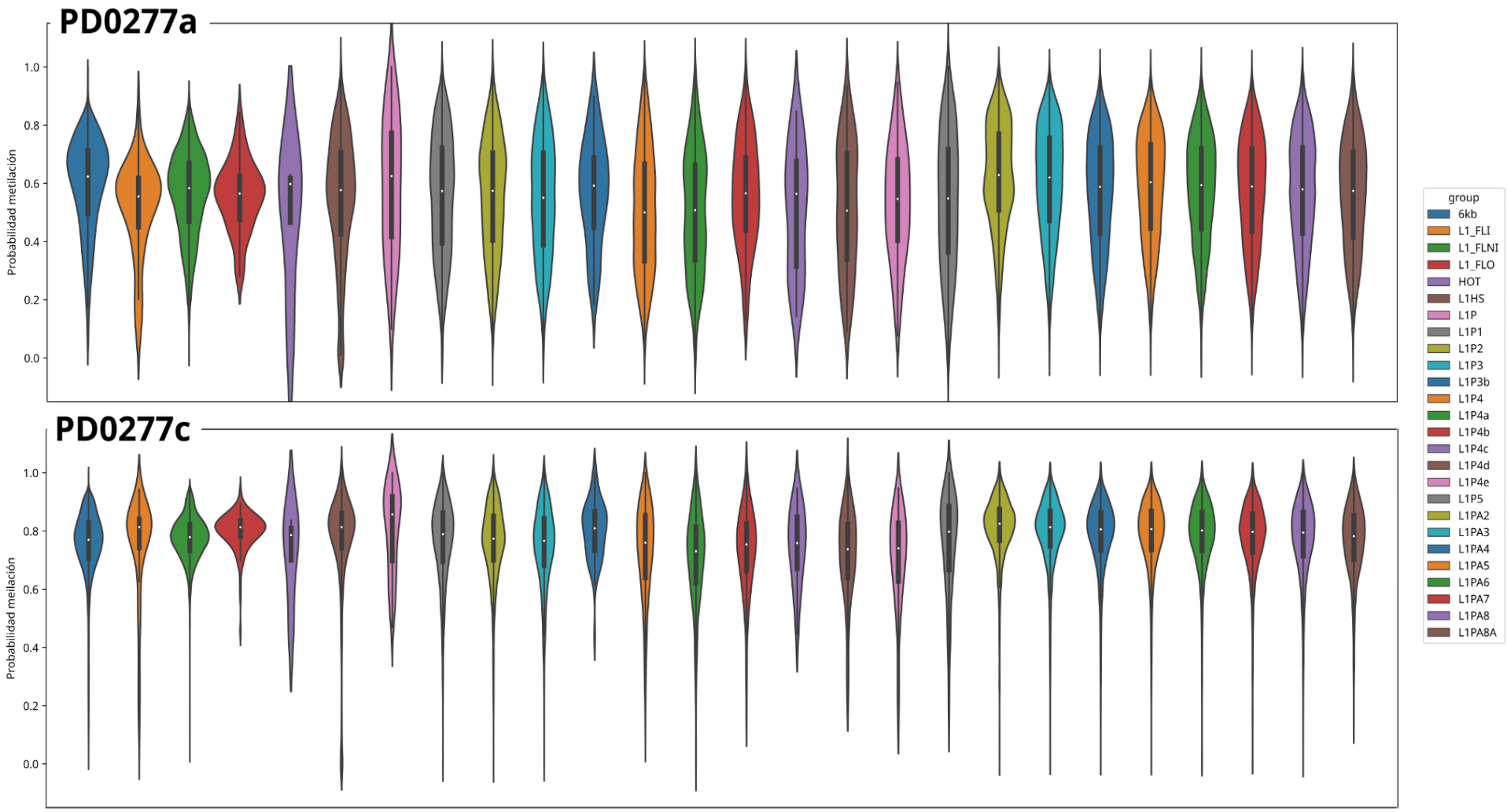


Figura 107: Gráfica de violín producida por Methylartist de la metilación en diversos LINE1 I. (PD0277a tumor y PD0277c tejido adyacente)

En términos generales, los resultados del análisis de metilación muestran que el tejido normal adyacente (PD0277c) presenta un nivel de metilación promedio mayor que su contraparte tumoral (PD0277a). Esta diferencia es más pronunciada en los elementos LINE1 que en otras regiones funcionales del genoma. Estos análisis están en plena concordancia con los obtenidos utilizando los programas, gráficas y estadísticas creado específicamente por nosotros para este cometido y son descritas en el apartado (8.1 Metilación general de las muestras).

Curiosamente, la línea celular NCI-BL2087, presenta un estado de metilación global mucho menor que el que su correspondiente pareja tumoral NCI-H2087 (Figura 108). Este mismo patrón opuesto también se observa en la otra línea celular NCI-H2009 y su correspondiente sangre NCI-BL2009, en aparente oposición a lo observado en las parejas tumor primario-tejido adyacente estudiadas en esta Tesis doctoral (Figura 123) y que se trata con más detalle en el apartado 8.1.1.1 metilación de las líneas celulares.

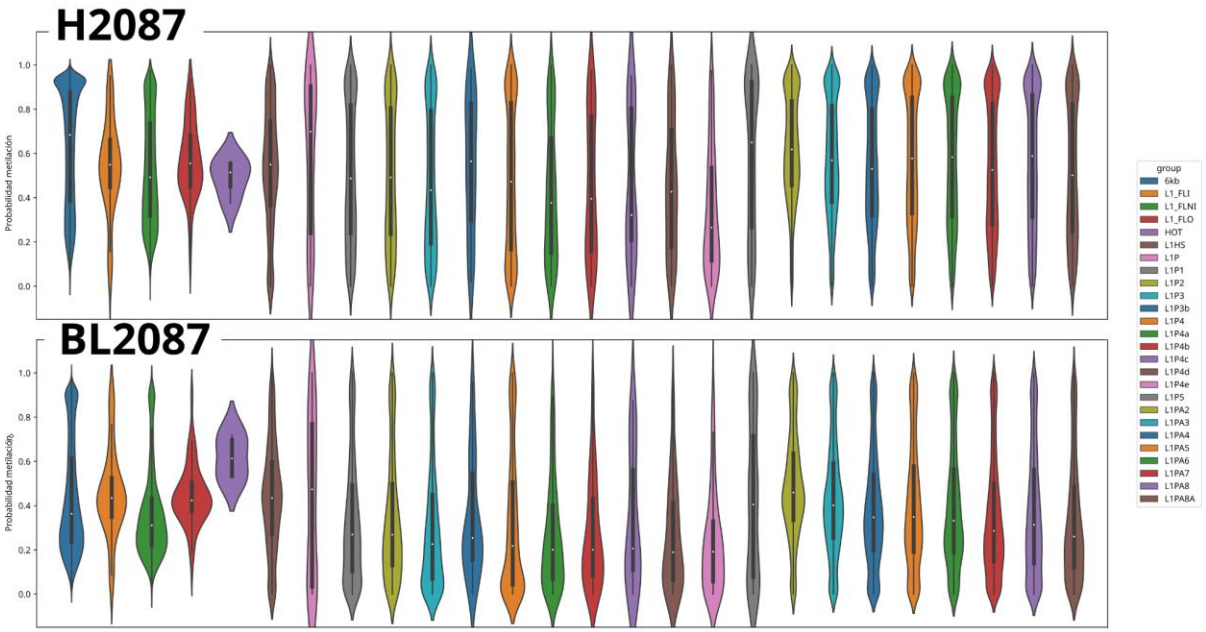


Figura 108: Gráfica de violín producida por Methylartist de la metilación en diversos LINE1 II. (NCI-H2087 y NCI-BL2087)

Pese a las innegables virtudes de Methylartist para la representación gráfica de este tipo de datos, en última instancia se limita a utilizar los resultados de apilamiento de metilación por posición de Megalodon, lo cual supone un problema en el caso concreto de regiones repetitivas y, por extensión en los LINE1. Al tratar de analizar los patrones de metilación de los LINE1, nos dimos cuenta de que dada su similitud, repetitividad y abundancia en el genoma en un mismo loci se apilaban equívocamente lecturas correspondientes a LINE1 de otras posiciones genómicas e, inclusive de otras categorías. Por ello, decidimos centrarnos en el análisis de los LINE1 íntegros (L1HS), obviando las demás familias de LINE1. Este apilamiento equívoco supone un considerable problema, ya que estas lecturas provenientes de distintos elementos fuente pueden presentar patrones de metilación muy variables y con ello alterar la evaluación de sus verdaderos perfiles epigenéticos. **Para garantizar un filtrado exhaustivo de las lecturas y evitar problemas de alineamiento, desarrollamos el programa Methyl-Meiga**, el cual garantiza que cada lectura pertenezca

únicamente a un LINE1 concreto y cuyo funcionamiento describiremos en apartados posteriores. Además, con el fin de tener un control más detallado y un conocimiento profundo de los datos, creamos un conjunto de *scripts* de Python para analizar, visualizar y generar estadísticas de los resultados. Es importante mencionar que uno de los objetivos secundarios de este trabajo fue aprender el funcionamiento de este tipo de programas y análisis, realizándolos personalmente. En todo caso ha de destacarse que en versiones posteriores del programa Methylartist se introdujeron nuevas funcionalidades para garantizar que los alineamientos de las lecturas en los LINE1 fueran unívocos, descartando aquellas lecturas con mapeos suplementarios que se anclaran simultáneamente a varias posiciones del genoma.

6.4.2 Comparativa de *scripts* gráficos ad hoc creados por mí.

En vista de las limitaciones de los programas de recreación gráfica a partir de datos de metilación de ONT sucintamente mencionadas en el apartado anterior y a fin de tener un mayor control y un conocimiento más profundo de los datos opté por diseñar mis propios programas para obtener análisis y gráficas. El objetivo era el de obtener un programa intuitivo y fácil de emplear y que permitiese caracterizar el estado de metilación global y el de ciertas características funcionales del genoma de una muestra o grupo de muestras, por ejemplo, los UTR, las regiones promotoras de los genes, los exones y las regiones codificantes (CDS) a partir de los archivos de salida de Megalodon. Además, se deseaba incluir funcionalidades para establecer el estado de metilación por haplotipo, lo que requería el archivo de metilación por lectura individual en lugar del archivo de apilamiento de datos de metilación por posición utilizado por programas como Methylartist.

Una de las funcionalidades que proporciona Megalodon es la estimación de la probabilidad de metilación (en tanto por uno) para cada citosina en un dinucleótido CpG presente en cada una de las lecturas de Nanopore. Es decir, el programa calcula la probabilidad de que una citosina esté bien metilada o hidroximetilada para cada lectura individual generada en la secuenciación, lo que permite obtener una visión detallada del estado de metilación de cada sitio a nivel de una sola molécula de ADN. Si bien este archivo de salida es idóneo para discriminar el estado de metilación por haplotipos, también genera un dilema sobre qué estrategia de procesamiento de datos de metilación había de seguirse: Si emplear el rango continuo de valores de probabilidad de modificación epigenética o si estos han de ser binarizados (0=no metilado; 1=metilado). Se hacía necesario pues hacer una serie de pruebas que marcaron el procedimiento a seguir para posteriores análisis, estas fueron:

1. Hacer los cálculos y gráficas usando los valores medios de probabilidad por cada posición CpG en todas las lecturas (agrupando los valores por posición).
    a. Media de todas las CpG. En la gráfica cada punto es la probabilidad de metilación de cada una de las posiciones CpG.
    b. Media ponderada de las medias de la probabilidad de metilación en ventanas de 10 kilo bases. En la gráfica cada punto es la media de cada ventana.
    c. Media ponderada de las medias de la metilación en ventanas de 50 CpG. En la gráfica cada punto es la media de cada ventana.
2. Binarizar los datos estableciendo un umbral de metilación ($p > 0.7$) y de no metilación ($p < 0.3$), asignando 1 o 0 respectivamente, y descartando los valores de probabilidad entre 0.3 y 0.7. En este caso, también se exploraron varias opciones:

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

a. Porcentaje del número de CpG metiladas respecto a todas las CpG (metiladas y no metiladas).
b. Porcentaje de las CpG metiladas respecto a todas las CpG (metiladas y no metiladas) por **lectura.** Media ponderada de las medias por lectura. En la gráfica cada punto es cada uno de los porcentajes por lectura.
c. Porcentaje de las CpG metiladas respecto a todas las CpG (metiladas y no metiladas) en ventanas de **10 Kilo bases**, en el análisis global, y por segmento de cada característica genómica, o LINE1 en sus respectivos análisis. Media (ponderada) de las medias por ventana o segmento. En la gráfica cada punto es cada uno de los porcentajes por ventana o segmento.

Ambas opciones tienen sus ventajas y desventajas y aportan información interesante/complementaria. Para los análisis iniciales nos centramos en la muestra tumoral PD0277a y su correspondiente tejido adyacente PD0277c.

6.4.2.1 Análisis usando datos de probabilidad por cada CpG

Para el análisis de los datos, se probaron tres opciones distintas. La opción 1a (figura 109 panel superior) computa la probabilidad promedio de metilación de todas las posiciones para generar las gráficas de manera sencilla y directa. Para ello, se calculó la media por posición de todas las lecturas que contienen dicha posición. Esta aproximación fue descartada por albergar un gran número de mediciones que incrementan sensiblemente el tiempo computacional, lo que lo hace poco práctico para obtener figuras de forma prospectiva.

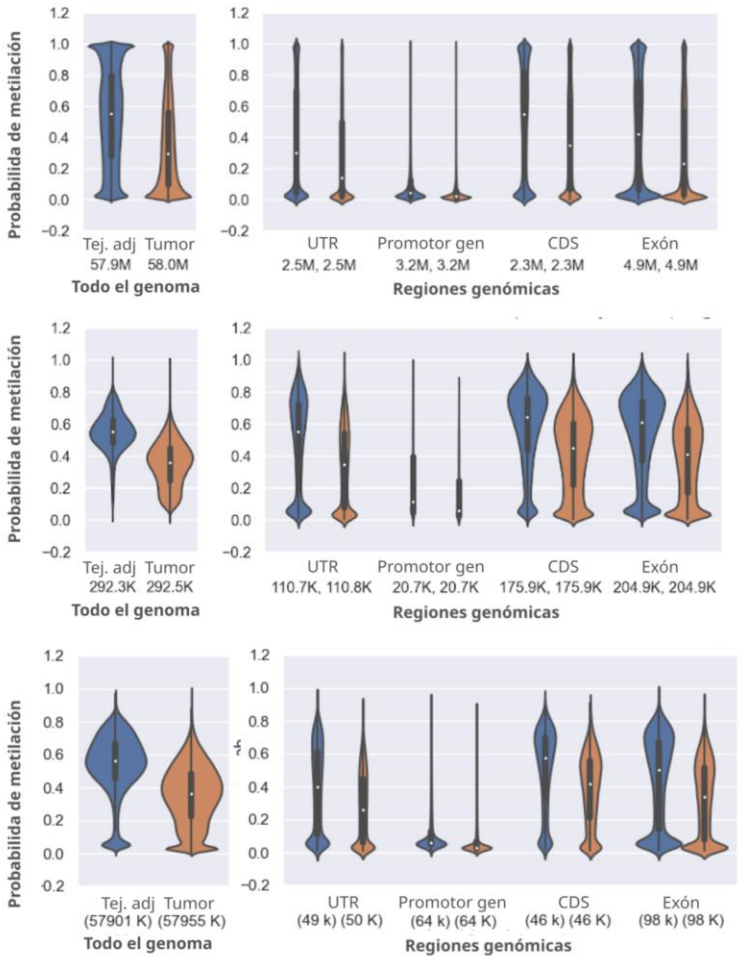


Figura 109: Pruebas de diferentes formas de representar la metilación con gráficas de violín.

En contraste con la opción anterior, la segunda aproximación 1b calcula las medias de probabilidad agrupándolas en ventanas de 10 Kb (figura 109, panel central). Si bien este enfoque tiene la ventaja de reducir el número de puntos a manejar, se debe tener en cuenta que puede presentar un sesgo basado en la densidad de CpGs a lo largo del genoma, lo que puede resultar en una gran oscilación con respecto al valor real de metilación en regiones deplecionadas de CpGs, como la heterocromatina constitutiva. En este sentido, la alternativa 1c, que calcula el promedio de

metilación en ventanas de 50 CpGs consecutivas, se presentaba como un buen compromiso entre las dos anteriores (figura 109 panel inferior) para obtener una visión más completa del conjunto de datos.

En las figuras anteriores los valores numéricos bajo cada uno de los gráficos de violín hacen referencia a la cantidad de puntos que contienen cada gráfica, ya sea el número de posiciones CpG o el de ventanas de las que se hace el promedio de metilación de las lecturas contenidas en las mismas. Para el procesamiento de todas estas gráficas y cálculos se utilizó el *script* <***..notebooks*/0_vision_general/get_data_from_megalodon_per_read_08.ipynb**>. Si bien estas aproximaciones fueron suficientes para realizar exploraciones preliminares de los datos de ONT, fueron descartadas en la generación de las gráficas finales, ya que las estadísticas de datos binarizados ofrecen una visión más concisa y con mucho menos ruido.

6.4.2.2 Análisis usando datos binarizados por cada CpG

En la segunda alternativa de representación gráfica (opción 2), se implementó una estrategia de binarización de los datos de probabilidad de modificación epigenética para cada CpG y lectura, ya sea metilación o hidroximetilación. En concreto, se asignó el valor 0 (no metilada) a aquellas CpGs con una probabilidad menor a 0.3, mientras que se les asignó el valor 1 a aquellas con una probabilidad mayor a 0.7 (ya sea metilada o hidroximetilada). Aquellas CpGs con una probabilidad de modificación entre ambos umbrales (0.3 y 0.7) fueron excluidas de la computación de la media de metilación, hidroximetilación.

Dentro de las opciones consideradas, la aproximación 2b de calcular la media de metilación por lectura inicialmente parecía ser la más adecuada. Durante el análisis preliminar de las inserciones somáticas, se identificó un patrón prometedor que será discutido en el apartado (8.4.1 Análisis de los patrones de metilación de inserciones somáticas de LINE1 por lectura). En teoría, esta estrategia permitiría detectar e incluso distinguir las lecturas asignables al tumor de la contaminación de células "normales" del tejido sano adyacente, basándose en los cambios globales de hipometilación atribuibles al tejido tumoral. Sin embargo, esta aproximación presenta un inconveniente en la generación de los gráficos de violín, ya que se observan ondulaciones bruscas en los bordes (ver figura 111, izquierda). Tras una investigación detallada, se descubrió que esto se debe al hecho de que las lecturas tienen un número fijo y limitado de sitios CpG (en promedio entre 50 y 100), así como un número reducido de observaciones. Como consecuencia, al calcular las medias de metilación, se obtiene una distribución discreta y no continua de valores decimales exactos, lo que afecta la continuidad de los gráficos de violín. Esta discretización se puede apreciar claramente en un histograma de los datos (ver figura 111, derecha). Aunque es posible mitigar este efecto reduciendo el suavizado de la distribución mediante el parámetro bw, como se muestra en la figura 111, centro. Esto modifica ligeramente la curva respecto a su distribución real. Dado que esta discretización genera curvas inusuales y no aporta información relevante en relación a la contaminación celular, se optó por utilizar un enfoque más sencillo, similar al empleado en estudios anteriores como el de (Ewing et al. 2020). En este nuevo enfoque, se calcula el porcentaje de CpG metiladas para los datos de metilación global promedio, y se divide el genoma en ventanas de 10 Kb para generar los gráficos y estadísticas correspondientes (opciones 2a y 2c). Como se puede observar en la figura 112, los gráficos resultantes son

mucho más suaves y continuos, sin los problemas de discretización mencionados anteriormente.

Es importante enfatizar que es imprescindible ponderar las medias en función de la cobertura para cada una de las posiciones CpG. Después de realizar los análisis anteriores, se observó que las medias de metilación de todas las CpGs difieren de las obtenidas al calcular las medias de las ventanas, ya sean de 10 Kb o de 50 CpGs consecutivas. Tras estas diferencias, en apariencia mínimas, subyace un error conceptual de gran magnitud, el que la media de las medias parciales de un conjunto de datos es siempre igual a la media de todos los datos (ver ecuación de la figura 110) De hecho, la estimación de la media de la metilación sólo sería correcta si las medias parciales tuviesen la misma representación en número de CpGs. Esto es especialmente preocupante en tumores, más proclives a mostrar CNVs en su genoma que puedan sesgar la estimación de la media de regiones sub o sobrerrepresentadas. De ahí que sea imprescindible ponderar cada media en función de su representatividad en el tamaño muestral global.

$$\mu_{\overline{x}} \neq \frac{\overline{x}_1 + \overline{x}_2 + \overline{x}_3 \ldots + \overline{x}_n}{n}$$

Figura 110: Media de las medias

Una vez tenido esto en cuenta las medias de metilación generales de las muestras coincidían con las medias (ponderadas) de las medias de las muestras por ventanas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

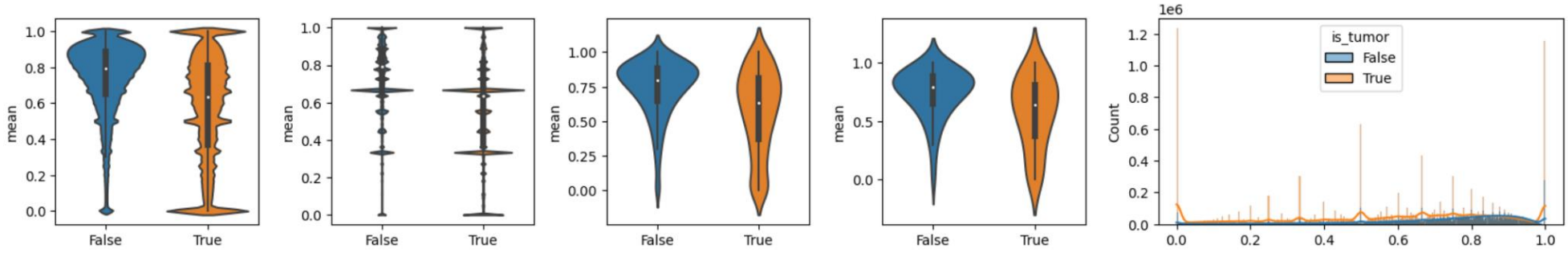


Figura 111: Pruebas de suavizado de las gráficas de violín. Datos de metilación binarizados.

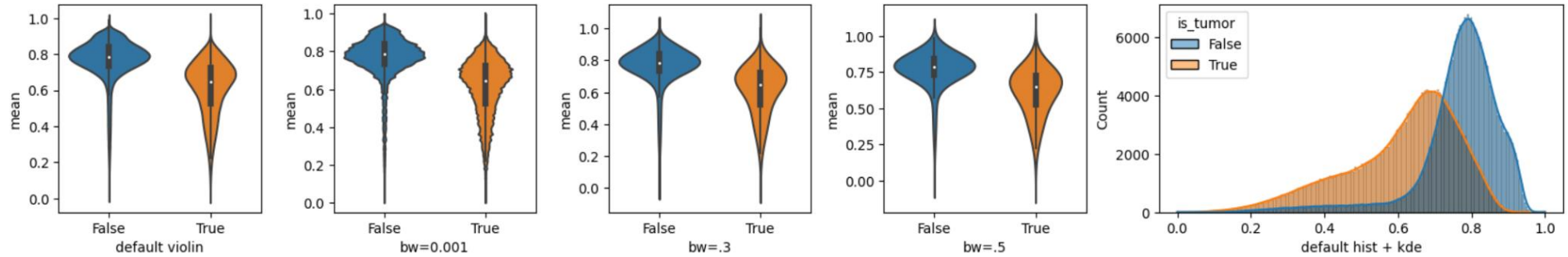


Figura 112: Pruebas de suavizado de las gráficas de violín. Datos de metilación binarizados, media en ventanas de 10K pb.

6.4.2.3 Descripción del método usado para generar gráficas y estadísticas

En base a las pruebas descritas en los apartados anteriores se optó por una estrategia de construcción de estadísticas y gráficas por binarización de los datos de metilación por lectura proporcionados por Megalodon. El *script*, íntegramente redactado en Python en un *notebook* de Jupyter, es demasiado grande para reproducirlo de forma íntegra en este manuscrito o incluirlo como anexo, por lo que está depositado en el repositorio adjunto junto con los restantes *scripts* manejados en esta Tesis doctoral.

<../notebooks/0_vision_general/20230221_vision_general_metilacion_binary_win_10k_04.ipynb>.

A continuación, se describen las secciones que lo componen y su funcionamiento general.

- Descripción del *script*, carga de librerías y definición de constantes.
- Sección con las funciones principales.
- Carga de los datos de metilación necesarios para el análisis, que incluyen:
    - Datos de metilación del genoma completo proporcionados por Megalodon "*Megalodon_per_read_per_position.txt*".
    - Datos de metilación de los elementos LINE1 germinales desde el archivo generado por "*Methyl-Meiga*".
    - Datos de metilación de los elementos LINE1 somáticos desde el archivo generado por "*Methyl-Meiga*" a partir de las calls somáticas obtenidas por "*MEIGA-LR*".
- *Reformateo de datos*. Se crea una tabla de datos en formato largo que contiene la metilación de las CpGs desglosadas según las características funcionales del genoma en las que se ubican. Los valores de metilación proporcionados por Megalodon se expresan como la probabilidad transformada como logaritmo en base 2 de metilación para cada CpG en cada lectura de ONT. El *script* binariza estos datos en base a los criterios antes expuestos en 0 o "no metilada" si la probabilidad es menor a 0.3, y 1 o "metilada" si la probabilidad es mayor a 0.7. Los datos que se encuentran entre estos umbrales se descartan.
- *Creación de Gráficas*. Cada punto en los gráficos de violín corresponde a la media por ventana de 10 Kb. Se generan gráficas para el genoma completo, las características funcionales y los elementos LINE1 germinales e inserciones somáticas proporcionados y se guardan en formato .png y .svg.

Análisis estadístico. Se calculan las medias a partir del porcentaje de metilación en las regiones de interés, ya sea todo el genoma, en las CpG contenidas en las distintas regiones funcionales (UTR, exón, intrón, intergénica…), y las CpG de los elementos LINE1 germinales y las inserciones somáticas derivadas de los mismos. Se realizan las comprobaciones pertinentes de normalidad y homocedasticidad de los datos con, respectivamente, un test de Shapiro-Wilk y un test de Levene. Finalmente, se realizan pruebas estadísticas de comparación de distribuciones de datos entre muestras tumorales y tejido adyacente. Se aplican tanto el estadístico T de Student (con corrección de Welch en caso de que los conjuntos a comparar tengan diferentes tamaños) como el estadístico U de Mann-Whitney-Wilcoxon, ambos corregidos mediante el método de Bonferroni. Estas pruebas se realizan en conjuntos de datos con n >= 3. Para conjuntos con n > 5000, se utiliza un subconjunto de 5000 elementos debido a limitaciones de las funciones empleadas.

7. RESULTADOS Y DISCUSIÓN II. MUESTRAS Y DATOS

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Imagen portada del capítulo: “Interpretación libre de una tinción eosina-hematoxilina de cáncer de pulmón” Acuarela sobre cartulina (feb 2023)

7.1 MUESTRAS

Las muestras usadas en este trabajo son tumores primarios de seres humanos, sus correspondientes tejidos normales y dos líneas celulares obtenidas también a partir de células humanas.

Las líneas celulares son de cáncer de pulmón NCI-H2087, NCI-H2009 con su correspondiente tejido sano, este caso sangre NCI-BL2087 y NCI-BL2009. Estas líneas celulares fueron seleccionadas por que contienen un alto número de retrotransposiciones y son descritas con más detalle más adelante.

7.1.1 Tumores primarios y su tejido adyacente

Los tumores primarios y sus correspondientes tejidos adyacentes (o en caso de no ser posible sangre) forman parte de dos conjuntos de muestras solicitadas por la estudiante de doctorado Sonia Zumalave. Las muestras proceden del FIDIS y del biobanco del País Vasco biobanko (https://www.biobancovasco.org/). Ambas cuentan con el permiso del CEIC. (The impact of L1-mediated somatic rearrangements in the origin and development of human cancer" promovido por José Manuel Castro Tubío, con código de registro 2018/586, y evaluado por el Comité el 19/12/2018) y fueron obtenidas con el consentimiento expreso de los pacientes.

Las muestras consisten en 150 tumores. De ellos 50 corresponden a carcinomas de células escamosas de pulmón, 50 a carcinomas de células escamosas de cabeza y cuello y 50 a adenocarcinomas de colon.

El ADN de estas muestras fue extraído por los técnicos de laboratorio, bajo la supervisión de Daniel García Souto (los protocolos usados se describen en los apartados 5.1.3 y siguientes).

Estas muestras se mandaron a secuenciar a **Macrogen (Corea del Sur https://dna.macrogen.com/)**. Se secuenciaron mediante la tecnología de "Illumina Pair End" con una cobertura media de 8x y una longitud de 150 pares de bases. Estos datos fueron usados para realizar una selección de las muestras con más actividad de retrotransposición. Esta selección se realizó usando el programa Meiga-SR (retrotest**)** (pendiente de publicación), específicamente desarrollado para la identificación de actividad en muestras tumorales usando lecturas cortas. Este programa ha sido desarrollado en nuestro grupo de investigación por la estudiante de doctorado Sonia Zumalave.

De las 150 muestras tumorales se seleccionaron 14. De diez de ellas (10) también se seleccionó su correspondiente tejido normal adyacente sano. La relación de muestras atendiendo al tipo de tumor es la siguiente: 5 corresponden a carcinomas de células escamosas de pulmón, 6 a carcinomas de células escamosas de cabeza y cuello y 3 a adenocarcinoma de colon. Los datos de otras 9 muestras de adenocarcinoma de esófago y 2 muestras de sangre de dos pacientes fueron obtenidos de una colaboración con el laboratorio de la doctora Rebecca Fitzgerald del departamento de oncología de la Universidad de Cambridge

(https://www.oncology.cam.ac.uk/directory/r-fitzgerald accedido 09.06.2022)

Se muestran a continuación su códigos, tipología y el tipo de secuenciación usado para el filtrado de las inserciones somáticas.

Identificación de la muestra	Tipo	Grupo	Filtrado
LP6005334-DNA_A03	Adenocarcinoma de esófago	Tumor	P6005333-DNA_A03-Illumina
LP6005935-DNA_C04	Adenocarcinoma de esófago	Tumor	P6005934-DNA_C04-Illumina
LP6008141-DNA_H01	Adenocarcinoma de esófago	Tumor	P6008138-DNA_H01-Illumina
LP6008221-DNA_E01	Adenocarcinoma de esófago	Tumor	P6008220-DNA_E01-Illumina
LP6008221-DNA_H02	Adenocarcinoma de esófago	Tumor	P6008220-DNA_H02-Illumina
LP6008269-DNA_A09	Adenocarcinoma de esófago	Tumor	P6008268-DNA_A09-ONT
LP6008269-DNA_B10	Adenocarcinoma de esófago	Tumor	P6008268-DNA_B10-Illumina
LP6008269-DNA_D08	Adenocarcinoma de esófago	Tumor	LP6008268-DNA_D08-ONT
LP6008334-DNA_H02	Adenocarcinoma de esófago	Tumor	LP6008335-DNA_D0- Illumina
PD0341a	Adenocarcinoma de colon	Tumor	PD0341c-Illumina
PD0351a	Adenocarcinoma de colon	Tumor	PD0351c-ONT
PD0353a	Adenocarcinoma de colon	Tumor	PD0353a-Illumina
12T_FIDIS	Carcinoma de células escamosas de cabeza y cuello	Tumor	12B_FIDIS-Illumina
PD0266a	Carcinoma de células escamosas de cabeza y cuello	Tumor	PD0266c-ONT
PD0270a	Carcinoma de células escamosas de cabeza y cuello	Tumor	PD0270c-ONT
PD0274a	Carcinoma de células escamosas de cabeza y cuello	Tumor	PD0274c-ONT
PD0277a	Carcinoma de células escamosas de cabeza y cuello	Tumor	PD0277c-ONT
PD0287a	Carcinoma de células escamosas de cabeza y cuello	Tumor	PD0287c-ONT
PD0299a	Carcinoma de células escamosas de pulmón	Tumor	PD0299a-Illumina
PD0307a	Carcinoma de células escamosas de pulmón	Tumor	PD0307c-ONT
PD0312a	Carcinoma de células escamosas de pulmón	Tumor	PD0312c- ONT
PD0319a	Carcinoma de células escamosas de pulmón	Tumor	PD0319c- ONT
PD0331a	Carcinoma de células escamosas de pulmón	Tumor	PD0331c- ONT
LP6008268-DNA_A09	Sangre	Sangre	
LP6008268-DNA_D08	Sangre	Sangre	
PD0351c	Tejido colon	Tejido adyacente	
PD0266c	Células escamosas de cabeza y cuello	Tejido adyacente	
PD0270c	Células escamosas de cabeza y cuello	Tejido adyacente	
PD0274c	Células escamosas de cabeza y cuello	Tejido adyacente	

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Identificación de la muestra	Tipo	Grupo	Filtrado
PD0277c	Células escamosas de cabeza y cuello	Tejido adyacente	
PD0287c	Células escamosas de cabeza y cuello	Tejido adyacente	
PD0307c	Células escamosas de pulmón	Tejido adyacente	
PD0312c	Células escamosas de pulmón	Tejido adyacente	
PD0319c	Células escamosas de pulmón	Tejido adyacente	
PD0331c	Células escamosas de pulmón	Tejido adyacente	
Identificación de la muestra	Tipo	Grupo	
NCI-H2087	Carcinoma de pulmón de células no pequeñas	Línea celular	NCI-BL2087-ONT
NCI-H2009	Carcinoma de pulmón de células no pequeñas	Línea celular	NCI-BL2009-ONT
NCI-BL2087	Linfocitos B inmortalizados	Línea celular	
NCI-BL2009	Linfocitos B inmortalizados	Línea celular	

Tabla 17: Muestras usadas en el presente estudio.

Todas estas muestras también fueron enviadas a secuenciar a Macrogen con una cobertura media de 28x y una longitud de 150 pares de bases. Posteriormente fueron alineadas contra el genoma de referencia HG38 usando el protocolo (5.2.1 Alineación de datos de secuenciación de Illumina (lecturas cortas “pair end”))

También fueron secuenciadas en nuestro laboratorio usando secuenciadores MinION de ONT, preparándose varias librerías por muestra (5-10) siguiendo el protocolo proporcionado por la compañía adaptado para obtener lecturas largas, descrito en (5.1.12 Secuenciación de nanoporos mediante la tecnología de Oxford Nanopore Technologies). y alineadas contra el genoma Hg38 usando el protocolo (5.2.3 Alineación de datos de secuenciación MinION (ONT lecturas largas))

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Se muestran a continuación los datos estadísticos de los resultados de la secuenciación de lecturas largas (ONT)

Grupo	Tipo	Cobertura Media
Línea celular	Carcinoma de pulmón de células no pequeñas	29.65
	Linfocitos B inmortalizados	16.96
Tejido adyacente	Adenocarcinoma de colon	29.11
	Células escamosas de cabeza y cuello	27.87
	Células escamosas de pulmón	29.84
	Sangre	15.38
Tumor	Adenocarcinoma de colon	29.29
	Adenocarcinoma de esófago	27.58
	Carcinoma de células escamosas de cabeza y cuello	34.88
	Carcinoma de células escamosas de pulmón	31.20

Tabla 18: Estadísticas de los resultados de la secuenciación por ONT de las muestras.

El programa usado para procesar esta información, junto con más gráficas, se encuentra en el repositorio asociado a esta Tesis y es el siguiente:

<../notebooks/0_vision_general/20230315_samples_statistics_01.ipynb>

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

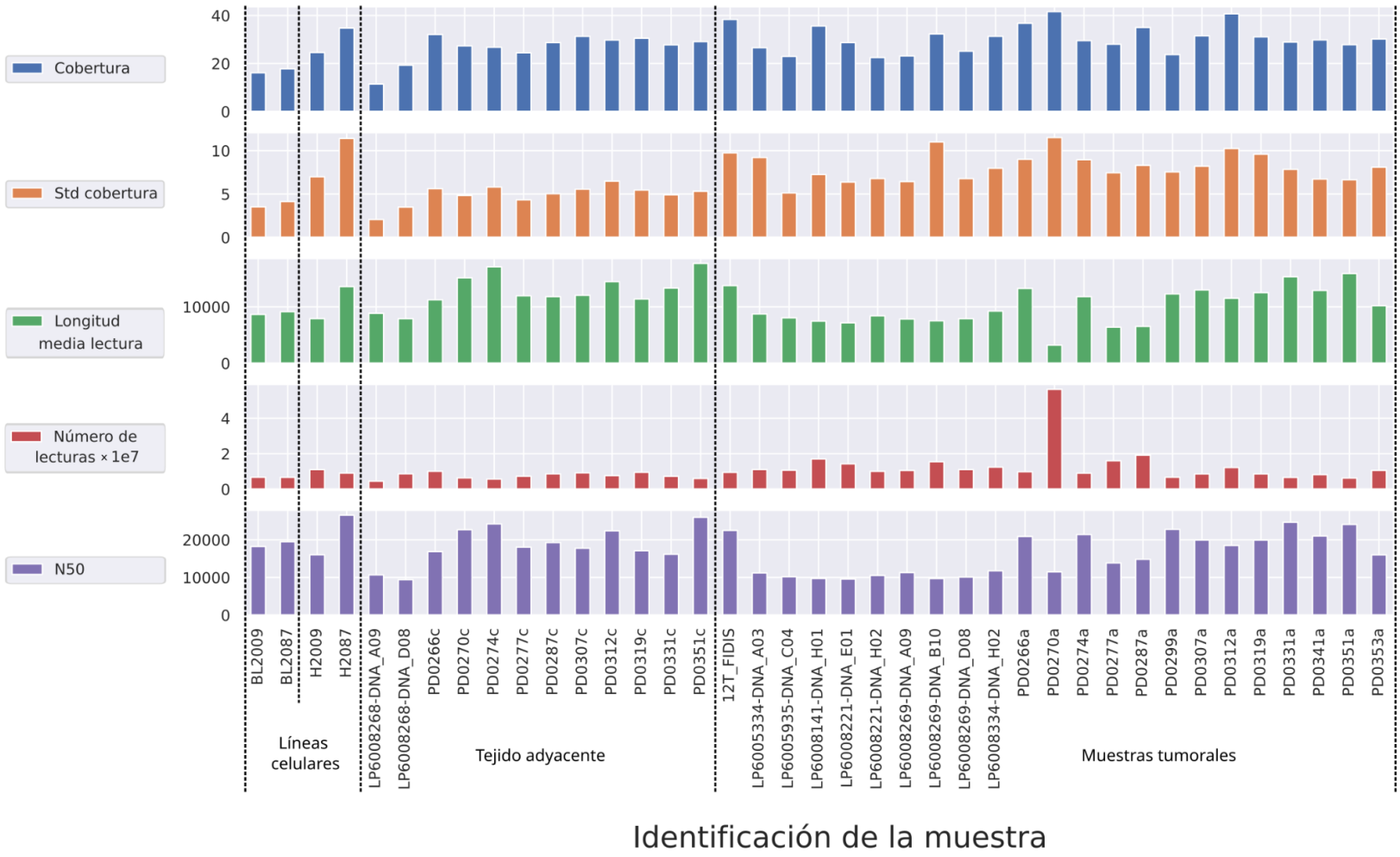


Figura 113: Estadísticas de la secuenciación por ONT de las muestras de la cohorte.

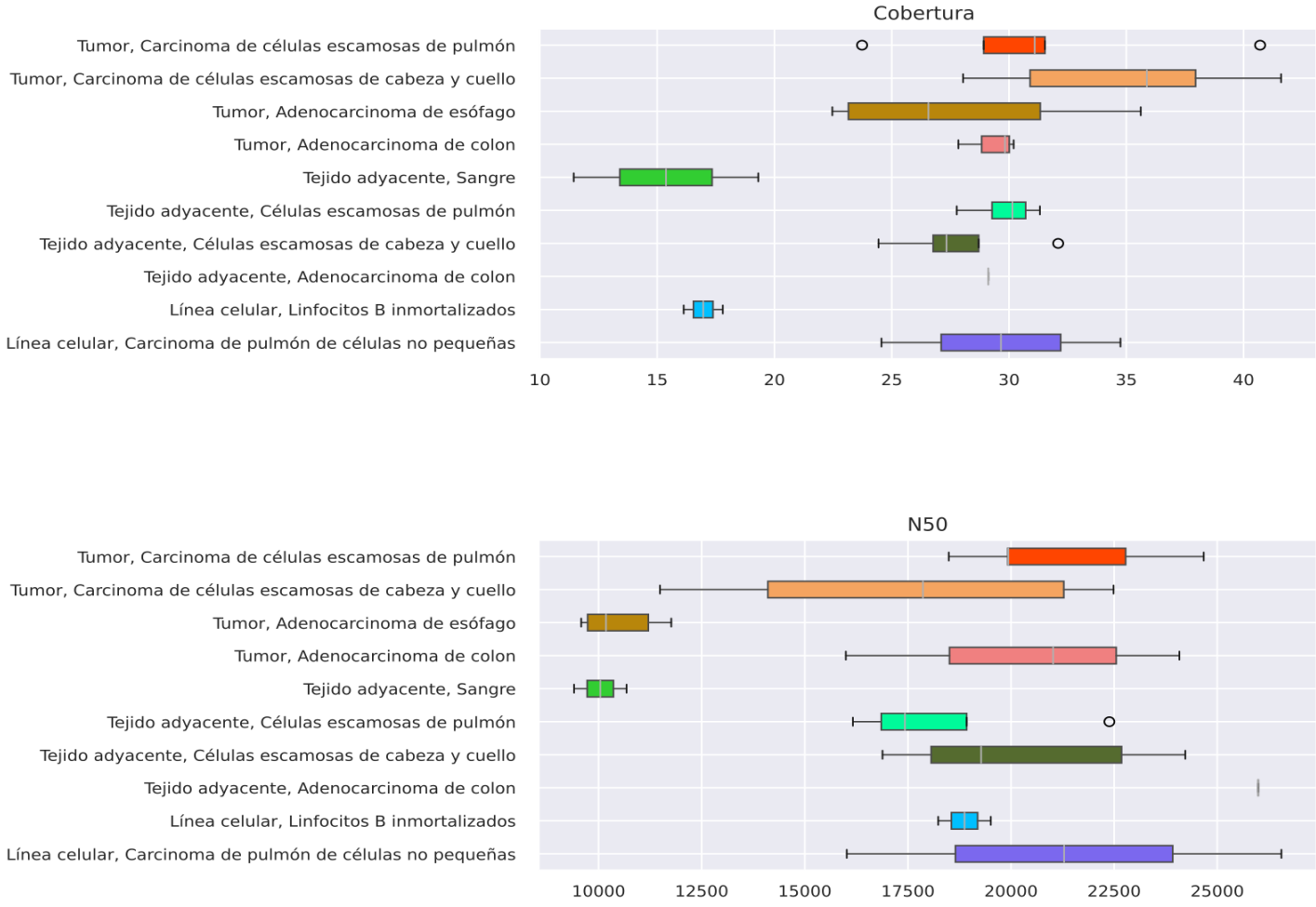


Figura 114: Gráficas de bigote de la cobertura (panel superior) y longitud de las lecturas (panel inferior) agrupados por tipo de tumor.

7.1.1.1 Pureza de los tumores

La mayor parte de los cortes histológicos de tumores contienen no sólo células tumorales, sino también células del estroma, del tejido adyacente, del sistema inmune, etc. En muchos análisis genómicos es fundamental conocer la pureza y la ploidía de cada una de las muestras. Para este conjunto de muestras este estudio fue realizado por los estudiantes de doctorado Sonia Zumalave y Iago Otero, usando el programa Battemberg (`Versión: 2.2.10`). El algoritmo de Battemberg está específicamente diseñado para detectar cambios en el número de copias subclonales en datos de secuenciación NGS (Nik-Zainal et al. 2012), proporcionando además datos de pureza. Para ello se usaron los datos de secuenciación de Illumina que disponemos de estas muestras, así como el genoma de referencia hg38.

MUESTRA	PUREZA	PLOIDÍA	PSI
LP6005334-DNA_A03	0.68	4.36	4.44
LP6005935-DNA_C04	0.92	2.05	2.06
LP6008141-DNA_H01	0.42	3.54	3.46
LP6008221-DNA_E01	0.83	2.60	2.93
LP6008221-DNA_H02	0.66	1.61	1.98
LP6008269-DNA_A09	0.44	3.50	3.10
LP6008269-DNA_B10	0.87	2.49	2.66
LP6008269-DNA_D08	0.77	2.52	2.59
LP6008334-DNA_H02	0.73	2.47	2.55
PD0341a	0.68	4.10	4.14
PD0351a	0.73	3.03	3.09
PD0353a	0.79	2.43	2.56
12T_FIDIS	0.74	2.45	2.82
PD0266a	0.76	2.90	2.72
PD0270a	0.78	2.86	2.74
PD0274a	0.93	1.72	1.76
PD0277a	0.72	1.56	1.57
PD0287a	0.54	3.30	3.09
PD0299a	0.71	3.42	3.24
PD0307a	0.71	2.28	2.96
PD0312a	0.54	2.91	2.75
PD0319a	0.83	3.42	3.03
PD0331a	0.77	2.85	3.00

Tabla 19: Pureza de las muestras tumorales (obtenida con Battemberg).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.1.2 Líneas celulares

Las líneas celulares de cáncer de pulmón constituyen una fuente invaluable de información en la investigación biomédica y traslacional. Debido a que existe una alta similitud entre las líneas celulares de pulmón y los tumores de los cuales proceden, las primeras son muy adecuadas para el estudio de los tumores de pulmón (Gazdar et al. 2010). Una de las grandes ventajas de las líneas celulares es que permiten obtener grandes cantidades de ADN de alta pureza, indispensable para la correcta caracterización de los métodos y protocolos, que después se pueden se pueden extrapolar y utilizar en los tumores primarios donde la disponibilidad de ADN es mucho más limitada.

Las líneas celulares usadas en el presente trabajo fueron obtenidas directamente de alícuotas procedentes del Instituto Sanger, traídas a nuestro laboratorio por el Dr. Jose C. Tubio, donde son activamente mantenidas y cultivadas. El ADN de las muestras fue extraído usando el protocolo 5.1.4 Extracción de ADN a partir de líneas celulares y mandado a secuenciar en varias ocasiones a Magrogen (nextseq 90GB)

Actualmente se pueden obtener del instituto American Type Culture Collection (ATCC). **(https://www.atcc.org/).**

El instituto Wellcome Trust Sanger Institute **(www.sanger.ac.uk)** mantiene la base de datos pública COSMIC **(https://cancer.sanger.ac.uk/cosmic)** con un catálogo de mutaciones somáticas en cáncer. También mantiene una base de datos de líneas celulares de donde se ha obtenido la información de estas líneas celulares. **(https://cancer.sanger.ac.uk/cell_lines)**

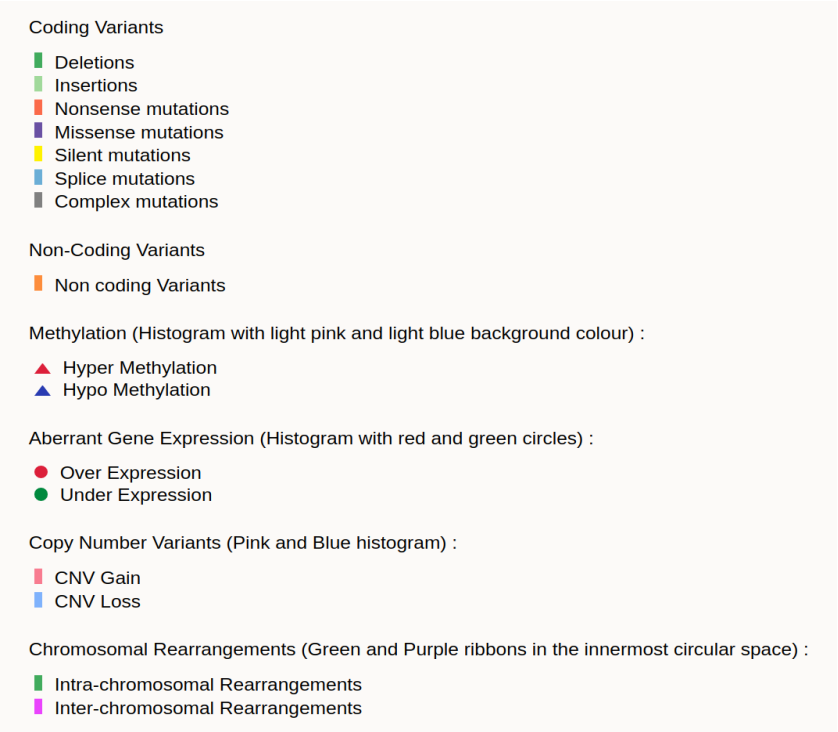


Figura 115: Leyenda de las figuras 116 y 1117 (circos plot de las líneas celulares).

7.1.2.1 NCI-H2087

Esta línea celular procede de una metástasis de un carcinoma de pulmón de células no pequeñas obtenida de un nódulo linfático. Es un tumor de grado I. El paciente era un varón caucásico, fumador de 69 años.

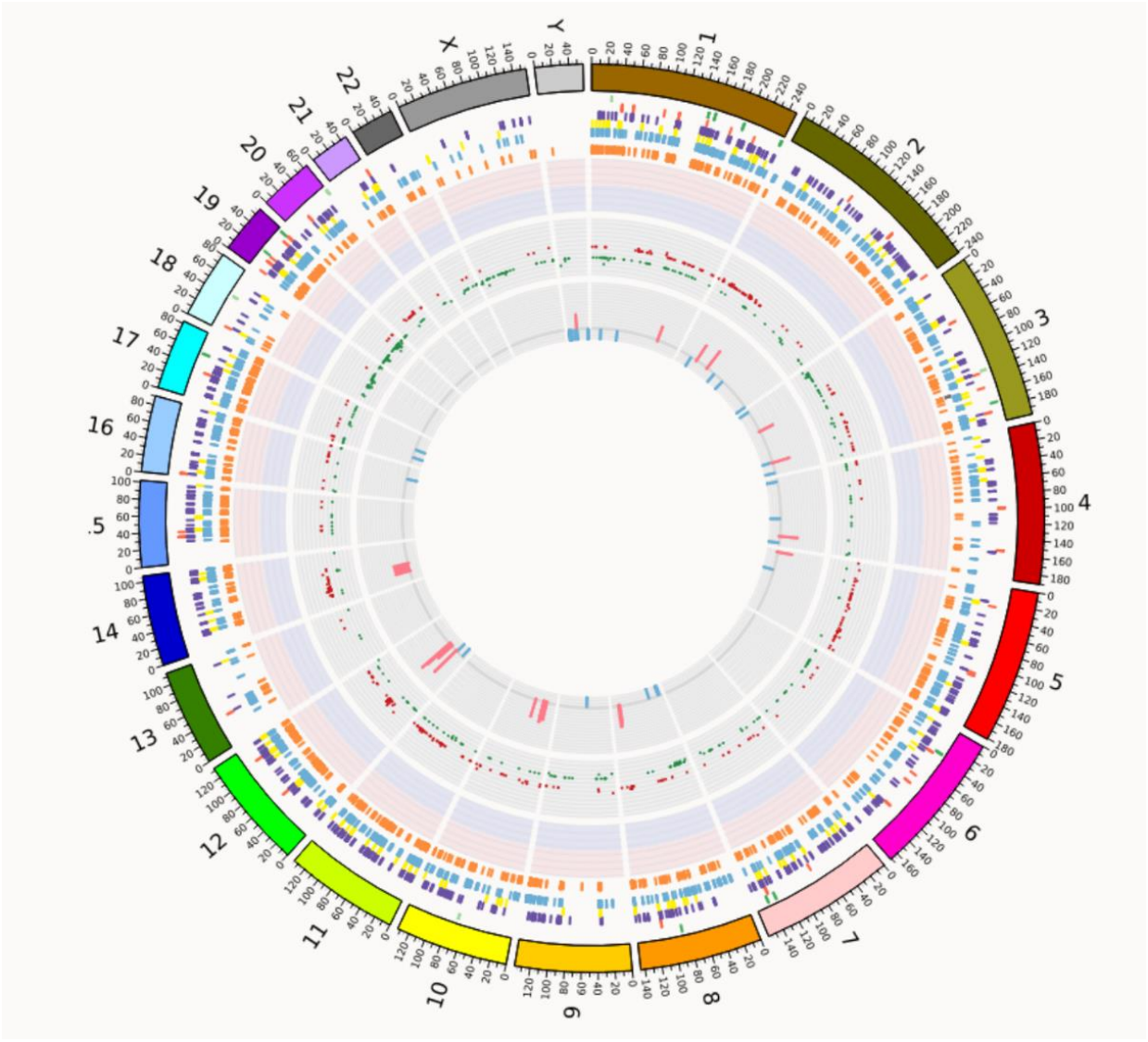


Figura 116: Circos plot con información de la línea celular NCI-H2087. https://cancer.sanger.ac.uk/cell_lines/sample/overview?id=724834 (accedido 29.10.2022)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.1.2.2 NCI-H2009

Esta línea celular procede de una metástasis de un carcinoma de pulmón de células no pequeñas. Es un tumor de grado IV. El paciente era una mujer caucásica, fumadora de 58 años.

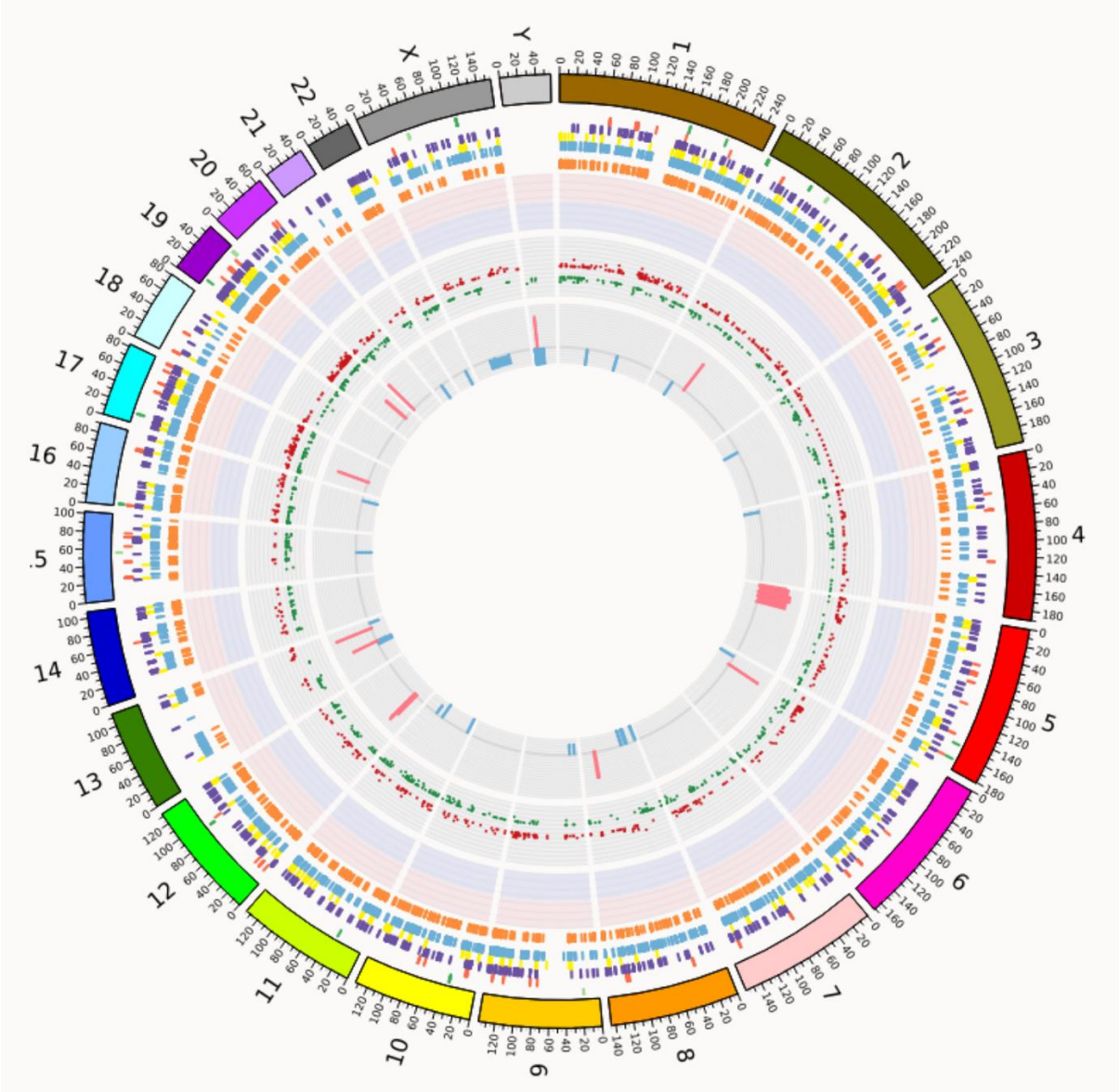


Figura 117: Circos plot con información de la línea celular NCI-H2009. https://cancer.sanger.ac.uk/cell_lines/sample/overview?id=724873 (accedido 29.10.2022)

7.1.2.3 NCI-BL2087 y NCI-BL2009.

Estas líneas celulares se obtuvieron a partir de los linfocitos B (sangre periférica) de los respectivos pacientes e inmortalizadas mediante la transfección del Virus Epstein-Barr.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.2 CONJUNTO DE DATOS

Para la realización de este trabajo se usaron datos de dominio público y datos generados por nosotros. Estos datos consisten en:

7.2.1 Genoma de referencia

El genoma de referencia usado para este trabajo es el GRCh38.p14 (Genome Reference Consortium Human Build 38 patch release 14 (GRCh38.p14) descargado de

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz

7.2.2 Secuencia consenso L1HS

La secuencia consenso de LINE1 se utilizó para generar los cromosomas artificiales contra los que alinear los datos de las muestras, en aquellos casos donde no se tenía la secuencia exacta.

```
AAGAAAATGAAAGTGTGAGAGGAGGAGCCAAGATGGCTGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACG
CAGAAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTCACTAGGGAGTGCCAGACAGTGGGCGCAGGCCA
GTGGGTGCGCGCGCATCTTGAGAGCTTGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGTTCC
CTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTCCCACCCGAATATTGCGCTTTTCAGACCG
GCTTAAGAAACGGCGCACCACGAGACTATATCCCACACCTGGCTCAGAGGGTCCTACGCCCACGGAATCTCGCTGATTGCT
AGCACAGCAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGGAGGGGCGCCCGCCATTGCCCAGGCTTGCTTAG
GTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAGCCCACCACAGCTCAAGGAGGCCTGCCTGCCTCTGTAGGCTCCAC
CTCTGGGGGCAGGACACAGACAAACAAAAAGACAGCAGTAACCTCTGCAGACTTAAGTGTCCCTGTCTGACAGCTTTGAAG
AGAGCAGTGGTTCTCCCAGCACGCAGCTGGAGATCTGAGAATGGGCAGACTGCCTCCTCAAGTGGGTCCCTGACCCCTGAC
CCCCAGCAGCCTAACTGGGAGGCACCCCCAGCAGGGGCACACTGACACCTCACACGGCAGGGTATTCCAACAGACCTGCAG
CTGAGGGTCCTGTCTGTTAGAAGGAAAACTAACAACCAGAAAGGACATCTACACTGAAAACCCATCTGTACATCACCATCA
TCAAAGACCAAAAGTAGATAAAACCACAAAGATGGGGAAAAAAACAGAACAGAAAAGCTGGAAACTCTAAAACGCAGAGCG
CCTCTCCTCCTCCAAGGAACGCAGTTCCTCACCAGCAACAGAACAAAGCTGGATGGAGAATGATTTTGACGAGCTGAGAGA
AGAAGGCTTCAGACGATCAAATTACTCTGAGCTACGGGAGGACATTCAAACCAAAGGCAAAGAAGTTGAAATCTTTGAAAA
AATTTAGAAGAATGTATAACTAGAATAACCAATACAGAGAAGTGCTTAAAGGAGCTGATGGAGCTGAAAACCAAGGCTCGA
GAACTACGTGAAGAATGCAGAAGCCTCAGGAGCCGATGCGATCAACTGGAAGAAAGGGTATCAGCAATGGAAGATGAAATG
AATGAAATGAAGCGAGAAGGGAAGTTTAGAGAAAAAGAATAAAAGAAATGAGCAAAGCCTCCAAGAAATATGGGACTATGT
GAAAAGACCAAATCTACGTCTGATTGGTGTACCTGAAAGTGATGTGGAGAATGGAACCAAGTTGGAAAACACTCTGCAGGA
TATTATCCAGGAGAACTTCCCCAATCTAGCAAGGCAGGCCAACGTTCAGATTCAGGAAATACAGAGAACGCCACAAAGATA
CTCCTCGAGAAGAGCAACTCCAAGACACATAATTGTCAGATTCACCAAAGTGGAAATGAAGGAAAAAATGTTAAGGGCAGC
CAGAGAGAAAGGTCGGGTTACCCTCAAAGGAAAGCCCATCAGACTAACAGCGGATCTCTCGGCAGAAACCCTACAAGCCAG
AAGAGAGTGGGGGCCAATATTCAACATTCTTAAAGAAAGAATTTTCAACCCAGAATTTCATATCCAGCCAAACTAAGCTTC
ATAAGTGAAGGAGAAATAAAATACTTTATAGACAAGCAAATGCTGAGAGATTTTGTCACCACCAGGCCTGCCCTAAAAGAG
CTCCTGAAGGAAGCGCTAAACATGGAAAGGAACAACCGGTACCAGCCGCTGCAAAATCATGCCAAAATGTAAAGACCATCG
AGACTAGGAAGAAACTGCATCAACTAATGAGCAAAATCACCAGCTAACATCATAATGACAGGATCAAATTCACACATAACA
ATATTAACTTTAAATATAAATGGACTAAATTCTGCAATTAAAAGACACAGACTGGCAAGTTGGATAAAGAGTCAAGACCCA
TCAGTGTGCTGTATTCAGGAAACCCATCTCACGTGCAGAGACACACATAGGCTCAAAATAAAAGGATGGAGGAAGATCTAC
CAAGCCAATGGAAAACAAAAAAAGGCAGGGGTTGCAATCCTAGTCTCTGATAAAACAGACTTTAAACCAACAAAGATCAAA
AAGAGACAAAGAAGGCCATTACATAATGGTAAAGGGATCAATACAACAAGAGGAGCTAACTATCCTAAATATTTATGCACC
CAATACAGGAGCACCCAGATTCATAAAGCAAGTCCTGAGTGACCTACAAAGAGACTTAGACTCCCACACATTAATAATGGG
AGACTTTAACACCCCACTGTCAACATTAGACAGATCAACGAGACAGAAAGTCAACAAGGATACCCAGGAATTGAACTCAGC
TCTGCACCAAGCAGACCTAATAGACATCTACAGAACTCTCCACCCCAAATCAACAGAATATACATTTTTTCAGCACCACAC
CACACTTATTCCAAAATTGACCATATAGTTGGAAGTAAAGCTCTCCTCAGCAAATGTAAAAGAACAGAAATTATAACAAAC
TATCTCTCAGACCACAGTGCAATCAAACTAGAACTCAGGATTAAGAATCTCACTCAAAGCCACTCAACTACATGGAAACTG
AACAACCTGCTCCTGAATGACTACTGGGTACATAACGAAATGAAGGCAGAAATAAAGATGTTCTTTGAAACCAACGAGAAC
AAAGACACCACATACCAGAATCTCTGGGACGCATTCAAAGCAGTGTGTAGAGGGAAATTTATAGCACTAAATGCCCACAAG
AGAAAGCAGGAAAGATCCAAAATTGACACCCTAACATCACAATTAAAAGAACTAGAAAAGCAAGAGCAAACACATTCAAAA
GCTAGCAGAAGGCAAGAAATAACTAAAATCAGAGCAGAACTGAAGGAAATAGAGACACAAAAACCCTTCAAAAAATCAATG
AATCCAGGAGCTGGTTTTTCGAAAGGATCAACAAAATTGATGGACCGCTAGCAAGACTAATAAAGAAAAAAAGAGAGAAGA
ATCAAATAGACACAATAAAAAATGATAAAGGGGATATCACCACTGATCCCACAGAAATACAAACTACCATCAGAGAATACT
```

```
ACAAACACCTCTACGCAAATAAACTAGAAAATCTAGAAGAAATGGATACATTCCTCGACACATACACTCTCCCAAGACTAA
ACCAGGAAGAAGTTGAATCTCTGAATAGACCAATAACAGGCTCTGAAATTGTGGCAAAAATCAATAGTTTACCAACCAAAA
AGAGTCCAGGACCAGATGGATTCACAGCCGAATTCTACCAGAGGTACAAGGAGGAACTGGTACCATTCCTCCTGAAACTAT
TCCAATCAATAGAAAAAGAATCCTCCTAACTCATTTTATGAGGCCAGCATCATTCTGATACCAAAGCCGGGCAGAGACACA
ACCAAAAAAAAAGAATTTTAGACCAATATCCTTGATGAACATTGATGCAAAAATCCTCAATAAAATACTGGCAAACCGAATC
CAGCAGCACATCAAAAAGCTTATCCACCATGATCAAGTGGGCTTCATCCCTGGGATGCAAGGCTGGTTCAATATACGCAAA
TCAATAAATGTAATCCAGCATATAAACAGAGCCAAAGACAAAAACCACATGATTATCTCAATAGATGCAGAAAAAGCCTTT
GACAAAATTCAACAACCCTTCATGCTAAAAACTCTCAATAAATTAGGTATTGATGGGACGTATTTCAAAATAATAAGAGCT
ATCTATGACAAACCCACAGCCAATATCATACTGAATGGGCAAAAACTGGAAGCATTCCCTTTGAAAACTGGCACAAGACAG
GGATGCCCTCTCTCACCGCTCCTATTCAACATAGTGTTGGAAGTTCTGGCCAGGGCAATCAGGCAGGAGAAGGAAATAAAG
GGTATTCAATTAGGAAAAGAGGAAGTCAAATTGTCCCTGTTTGCAGACGACATGATTGTTTATCTAGAAAACCCCATCGTC
TCAGCCCAAAATCTCCTTAAGCTGATAAGCAACTTCAGCAAAGTCTCAGGATACAAAATCAATGTACAAAAATCACAAGCA
TTCTTATACACCAACAACAGACAAACAGAGAGCCAAATCATGAGTGAACTCCCATTCACAATTGCTTCAAAGAGAATAAAA
TACCTAGGAATCCAACTTACATGGGATGTGAAGGACCTCTTCAAGGAGAACTACAAACCACTGCTCAAGGAAATAAAAGAG
GACACAAACAAATGGAAGAACATTCCATGCTCATGGGTAGGAAGAATCAATATCGTGAAAATGGCCATACTGCCCAAGGTA
ATTTACAGATTCAATGCCATCCCCATCAAGCTACCAATGACTTTCTTCACAGAATTGGAAAAACTACTTTAAAGTTCATAT
GGAACCAAAAAAGAGCCCGCATCGCCAAGTCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCATCACACTACCTGACTTCA
GACTATACTACAAGGCTACAGTAACCAAAAACAGCATGGTACTGGTACCAAAACAGAGATATAGATCAATGGAACAGAACA
GAGCCCTCAGAAATAATGCCACATATCTACAACTATCTGATCTTTGACAAACCTGAGAAAAACAAGCAATGGGGAAAGGAT
TCCCTATTTAATAAATGGTGCTGGGAAAACTGGCTAGCCATATGTAGAAAGCTGAAACTGGATCCCTTCCTTACACCTTAT
ACAAAAATCAATTCAAGATGGATTAAAGATTTTAAACGTTAGACCTAAAAACCATAAAACCCTAGAAGAAAACCTAGGCAT
TACCATTCAGGACATAGGCGTGGGCAAGGACTTCATGTCCAAAACACCAAAAGCAATTGCAACAAAAGCCAAAATTGACAA
ATGGGATCTAATTAAACTAAAGAGCTTCTGCACAGCAAAAGAAACTACCATCAGAGTGAACAGGCAACCTACAACATGGGA
AAAAATTTCGCAACCTACTCATCTGACAAAGGGCTAATATCCAGAATCTACAATGAACTCAAACAAATTTACAAGAAAAAA
AACAAACAACCCCATCAAAAAGTGGGCGAAGGACATGAACAGACACTTCTCAAAAGAAGACATTTATGCAGCCAAAAAACA
CATGAAGAAATGCTCATCATCACTGGCCATCAGAGAAATGCAAATCAAAACCACTATGAGATATCATCTCACACCAGTTAG
AATGGCAATCATTAAAAAGTCAGGAAACAACAGGTGCTGGAGAGGATGTGGAGAAATAGGAACACTTTTACACTGTTGGTG
GGACTGTAAACTAGTTCAACCATTGTGGAAGTCAGTGTGGCGATTCCTCAGGGATCTAGAACTAGAAATACCATTTGACCC
AGCTATCCCATTACTGGGTATATACCCAAAGGACTATAAATCATGCTGCTATAAAGACACATGCACACGTATGTTTATTGC
GGCACTATTCTAATAGCAAAGACTTGGAACCAACCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATA
TACACCATGGAATACTATGCAGCCATAAAAATGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTGAAACCATCAT
TCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCACTCATAGGTGGGAATTGAACAATGAGAACACT
TGGACACAGGAAGGGGAATATCACACTCTGGGGACTGTGGTGGGGTCGGGGGAGGGGAGGGATAGCATTGGGAGATATACC
TAATGCTAGATGACACGTTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACCTGCACAATGTGCACATG
TACCCTAAAACTTAGAGTATAATAAAAAATTAAAAAAAAAAAAAAAAA
```

Tabla 20: Secuencia consenso L1HS.

Debido a su tamaño, las tablas con las secuencias consenso de las familias Ta0, Ta1, Pre-Ta se adjuntan en el repositorio que acompaña a este manuscrito.

7.2.3 LINE1 objeto de análisis

Para este trabajo estamos interesados en aquellos LINE1 capaces de movilizarse, es decir que pudieran están activos en cáncer. Como se ha comentado anteriormente en el genoma humano existen entre 500,000 a 800,000 copias de LINE1. De ellas la gran mayoría son no funcionales, debido a

- Truncamiento 5'
- Inversión de parte del elemento
- Mutaciones acumuladas a lo largo de la evolución

Por ello se ha tenido que hacer una selección previa de elementos con capacidad (a priori) de movilización, esto es que tengan la longitud completa, es decir sin truncamiento (full length > 6000pb), con la región promotora intacta, que no tengan inversiones internas y con los marcos de lectura (ORF) también intactos. A partir de ahora, en este trabajo, a los

USC DE SANTIAGO DE COMPOSTELA

LINE1 que cumplan estos criterios los denominaremos "íntegros". Esta selección se realizó tras explorar varias alternativas como se describe a continuación.

7.2.3.1 Análisis inicial con RepeatMasker obtenido de UCSC

Existen varias fuentes de información de repeticiones en genomas humanos, una de las más comúnmente usada es **RepeatMasker** **(Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org)** (Tarailo-Graovac y Chen 2009). Con esa denominación se hace referencia tanto a la web, como al programa usado para generar los datos como a la anotación producida por el programa. Inicialmente, para hacerme una idea global de los LINE1 en el genoma humano (abundancia, tipo, integridad, familia) se realizaron varios análisis preliminares ejecutando el programa RepeatMasker sobre el genoma de referencia Hg38. No obstante, por motivos de reproducibilidad decidí bajar los datos directamente del repositorio "*UCSC genome browser*" del Instituto de Genómica de la universidad de California Santa Cruz. Para ello se usó la siguiente cadena

https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1421614261_RChMHHVIr63jEx8ZEvspvGbzACAv&clade=mammal&org=Human&db=hg38&hgta_group=rep&hgta_track=knownGene&hgta_table=0&hgta_regionType=genome&position=chrX%3A15%2C560%2C138-15%2C602%2C945&hgta_outputType=primaryTable&hgta_outFileName=

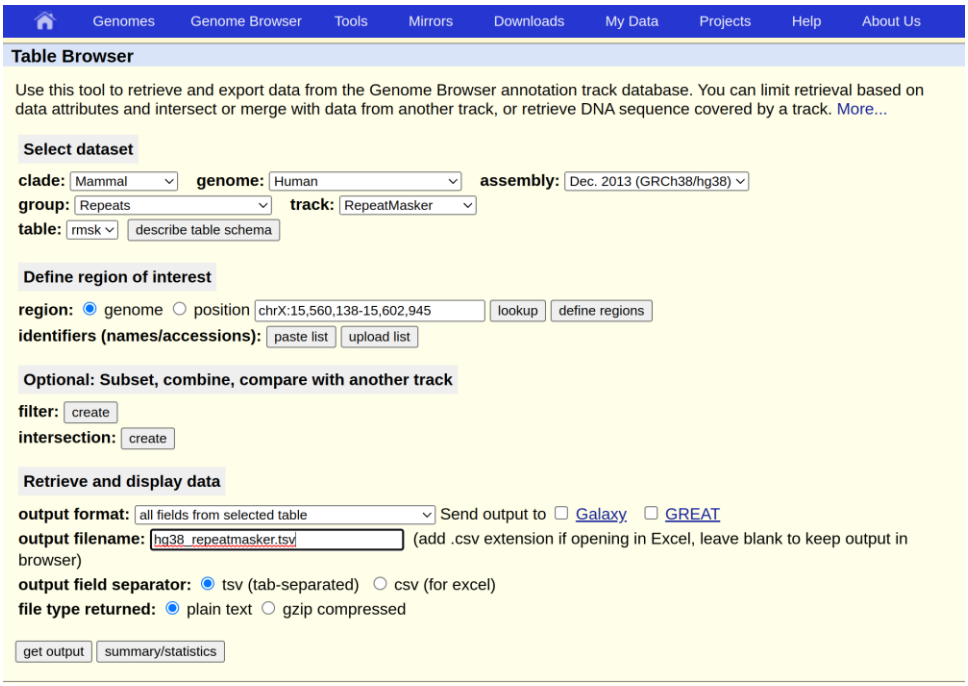


Figura 118: Descarga de los datos del RepeatMasker del UCSC. (accedido 19.03.2021)

El fichero presenta el siguiente esquema:

Schema for RepeatMasker - Repeating Elements by RepeatMasker

Database: hg38 **Primary Table:** rmsk **Row Count:** 5,633,664 **Data last updated:** 2021-09-03
Format description: RepeatMasker .out record

field	example	SQL type	description
bin	585	smallint(5) unsigned	Indexing field to speed chromosome range queries.
swScore	463	int(10) unsigned	Smith Waterman alignment score
milliDiv	13	int(10) unsigned	Base mismatches in parts per thousand
milliDel	6	int(10) unsigned	Bases deleted in parts per thousand
milliIns	17	int(10) unsigned	Bases inserted in parts per thousand
genoName	chr1	varchar(255)	Genomic sequence name
genoStart	10000	int(10) unsigned	Start in genomic sequence
genoEnd	10468	int(10) unsigned	End in genomic sequence
genoLeft	-248945954	int(11)	-#bases after match in genomic sequence
strand	+	char(1)	Relative orientation + or -
repName	(TAACCC)n	varchar(255)	Name of repeat
repClass	Simple_repeat	varchar(255)	Class of repeat
repFamily	Simple_repeat	varchar(255)	Family of repeat
repStart	1	int(11)	Start (if strand is +) or -#bases after match (if strand is -) in repeat sequence
repEnd	471	int(11)	End in repeat sequence
repLeft	0	int(11)	-#bases after match (if strand is +) or start (if strand is -) in repeat sequence
id	1	char(1)	First digit of id field in RepeatMasker .out file. Best ignored.

Tabla 21: Esquema de los datos del RepeatMasker descargados de la UCSC.

Tras una serie de filtrados nos quedamos solo con las familias de LINE1 de primates y de ellos solo los correspondientes a humanos (L1HS) (83575 elementos), de ellos nos quedamos con los que son mayores de 6000 pb (6517 elementos). No obstante, este sistema no nos permite distinguir los que son íntegros al no tener información del estado de los marcos de lectura abiertos (ORF) o si contenían inversiones o mutaciones puntuales. Por eso decidí recurrir a bases de datos preexistentes de LINE1 en humanos como se explicará en el siguiente apartado.

Nota: las tablas y datos usados se encuentran el repositorio en el directorio

<../1_hg38_repeatmasker/ >

7.2.3.2 Análisis inicial con datos obtenidos de la web de L1base y RepeatMasker

El “Institute for Medical Genetics and Human Genetics, Computational Biology and Department of Radiology, Charité Universitätsmedizin Berlín” mantiene una base de datos “L1Base (v2)” (Penzkofer et al. 2017) especializada de secuencias de LINE1 con actividad en humanos (y roedores). La web permite la visualización de los LINE1 y sus características, por lo que resulta muy útil e informativo para la exploración de los LINE1.

Alberga también una serie de ficheros:

1. LINE1 completos e intactos en los dos ORF:

[FLI-L1] Human Full-Length, Intact LINE1 Elements (Ens84.38) 146 elementos, actualizada en 2016-06-01)

2. LINE-2 con ORF2 intacto, pero con ORF1 interrumpido:

[ORF2-L1] Human ORF2 Intact LINE1 Elements (Ens84.38) (107 elementos. Actualizada en 2016-07-09)

3. Además, debido a su potencial regulador, también se incluyen en la base de datos las LINE1 no intactas (FLnI-L1) de longitud completa (>6000 pb):

[FLnI-L1] Human Full-Length > 4500nt LINE1 Elements (Ens84.38) (13418 elementos. Actualizada: 2016-09-27)

(https://l1base.charite.de/l1base.php accedida 30.03.2021)

Los ficheros contienen abundante de información acerca de los LINE1 de gran importancia para este trabajo, como, las islas CpG, motivos de unión a reguladores, la familia y subfamilia, etc.

```
ID, Comment, Chr, Start, End, Query Start, Query End, Strand, ORF1 conserved,
ORF1 Start, ORF1 End, ORF1 Q-Start, ORF1 Q-End, ORF1 gaps, ORF1 frameshifts, ORF1
stops, ORF2 conserved, ORF2 Start, ORF2 End, ORF2 Q-Start, ORF2 Q-End, ORF2 gaps,
ORF2 frameshifts, ORF2 stops, Ta SSVs, Ta0/Ta1 SSVs, Ta1-nd/d, L1M/L1PA
Discrimination, Poly-A, PolyA Signal, Runx3 Site, Runx3 ASP, SRY Site 1, SRY Site
2, YY1 BoxA+BoxA, TF nkx-2.5, TF nkx-2.5B, find TSDs, REKG235, ARR260, YPAKLS282,
N14, E43, Y115, D145, N147, T192, D205, SDH228, R363, FADD700, HMKK1091, SSS1096,
I1220, S1259, ORF StartStop, G-C Content, Repeat Annontation, Genome Vicinity
(ENS), ORF1 %A, ORF1 %T, ORF2 %A, ORF2 %T, ORF1&2 %A, ORF1&2 %T, CpG Islands,
ORF1 CAI, ORF2 CAI, Intactness Score
```

Tabla 22: Campos de la base de datos L1base.

Estos ficheros presentan un problema y es que las coordenadas que facilita son imprecisas dando las posiciones de inicio de los elementos aproximadamente 1000 pb antes de la posición real y las posiciones finales aproximadamente 1000 pb después de la posición real. Methyl-Meiga necesita coordenadas precisas (a nivel de nucleótido) para funcionar correctamente. Para obtenerlas fusionamos la información de L1base con la del RepeatMasker (fichero con anotación), que contiene coordenadas precisas. El fichero fue descargado de su página web oficial **(https://www.repeatmasker.org/species/hg.html)**

Existen 2 ficheros referentes a las repeticiones del genoma de referencia hg38

- hg38 - Dec 2013 - RepeatMasker open-4.0.5 - Repeat Library 20140131
- hg38 - Dec 2013 - RepeatMasker open-4.0.6 - Dfam 2.0

El último es específico de elementos transponibles y fue el que descargué

https://www.repeatmasker.org/genomes/hg38/RepeatMasker-rm406-dfam2.0/hg38.fa.out.gz (accedido 21.03.2021)

El formato del fichero una vez descomprimido es el siguiente.

```
    SW  perc perc perc  query      position in query           matching       repeat
position in  repeat
 score  div. del. ins.  sequence    begin     end    (left)    repeat
class/family         begin  end (left)       ID

    64  24.6 17.7  2.5  chr1        11502   11676 (248944746) C  L1MC
LINE/L1              (2234) 5648   5448      3
    11  35.5  8.6 15.1  chr1        18418   18649 (248937773) +  L1M2b
LINE/L1                744  962 (6247)      9
   199  22.2  2.7  1.0  chr1        29901   30196 (248926226) +  L1MB3
LINE/L1               5882 6182    (1)     25
```

Tabla 23: Cabecera del fichero de transposones RepeatMasker.

Mediante comandos bash obtuve un fichero final, fusionando la información de L1base con las coordenadas del RepeatMasker.

```
cd /media/javi/SSD1T/05_transposons/hg38-Dec_2013-RepeatMasker_open-4.0.6-Dfam_2.0/
head -n 3 /media/javi/SSD1T/05_transposons/hg38-Dec_2013-RepeatMasker_open-4.0.6-
Dfam_2.0/hg38.fa.out > head.txt
cat /media/javi/SSD1T/05_transposons/hg38-Dec_2013-RepeatMasker_open-4.0.6-
Dfam_2.0/hg38.fa.out |grep "LINE/L1" > LINE1_hg38_dfam.out

cat head.txt  LINE1_hg38_dfam.out > LINE1_hg38_dfam_header.out
awk 'BEGIN {OFS="\t"} {print $5,$6,$7,$9,$10}' LINE1_hg38_dfam.out >
LINE1_hg38_dfam.bed
cd /media/javi/SSD1T/05_transposons/01_L1BASE

cat *.tsv > all_L1_base.tsv
sort -k1.1V -k2,2n   all_L1_base.tsv > all_L1_base.sorted.tsv

#elimate duplicates
awk '{if($2 != start && $3 != end){print $0}; start=$2; end=$3}'
all_L1_base.sorted.tsv > all_L1_base.sorted.unique.tsv


bedtools intersect -b /media/javi/SSD1T/05_transposons/hg38-Dec_2013-
RepeatMasker_open-4.0.6-Dfam_2.0/LINE1_hg38_dfam.bed -a
/media/javi/SSD1T/05_transposons/01_L1BASE/all_L1_base.sorted.unique.tsv -wo >
intersection.bed

awk 'NR<=1{chr=$1; start=$2; maxOverl=$15} {if($1 == chr && $2 == start && $15 >=
maxOverl) {maxOverl=$15; linePrint=$0} else if($1 == chr && $2 == start && $15 <
maxOverl){next} else if($1 != chr
|| $2 != start){chr=$1; start=$2; maxOverl=$15; print linePrint;
linePrint=$0}}END{print linePrint} ' intersection.bed > intersection.unique.bed
```

Código 9: Fusión de los datos de RepeatMasker con los de L1Base.

El fichero resultante se encuentra en el repositorio en:

<../2_datos_de_repeatmasker/01_L1BASE/intersection.unique.bed>

Este fichero contiene 12392 elementos de los cuales 338 son L1H de más de 6000 pb.

chr1	437436	447857	UID-7124		0	+	442278	447141	128,128,128
chr1	440934	447355	+	L1PA7	6421				
chr1	672412	682833	UID-7125		0	+	677254	682117	128,128,128
chr1	675910	682331	+	L1PA7	6421				
chr1	4030826	4040471	UID-6947		0	+	4035007	4039928	128,128,128
chr1	4034321	4043199	+	L1PB1	6150				
chr1	7411258	7421422	UID-6882		0	-	7411978	7416879	128,128,128
chr1	7411759	7417922	C	L1PA4	6163				

Tabla 24: Muestra del fichero de LINE1 con coordenadas de los LINE1 precisas.

Mientras se estaba realizando este análisis dos compañeros del grupo de investigación, Bernardo Rodriguez-Martín y Martín Santamarina crearon una lista curada de LINE1 germinales íntegros a partir de la bibliografía existente y datos propios, lista que fue utilizada en el artículo (Ebert et al. 2021). Esa lista curada de LINE1 es el punto de partida para nuestra base de datos de LINE1 íntegros, tal como se comenta en el siguiente apartado.

7.2.3.3 Base de datos definitiva de LINE1 germinales, íntegros

A continuación, se describe como se ha obtenido la base de datos de LINE1 germinales íntegros usada en el presente trabajo. Esta contiene los LINE1 germinales íntegros, es decir que contienen la región promotora, los dos marcos de lectura abiertos (ORF) y no contienen inversiones internas. La mayor parte de ellos son L1HS y algún L1PA2. Entre ellos se encuentran los LINE1 de los cuales se conoce que presentan actividad.

Para el artículo "Haplotype-resolved diverse human genomes and integrated analysis of structural variation" (Ebert et al. 2021) los autores crearon una lista curada de LINE1 íntegros (abf7117_ebert_tables-s1-s56.xlsx), formada por LINE1 íntegros y por LINE1 que se sabe son activos en humanos (*source elements*). Esta lista la compusieron a partir de la literatura y los datos de secuenciación propios de varios individuos obtenidos con PacBio Hi-FI.

Mediante un libro de Python se extrajo la información y se combinaron las tablas

1. **Tabla S23. Compendium of FL-L1s known to be active**
2. **Tabla S22. Collection of sequence-resolved full-length L1s (FL-L1s)**

para obtener el fichero: **<../03_LINE1_fuente/FL-L1_allConsensus_GRCh38_v1.tsv>** que contiene 693 LINE1 de los cuales 18 no conocemos su orientación. Aunque se hicieron pruebas para obtenerla no fueron concluyentes, debido a que las muestras que tenemos no contenían esos elementos, los cuales deben estar a muy baja frecuencia en la población.

Tal como comentan los autores

- *Las coordenadas están basadas en HG38.*
- *Se incluyen elementos íntegros existentes en el genoma de referencia y alguno que no lo están*
- *Este archivo no representa toda la diversidad de secuencias de LINE1 íntegros en los seres humanos; es solo una base de datos enriquecida en elementos LINE1 íntegros activos y elementos resueltos en secuencia.*
- *Las secuencias consenso disponibles para cada elemento no necesariamente coinciden completamente con la presente en su muestra.*
- *Actualmente, 57 elementos no tienen secuencia disponible*

Para el fichero de entrada usado para Methyl-Meiga: **<../3_LINE1_fuente/FL-L1_for_Me-Me_Consensus_GRCh38_v1.tsv>** se descartaron estos 18 elementos sin orientación quedándonos con 675 elementos. Se conoce la secuencia consenso de 618. En el resto de los elementos, 57 se ha usado la secuencia de la familia (si se conoce) o si no la secuencia consenso genérica del LINE1.

Nota: el programa utilizado fue:

<../3_LINE1_fuente/scripts/Create_FL-L1_allConsensus_GRCh38_v1.tsv.ipynb>

7.2.4 Patrones de dinucleótidos CpG e Islas CpG en los LINE1

Durante la realización de esta Tesis fue necesario la visualización de gran cantidad de imágenes con los patrones de metilación de los LINE1 y las regiones adyacentes. Por eso era

de crucial importancia ser capaz de reconocer los LINE1. Un método al que recurrimos en la mayoría de las ocasiones, fue mediante la identificación de los patrones de dinucleótidos de CpG. Para ello se recurrió a la herramienta on line **"DataBase of CpG islands and Analytical Tool (DBCAT**) "**http://dbcat.cgm.ntu.edu.tw/** (accedida sep. 2020). Esta herramienta además de proporcionar las posiciones de las CpG nos muestra si en la región analizada existe una isla CpG. Se aplicó a las diferentes familias de LINE1 en humanos L1HS (Ta-1, Ta-0, Pre-Ta), L1PA1-L1P8, aunque por brevedad sólo se muestran a continuación las más frecuentemente observadas en los LINE1 objeto de investigación.

7.2.4.1 Ta-1

Esta secuencia de 6031 nucleótidos presenta una isla CpG en la posición 17 - 536 de longitud 520 pb y un contenido GC 60% tal como se puede observar en la siguiente figura

GGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGT
GAGCGACGCAGAAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTC
ACTAGGGAGTGCCAGACAGTGGGCGCAGGCCAGTGTGTGTGCGCACCGTGCGCGA
GCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGTT
CCCTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTCC
CACCCGAATATTGCGCTTTTCAGACCGGCTTAAGAAACGGCGCACCACGAGACTATAT
CCCACACCTGGCTCGGAGGGTCCTACGCCCACGGAATCTCGCTGATTGCTAGCACAG
CAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGAGGGGCGCCCGCCA
TTGCCCAGGCTTGCTTAGGTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAGC
CCACCACAGCTCAAG

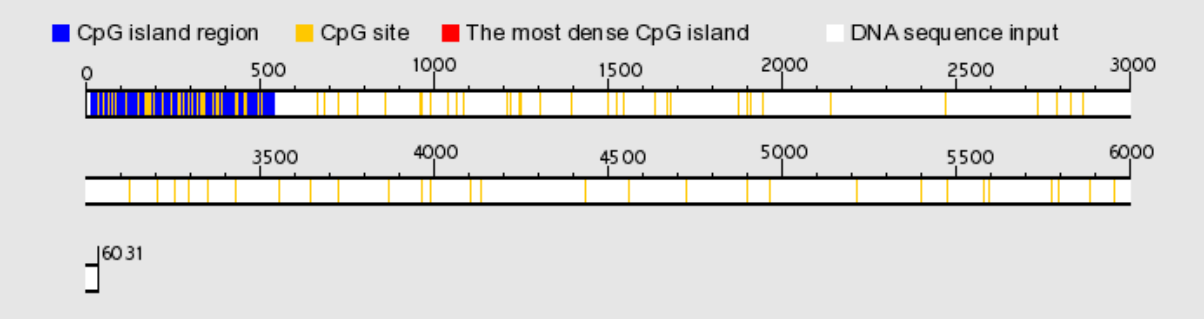


Figura 119: Disposición de las CpG e isla CpG del elemento Ta-1. Secuencia con las CpG marcadas (arriba) y las islas CpG (abajo)

7.2.4.2 Ta-0

Esta secuencia de 6032 nucleótidos presenta una isla CpG en la posición 17 - 534 de longitud 518 pb y un contenido GC 60% tal como se puede observar en la siguiente figura

GGAGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGT
GAGCGACGCAGAAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCT
CACTAGGGAGTGCCAGACAGTGGGCGCAGGCCAGTGTGTGTGCGCACCGTGCGCG
AGCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGT
TCCCTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTC
CCACCCGAATATTGCGCTTTTCAGACCGGCTTAAGAAACGGCGCACCACGAGACTATA
TCCCACACCTGGCTCAGAGGGTCCTACGCCCACGGAATCTCGCTGATTGCTAGCACA
GCAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGAGGGGCGCCCGCC
ATTGCCCAGGCTTGCTTAGGTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAG
CCCACCACAGCTC

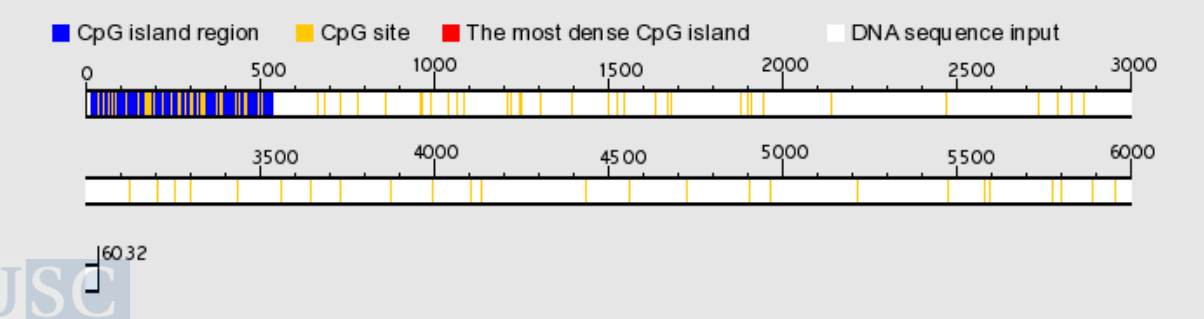


Figura 120: Disposición de las CpG e isla CpG del elemento Ta-0. Secuencia con las CpG marcadas (arriba) y las islas CpG (abajo)

7.2.4.3 Pre-Ta-0

Esta secuencia de 6034 nucleótidos presenta una isla CpG en la posición 19 - 536 de longitud 518 pb y un contenido GC 60% tal como se puede observar en la siguiente figura

GGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGT
GAGCGACGCAGAAGACGGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCT
CACTAGGGAGTGCCAGACAGTGGGCGCAGGCCAGTGTGTGTGCGCACCGTGCGCG
AGCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGT
TCCCTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTC
CCACCCGAATATTGCGCTTTTCAGACCGGCTTAAGAAACGGCGCACCACGAGACTATA
TCCCACACCTGGCTCAGAGGGTCCTACGCCCACGGAATCTCGCTGATTGCTAGCACA
GCAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGAGGGGCGCCCGCC
ATTGCCCAGGCTTGCTTAGGTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAG
CCCACCACAGCTC

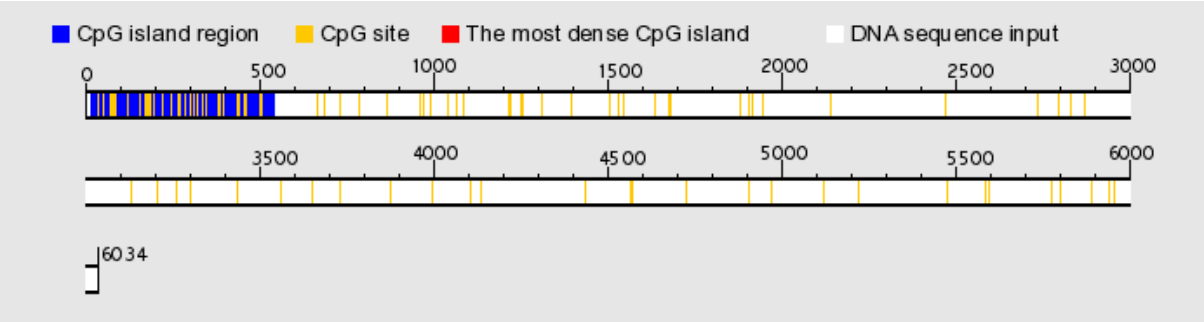


Figura 121: Disposición de las CpG e isla CpG del elemento pre-Ta.
Secuencia con las CpG marcadas (arriba) y las islas CpG (abajo)

7.2.4.4 L1PA1 (L1HS)

Esta secuencia de 6027 nucleótidos presenta una isla CpG en la posición 10 - 529 de longitud 520 pb y un contenido GC 60% tal como se puede observar en la siguiente figura

GGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGT
GAGCGACGCAGAAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTC
ACTAGGGAGTGCCAGACAGTGGGCGCAGGCCAGTGTGTGTCGCACCGTGCGCGA
GCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGTT
CCCTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTCC
CACCCGAATATTGCGCTTTTCAGACCGGCTTAAGAAACGGCGCACCACGAGACTATAT
CCCACACCTGGCTCGGAGGGTCCTACGCCCACGGAATCTCGCTGATTGCTAGCACAG
CAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGAGGGGCGCCCGCCA
TTGCCCAGGCTTGCTTAGGTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAGC
CCACCACAGCTCAAG

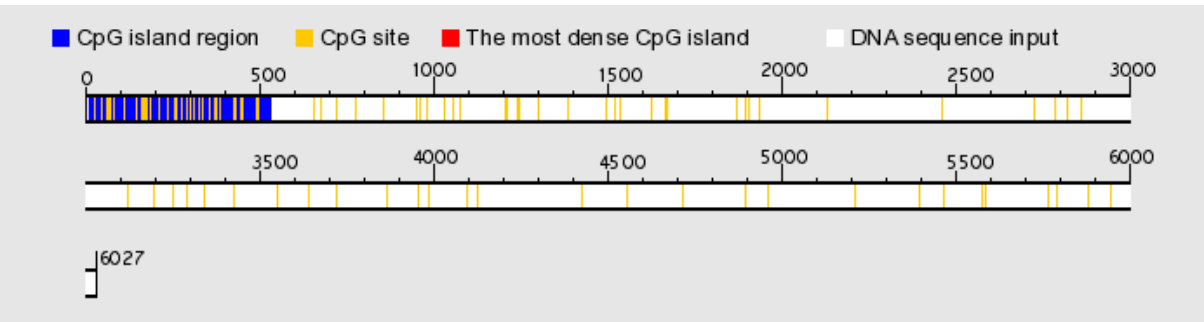


Figura 122: Disposición de las CpG e isla CpG del elemento L1HS.
Secuencia con las CpG marcadas (arriba) y las islas CpG (abajo)

7.2.5 Datos de las muestras tumorales, tejido adyacente y/o sangre

Tal como se comentaba en el apartado anterior (7.1), nuestro juego de datos consiste en los datos de lecturas cortas (Illumina) de todas las muestras tumorales, tejidos adyacentes y/o sangre, datos de metilación (WGBS) de las líneas celulares tumorales y datos de lecturas largas (ONT) de las siguientes muestras.

- 23 tumores
  - 9 tumores de esófago.
  - 3 carcinomas de colon.
  - 6 de carcinomas de células escamosas de cabeza y cuello.
  - 5 de carcinomas de células escamosas de pulmón.
- 10 tejidos adyacentes
  - 1 de colon.
  - 5 de células escamosas de cabeza y cuello.
  - 4 de células escamosas de pulmón.
- 2 muestras de linfocitos correspondientes a algunas de las muestras de esófago.
- 2 líneas celulares tumorales de pulmón.
- 2 líneas celulares de linfocitos inmortalizados correspondientes a las líneas celulares de pulmón.

El no disponer de un conjunto me muestras homogéneo, nos ha obligado a usar distintas aproximaciones, por ejemplo, a la hora de buscar las inserciones somáticas de LINE1 ejecutamos MEIGA-LR en modo pareado con aquellas muestras de las que disponemos de tejido normal para filtrar, esto es 12 tumores primarios (10 tejidos adyacente y dos sangres) y las dos líneas celulares (sangre). Para el resto de las muestras se ejecutó MEIGA-LR en modo pareado, pero usando como datos de la muestras “normal” la PD0277c y realizando después un filtrado de los LINE1 germinales con los datos de Illumina + xTea de su correspondiente tejido normal.

Para comparar los patrones de metilación del tumor con su tejido adyacente, no se puede usar los datos de secuenciación de la sangre, como se hace con las inserciones somáticas de LINE1 (y las mutaciones y variaciones estructurales), pues los patrones de metilación podrían verse influenciados por el tipo de tejido, más que por la variación entre tejido sano / tumor. (Los análisis se realizaron igualmente, para comprobar si era posible filtrar la componente debida a las diferencias de tejido de las diferencias producidas por la transición de tejido normal a tumoral).

Los datos de secuenciación obtenidos fueron procesados usando *pipelines* diseñados en nuestro grupo de investigación por Jorge Zamora, Sonia Zumalave, Iago Otero, Jenifer Brea, y yo mismo. Estas pipelines están basadas en snakemake **(https://snakemake.github.io/ accedida 09.06.2020)** y están específicamente diseñadas para aprovechar al máximo los recursos computacionales del Centro de Supercomputación de Galicia (CESGA) y su sistema de colas SLURM. Se usaron dos conjuntos distintos de pipelines dependiendo de la tecnología usada para secuenciar los datos (descritas en 5.2 Pipelines y programas bioinformáticos).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

```
/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_javi/2_meme/05
_input_data_MEIGA-LR.2.1.0/PD0266a_5_PD0266c cat launch.sh
#!/bin/sh
#SBATCH -c 24
#SBATCH -t 30:00:00
#SBATCH --mem 120GB
#SBATCH --error
/mnt/netapp2/genomas_enfermedad/1/1_projects/3051_MEIGA_SCREENING-
MN/9_various/03_MEIGA-LR/2.1.0/PD0266a_5_PD0266c/MEIGA.err
#SBATCH --output
/mnt/netapp2/genomas_enfermedad/1/1_projects/3051_MEIGA_SCREENING-
MN/9_various/03_MEIGA-LR/2.1.0/PD0266a_5_PD0266c/MEIGA.out
#SBATCH --mail-type=end

#SBATCH --mail-user=sonia.zumalave@gmail.com
source /mnt/netapp2/uscmg_aplic/1_internal_tools/26_MEIGA-LR/env/activate
export PYTHONHASHSEED=0
mprof run --include-children --multiprocess --python
/mnt/netapp2/uscmg_aplic/1_internal_tools/26_MEIGA-LR/releases/2.1.0/MEIGA-LR.py
/mnt/netapp2/genomas_enfermedad/1/1_projects/3051_MEIGA_SCREENING-
MN/2_alignment/PD0266a_5/PD0266a_5.sor
ted.bam ont
/mnt/netapp2/genomas_enfermedad/1/0_reference/0_hg38/GCA_000001405.15_GRCh38_no_a
lt_analysis_set.fna /mnt/netapp2/uscmg_aplic/1_internal_tools/26_MEIGA-
LR/databases/GRCh38_v2/meiga --normalBam /mnt/netapp2/genomas_enfermedad/1
/1_projects/3051_MEIGA_SCREENING-MN/2_alignment/PD0266c/PD0266c.sorted.bam --
minClusterSize 4 --minReads 5 --polishing-rounds 1 --source-families 'L1' --
minMAPQ 15 --refs
chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,chr11,chr12,chr1
3,chr14,chr15,chr16,chr17,chr18,chr19,chr20,chr21,chr22,chrX,chrY -p
$SLURM_CPUS_PER_TASK --gene-annot-dir
/mnt/netapp2/uscmg_aplic/1_internal_tools/26_MEIGA-
LR/databases/GRCh38_v2/annovarDb -o /mnt/netapp2/genomas_enfermedad/1/1_projects
/3051_MEIGA_SCREENING-MN/9_various/03_MEIGA-LR/2.1.0/PD0266a_5_PD0266c
```

Código 10: Ejemplo de lanzamiento de MEIGA-LR en modo pareado.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.2.6 Inserciones somáticas de LINE1

Como se comentó anteriormente debido a que se han usado dos aproximaciones para la obtención de las inserciones somáticas de LINE1, estas se muestran por separado. Por un lado están las inserciones somáticas obtenidas directamente de MEIGA-LR en modo pareado, filtrando con el tejido adyacente (o sangre) secuenciado con lecturas largas (ONT), y por otro lado las inserciones somáticas de las muestras de las cuales sólo teníamos datos de lecturas largas para el tumor, en cuyo caso, se lanzó MEIGA-LR usando como pareja "normal" la muestra PD0277c, posteriormente hizo falta filtrar las inserciones somáticas obtenidas con los datos del tejido adyacente (o sangre) secuenciado con lecturas cortas (Illumina) y las inserciones somáticas obtenidas de ejecutar xTea sobre ellos.

Se muestra a continuación una tablas (25) con los resultados totales.

	Inserciones somáticas L1	Inserciones "Solo"	Inserciones "Partnerer"	Inserciones "Huérfanas"	Inserciones con origen conocido	Inserciones "Partnerer" > 6000 pb
Filtradas con tejido adyacente (ONT)	7121	5315	993	813	1806	24
Filtradas con tejido adyacente (Illumina +xTea)	1085	903	82	100	182	2
Totales	8206	6218	1075	885	1988	26

Tabla 25: Inserciones somáticas de LINE1 en nuestra cohorte de tumores.

7.2.6.1 Inserciones somáticas de LINE1 obtenidos filtrando con datos de lecturas largas (ONT)

El número total de inserciones somáticas de LINE1 en estas muestras es de 7121, de ellas 1806 son "partener" o huérfanas, permitiéndonos conocer qué elemento fuente las originó. Además de estas últimas 24 tienen más de 6000 pares de bases, por lo que en algunos casos nos va a permitir hacer comparaciones del LINE1 insertado y el elemento fuente, sobre todo la región del promotor.

El resto de las inserciones 5315 corresponden a inserciones "solo" de ellas 77 son de más de 6000 pb, lo que nos va a permitir hacer un estudio "global" del estado de sus promotores. Aunque actualmente no podemos conocer por medio de las traducciones su elemento fuente, en nuestro grupo se está desarrollando un algoritmo para inferirlo a partir de posiciones diagnóstico, es decir mutaciones puntuales que poseen los distintos elementos fuente que los diferencia del resto y que son comunes al elemento fuente como al elemento insertado. Para ello se usa una estadística bayesiana.

El *script* usado para obtener las gráficas, junto con las gráficas en alta resolución y las correspondientes a inserciones huérfanas y "partener" se encuentran en el repositorio adjunto.

<../notebooks/01_vision_general/20230403_descripcion_insercciones_somaticas_MEIGA_3.0.0_all_donor_00.ipynb>

<../4_resultados/0.Metilacion_general/1_inserciones_somaticas_MEIGA-LR_3.0.0_all_donor/>

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

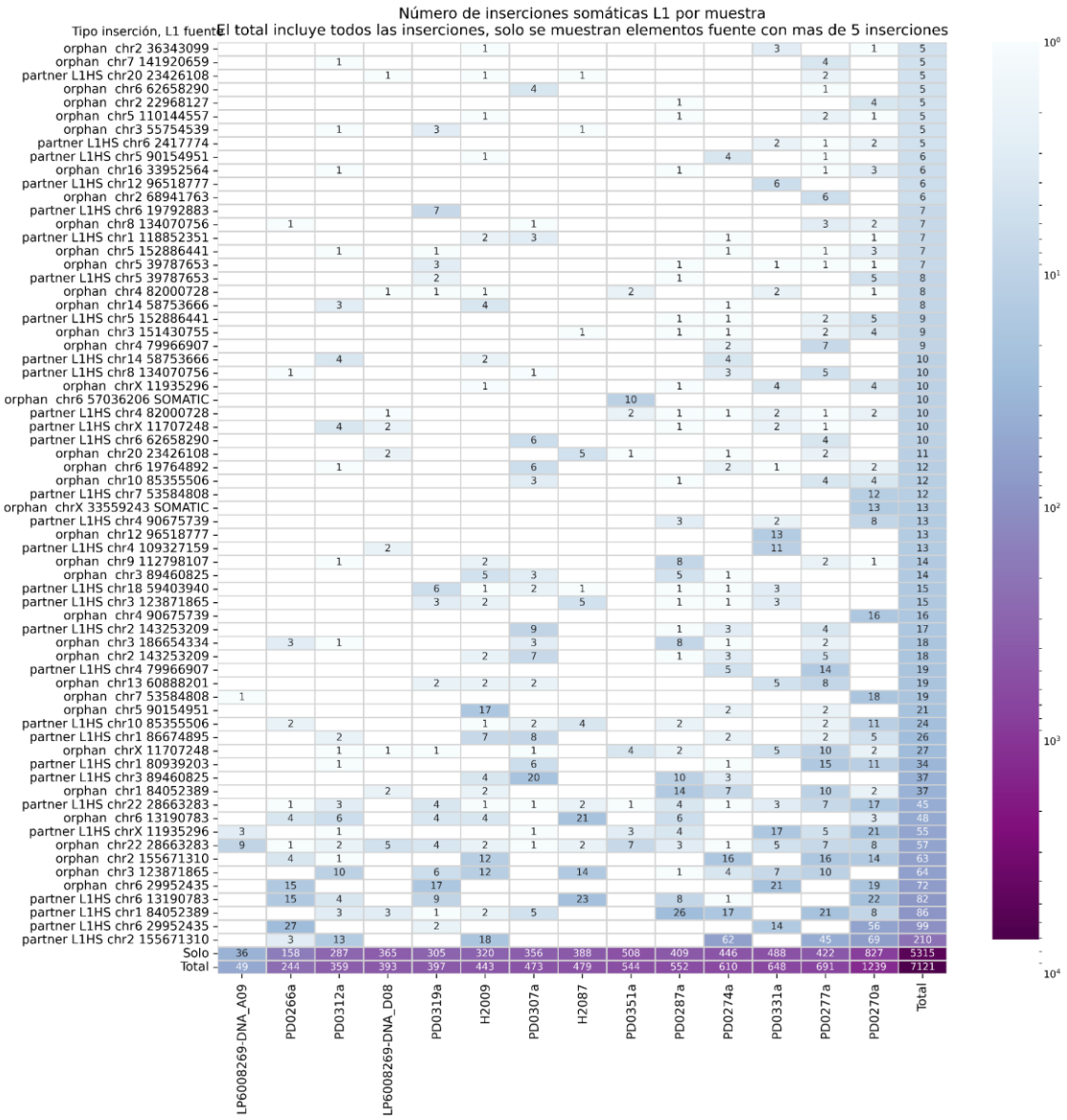


Tabla 26: Inserciones somáticas por muestra (muestras con tejido adyacente).

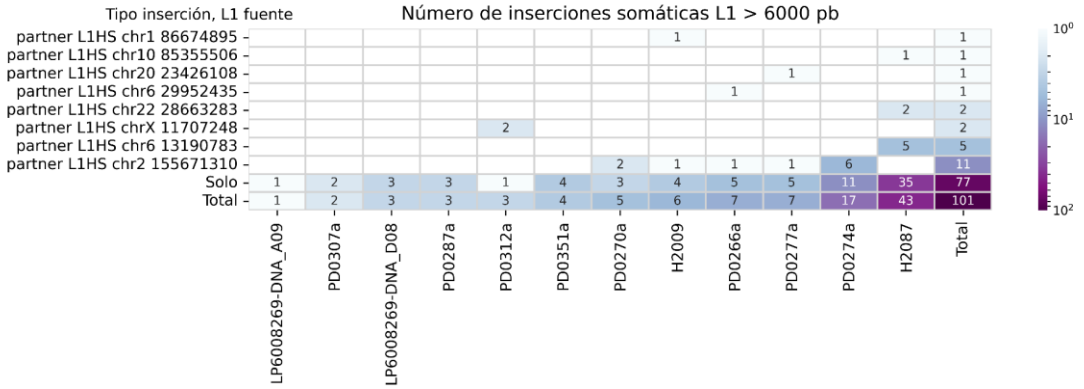


Tabla 27: Inserciones somáticas por muestra mayores de 6000 pares de bases (muestras con tejido adyacente).

7.2.6.2 Inserciones somáticas de LINE1 obtenidos filtrando con datos de lecturas cortas (Illumina + xTea)

Para las muestras de las cuales carecemos de datos de ONT de tejido adyacente (o sangre), hemos tenido que usar otra aproximación que consiste correr MEIGA-LR en modo pareado usando otra muestra no tumoral para eliminar la mayor parte posible de LINE1 germinales, en nuestro caso la muestra PD0277c. Posteriormente las inserciones somáticas fueron filtradas con los datos de las muestras de tejido adyacente correspondientes, obtenidos tras aplicar xTea (para lecturas cortas) sobre los datos de Illumina. Para ello se usó el siguiente código (bash).

<create_meiga_somatic_insertions_bed_filtering_xtea_paired.sh>

```
#!/bin/bash
# Script para  quitar (filtrar con xtea + illumina)  las inserciones germinales
de MEIGA-LR que se ha corrido en modo pareado con la muestra PD0277c
# (c) fjt 18.08.2022

module load bedtools
meiga_ouput_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/0
5_javi/2_meme/01_input_data_MEIGA-LR_.2.0.0/00_MEIGA-LR_2.0_somatic_tumor-
adjacent/
xtea_ouput_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05
_javi/2_meme/01_input_data_MEIGA-LR_.2.0.0/01_MEIGA-LR_2.0_somatic_tumor-
xtea_illumina/02_xTea_paired/
output_dir=/mnt/lustre/hsm/nlsas/notape/home/usc/mg/res_ge/00_personal_dir/05_jav
i/2_meme/01_input_data_MEIGA-LR_.2.0.0/01_MEIGA-LR_2.0_somatic_tumor-
xtea_illumina/02_xTea_paired/output/

dir=(
"12T_FIDIS_PD0277c"
"LP6005334-DNA_A03_PD0277c"
"LP6005935-DNA_C04_PD0277c"
"LP6008141-DNA_H01_PD0277c"
"LP6008221-DNA_E01_PD0277c"
"LP6008221-DNA_H02_PD0277c"
"LP6008269-DNA_B10_PD0277c"
"LP6008334-DNA_H02_PD0277c"
"PD0299a_PD0277c"
"PD0341a_PD0277c"
"PD0353a_PD0277c")

array=(
"12T_FIDIS/output/12T_FIDIS_12B_FIDIS/L1"
"LP6005334-DNA_A03/output/LP6005334-DNA_A03_LP6005333-DNA_A03/L1"
"LP6005935-DNA_C04/output/LP6005935-DNA_C04_LP6005934-DNA_C04/L1"
"LP6008141-DNA_H01/output/LP6008141-DNA_H01_LP6008138-DNA_H01/L1"
"LP6008221-DNA_E01/output/LP6008221-DNA_E01_LP6008220-DNA_E01/L1"
"LP6008221-DNA_H02/output/LP6008221-DNA_H02_LP6008220-DNA_H02/L1"
"LP6008269-DNA_B10/output/LP6008269-DNA_B10_LP6008268-DNA_B10/L1"
"LP6008334-DNA_H02/output/LP6008334-DNA_H02_LP6008335-DNA_D02/L1"
"PD0299a/output/PD0299a_PD0299b/L1"
"PD0341a/output/PD0341a_PD0341c/L1"
"PD0353a/output/PD0353a_PD0353c/L1")

for i in "${!array[@]}"; do
  printf "Processing %s and  %s\n" "${dir[i]}" "${array[i]}"
  mkdir -p "$output_dir""${dir[i]}"

  #INPUTS
```

```
   meiga="$meiga_ouput_dir""${dir[i]}"/INS_MEIGA.PASS.vcf

input_xtea_paired="$xtea_ouput_dir""${array[i]}"/aligned.sorted.marked_duplicates
_LINE1.vcf

   echo Input files
   ls $meiga
   ls $input_xtea_paired

   #OUTPUTS

output_xtea_paired="$output_dir""${dir[i]}"/aligned.sorted.marked_duplicates_LINE
1.vcf.sorted.bed
   ouput_file_bcftools="$output_dir""${dir[i]}"/INS_MEIGA.PASS.bed
   ouput_file_bedtools="$output_dir""${dir[i]}"/INS_MEIGA.PASS.SOMATIC.bed

ouput_file_meiga_somatic_L1HS="$output_dir""${dir[i]}"/INS_MEIGA.PASS.SOMATIC.L1H
S.tsv

   echo $output_xtea_paired
   echo $ouput_file_bcftools
   echo $ouput_file_bedtools
   echo $ouput_file_meiga_somatic_L1HS

   #create the bed files to intersect
   bcftools query -f '%CHROM\t%POS\t%END\t0\t%STRAND\n' "$input_xtea_paired"  |
sort -k1,1 -k2,2n > "$output_xtea_paired"
   bcftools query -f
'%CHROM\t%POS\t%END\t0\t%STRAND\tL1\t0\t1\t%ITYPE,%CYTOID\t%VTYPE\t%INSEQ\n'
"$meiga"  | sort -k1,1 -k2,2n > "$ouput_file_bcftools"

   bedtools intersect -a "$ouput_file_bcftools" -b <( awk '{print $1, $2-100,
$3+100}' "$output_xtea_paired" | tr ' ' '\t' | sort -k1,1 -k2,2n) -wa >
"$ouput_file_bedtools"
   echo "chrm  start   end     size    orientation     type    in_ref  is_somatic
family  tsd     consensus_seq" > "$ouput_file_meiga_somatic_L1HS"
   cat "$ouput_file_bedtools"  >> "$ouput_file_meiga_somatic_L1HS"
done
```

Código 11: Eliminación de inserciones germinales usando datos de xTea.

Los datos obtenidos se muestran en las siguientes tablas. En ellas podemos comprobar que se han descubierto 1085 inserciones somáticas, de ellas la mayoría (903) son inserciones "solo", con 52 de más de 6000 pb. Se han descubierto 182 inserciones "partener" y huérfanas, de las cuales al tener información del elemento fuente del cual provienen por lo que nos va a permitir estimar la actividad de los elementos germinales íntegros de estas muestras. De estas 182 inserciones, 2 son de más de 6000 pb, lo que nos va a permitir comparar, (en principio) el estado de metilación del elemento fuente y de su copia insertada.

En el repositorio se pueden visualizar estas gráficas y otras del número de inserciones "partener" y huérfanas, con alta resolución.

<../4_resultados/0.Metilacion_general/2_inserciones_somaticas_MEIGA-LR_2.0.0_Illumina-xtea/>

Así mismo se encuentra en él los *scripts* usados para generarlas

<../notebooks/0_vision_general/20230317_descripcion_insercciones_somaticas_filtradas_illumina_xtea_00.ipynb>

USC DE SANTIAGO DE COMPOSTELA

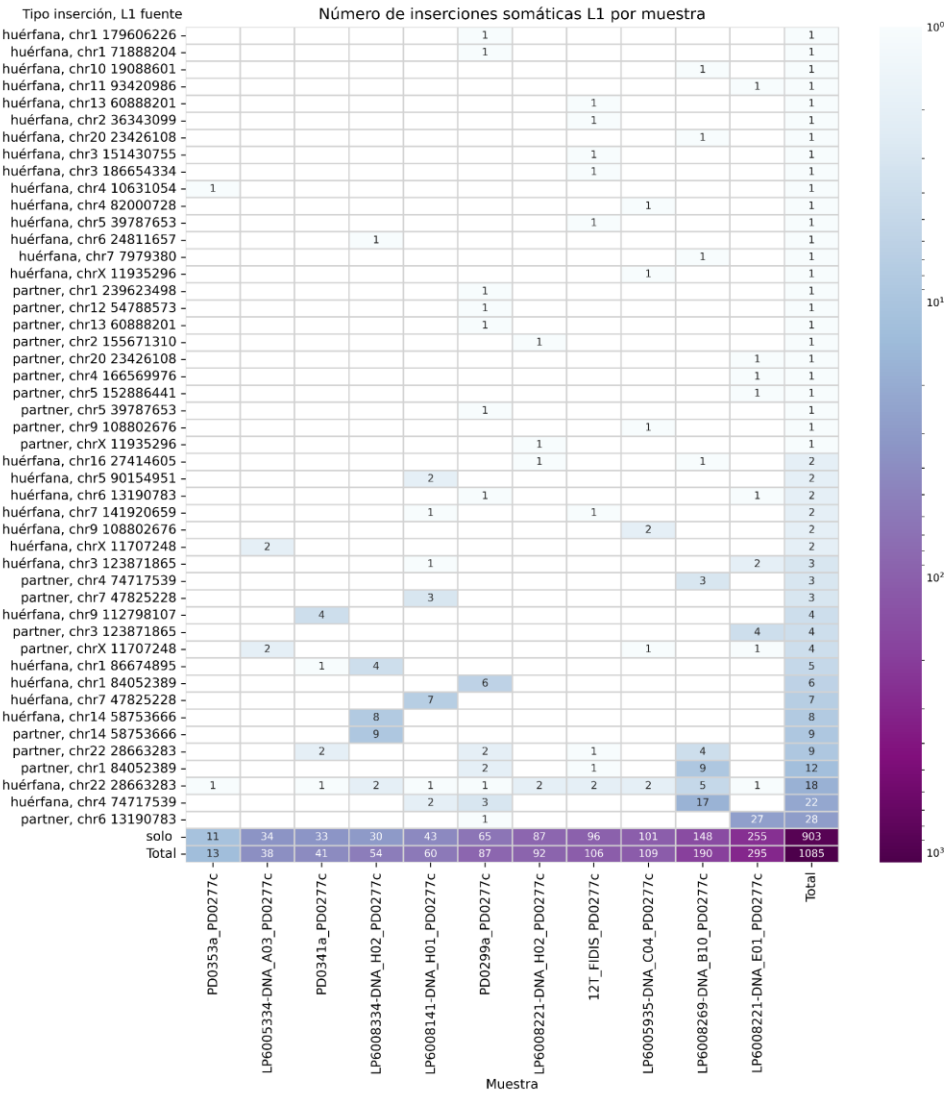


Tabla 28: Inserciones somáticas por muestra (muestras filtradas con xTea).

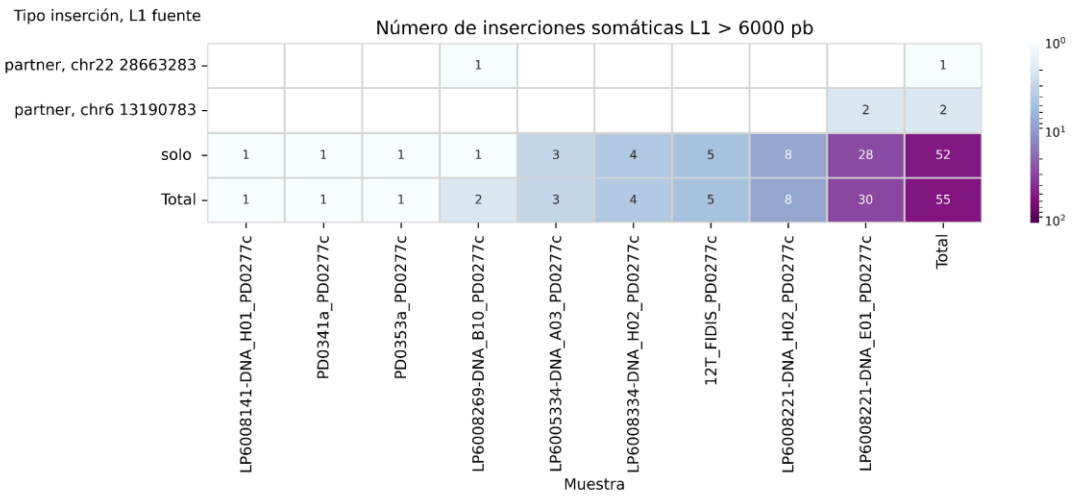


Tabla 29:Inserciones somáticas por muestra mayores de 6000 pares de bases (muestras filtradas con xTea)

8. RESULTADOS Y DISCUSIÓN III. RESULTADOS BIOLÓGICOS

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Vida”. Museo de Historia Natural USC. Acuarela y rotulador sobre libreta de dibujo (Dic 2018)

8.1 ANÁLISIS GENERAL DE LAS MUESTRAS

8.1.1 Metilación

Después de considerar los análisis mencionados en secciones anteriores (6.4 Comparativa de métodos gráficos y estadísticos de metilación general), tomé la decisión de utilizar datos binarizados en lugar de la probabilidad de modificación epigenética, ya sea metilación o hidroximetilación, por sitio CpG. Si bien esto supone la pérdida de aproximadamente el 10% de las observaciones, esta elección simplifica considerablemente los análisis, ya que los valores intermedios sólo introducen ruido en los resultados e inclusive, su eliminación no supone un impacto significativo en los cálculos finales de metilación de las características de interés. Además, desde un punto de vista biológico, tiene mucho más sentido, ya que cada sitio CpG en las cadenas de ADN sólo puede estar en uno de dos estados: metilado o desmetilado.

A continuación, se presenta un gráfico de violín resumiendo los niveles de metilación genómica calculados por ventanas de 10 Kb tanto de las parejas de tejido normal, ya sea sangre o tejido adyacente, y tumor primario como de las distintas líneas celulares analizadas en esta Tesis doctoral. Como se puede observar, la metilación global es en promedio mayor en el tejido normal que en los tumores, hecho ampliamente respaldado por la literatura científica (A. P. Feinberg y Vogelstein 1983; Florl et al. 2004; Jones y Baylin 2007; Esteller 2008; Sandoval y Esteller 2012; A. Feinberg 2014). En contraposición, las parejas de líneas celulares analizadas poseen un perfil de metilación aparentemente opuesto, con el equivalente "normal" marcadamente menos metilado que su correspondiente línea tumoral, tal y como ya se ha mencionado brevemente en la sección (6.4.1.1). Esto podría deberse al hecho de que estas líneas celulares son en realidad linfocitos inmortalizados, cuyo estado de metilación puede diferir del de los linfocitos no inmortalizados. De hecho, en el análisis disponemos de dos muestras de linfocitos primarios, correspondientes a pacientes con cáncer de esófago, denominadas LP6002868-D08 y LP6002868-A09, las cuales presentan un patrón de metilación consistente con los tejidos normales también mostrados en los gráficos. A destacar también que en términos generales los niveles de metilación son relativamente homogéneos entre las ventanas de 10 Kb de los genomas de las muestras "normales", mientras que en sus tumores poseen una distribución más dispersa dentro de la ya mencionada tendencia general a la hipometilación. Una posible explicación, además de la posible contaminación con tejido no tumoral, es que los tumores son heterogéneos en su genética y epigenética. Así, incluso si la tendencia general es la hipometilación, puede darse heterogeneidad entre distintas regiones genómicas. Asimismo, dentro de una muestra tumoral pueden coexistir varios subclones celulares con características epigenéticas divergentes. La excepción, nuevamente, son las líneas celulares NCI-BL2009 y NCI-BL2087, con una distribución más heterogénea y amplia, asemejándose de hecho más a las muestras tumorales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

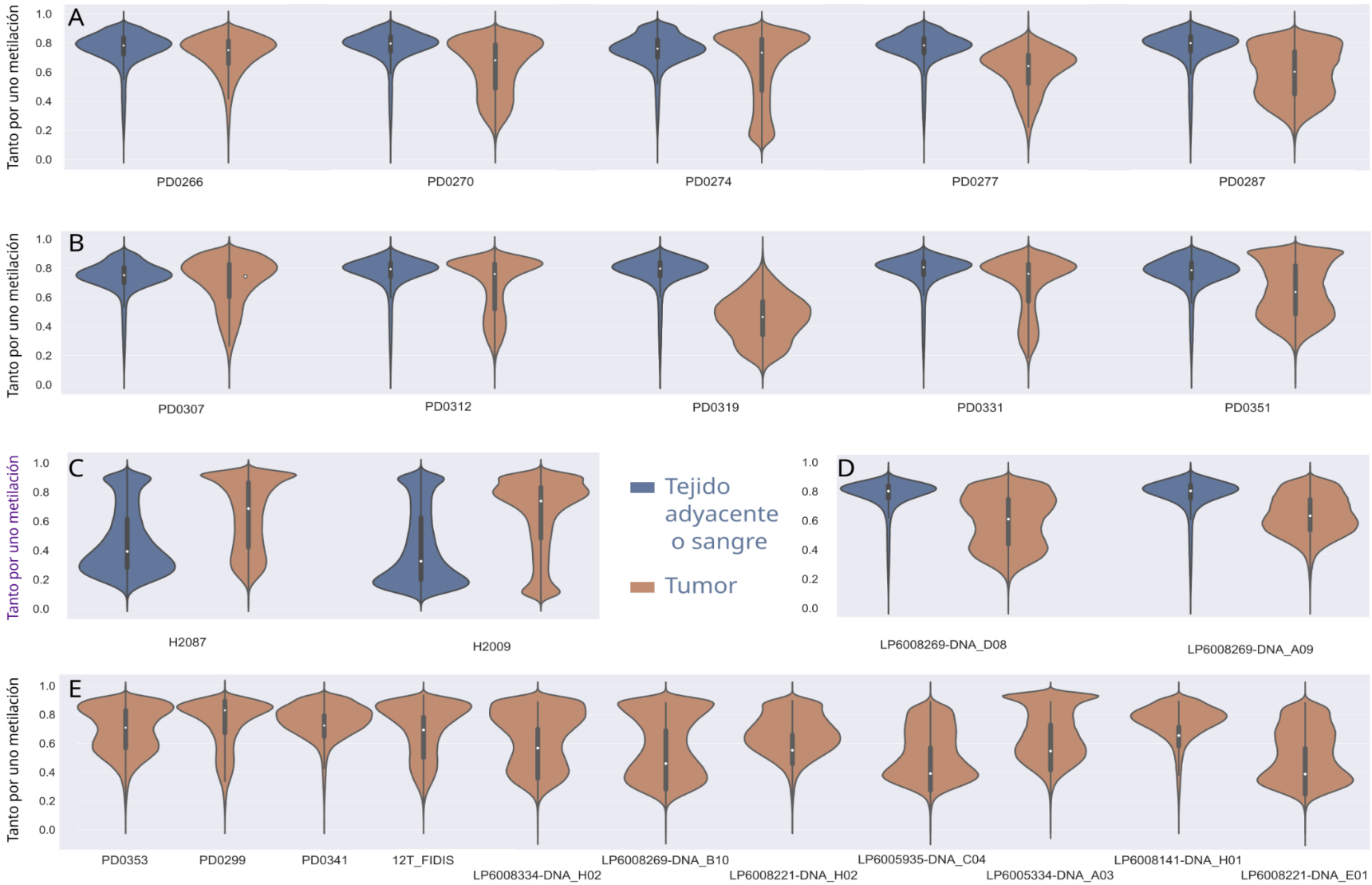


Figura 123: Gráfica de violín de la metilación global de tumores y tejido adyacente de nuestra cohorte (media por ventana de 10 Kpb).

La tabla con el resumen de los datos se muestra a continuación (tabla 30). En ella que se puede apreciar una disminución generalizada de la metilación en el tumor con respecto al tejido adyacente, siendo esta información complementaria a las gráficas de violín arriba expuestos.

Muestra	Media Tej. Normal	Des. standard	Media Tumor	Des. standard	Diferencia en la media
12T_FIDIS			0.679	0.467	
NCI-H2009	0.429	0.495	0.601	0.490	-0.172
NCI-H2087	0.481	0.500	0.661	0.473	-0.180
LP6005334-DNA_A03			0.675	0.468	
LP6005935-DNA_C04			0.522	0.500	
LP6008141-DNA_H01			0.698	0.459	
LP6008221-DNA_E01			0.503	0.500	
LP6008221-DNA_H02			0.644	0.479	
LP6008269-DNA_A09	0.740	0.438	0.642	0.479	0.098
LP6008269-DNA_B10			0.585	0.493	
LP6008269-DNA_D08	0.740	0.439	0.607	0.489	0.133
LP6008334-DNA_H02			0.619	0.486	
PD0266	0.714	0.452	0.690	0.462	0.024
PD0270	0.727	0.445	0.611	0.488	0.116
PD0274	0.716	0.451	0.613	0.487	0.103
PD0277	0.724	0.447	0.585	0.493	0.139
PD0287	0.721	0.449	0.581	0.493	0.140
PD0299			0.683	0.465	
PD0307	0.695	0.460	0.679	0.467	0.016
PD0312	0.716	0.451	0.660	0.474	0.056
PD0319	0.718	0.450	0.442	0.497	0.276
PD0331	0.723	0.447	0.671	0.470	0.053
PD0341			0.730	0.444	
PD0351	0.725	0.446	0.651	0.477	0.074
PD0353			0.663	0.473	

Tabla 30: Metilación global de las muestras de nuestra cohorte.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Permítanme destacar la odisea que fue recrear la figura 123, una experiencia que seguramente resultará familiar a todos aquellos que se hayan enfrentado a las alegrías y desafíos de la programación en Python. Nuestro *script* tuvo sus humildes inicios, afrontando la computación de una modesta selección de datos que, aunque reducida, creímos representativa. Una vez que se comprobó que el *script* estaba funcionando correctamente, llegó el momento de escalar para abarcar todo el conjunto de datos. Y ahí es cuando un proyecto que debería haberse completado en una semana se convirtió en una Ilíada de más de un mes. Incluso con la ayuda de un cuasi todopoderoso nodo del CESGA con 60 procesadores y 250 GB de memoria, el programa decidía cerrarse abruptamente sin más explicaciones que un desolador "*kernel died*". ¿No es acaso una frase reconfortante? Como bien expresó un sabio usuario anónimo de Python, "*descubrir la causa de la muerte del kernel en el notebook de Jupyter puede ser desalentador en ocasiones. Cuando el kernel sucumbe debido a problemas con alguna biblioteca, es posible que jamás se obtenga ninguna pista sobre la verdadera causa*".

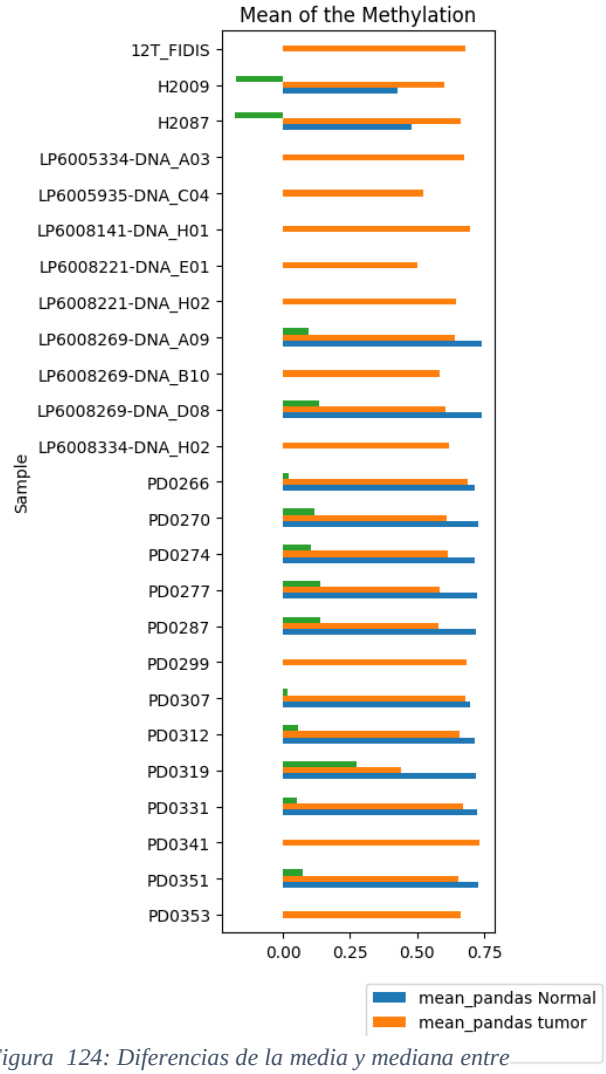


Figura 124: Diferencias de la media y mediana entre la metilación de tejido adyacente - metilación del tumor.

Y así fue, resultó imposible determinar qué causaba tan fatídico destino. Por lo tanto, las gráficas y estadísticas tuvieron que ser generadas individualmente para cada muestra, para luego armar la gráfica final uniendo las imágenes como si estuviéramos jugando a un rompecabezas, resultado que les he presentado en la figura 123.

8.1.1.1 Metilación de las líneas celulares

Debido a los extraños resultados de la metilación de las líneas celulares manejadas en este trabajo, decidí hacer una serie de comprobaciones complementarias. Para ello descargué los datos de metilación de bisulfito de todo el genoma (WGBS) procedentes de (Vidal et al. 2017) consistentes en varias réplicas de tejido pulmonar sano y un par de réplicas de una línea celular de un cáncer de pulmón de células no pequeñas (NCI-H1437). Los datos fueron procesados usando el protocolo descrito en la sección 5.2.4 (Alineación de datos de metilación. Pretratamiento con bisulfito + Illumina + Bismark). Los datos y los gráficos resultantes se muestran a continuación, y en ellos puede comprobarse que tanto el tejido normal como el tumor siguen los perfiles esperados estando el tejido normal mucho más metilado que el tumor. En contraposición, nuestras líneas celulares tumorales, NCI-H2009 y NCI-H2087, están casi tan metiladas como el tejido normal de pulmón y su equivalente

"tejido normal" los linfocitos inmortalizados NCI-BL2009 y NCI-BL2087 están mucho menos metilados que el tejido normal pulmonar, con una metilación correspondiente a un tejido tumoral. Entre otras alternativas se barajó que los datos de nuestras muestras de las líneas celulares de tumores y de sangre se habían intercambiado en algún punto, si bien comprobaciones ortogonales indican lo contrario. Los datos de metilación son concordantes entre los obtenidos en nuestro laboratorio por secuenciación de lecturas largas + Megalodon y los generados por WGBS + Bismark (secuenciados externamente en Macrogen). Además, los datos de los que se dedujo la metilación son compatibles con el número de eventos de inserciones somáticas de retrotransposones asignados a cada una de las muestras tumorales, siendo estas mucho más altas en las líneas celulares tumorales que en los linfocitos inmortalizados. También son compatibles con los datos de mutaciones puntuales y el número de copias (*copy number*) obtenidos por GATK (Genome Analysis Toolkit) HaplotypeCaller y Battemberg respectivamente. Una vez descartados posibles tras papeleos de muestras y artefactos técnicos o informáticos, hemos de asumir que la hipermetilación de estas líneas celulares es real e intrínseco a las mismas, sin que sepamos la verdadera causa de ello.

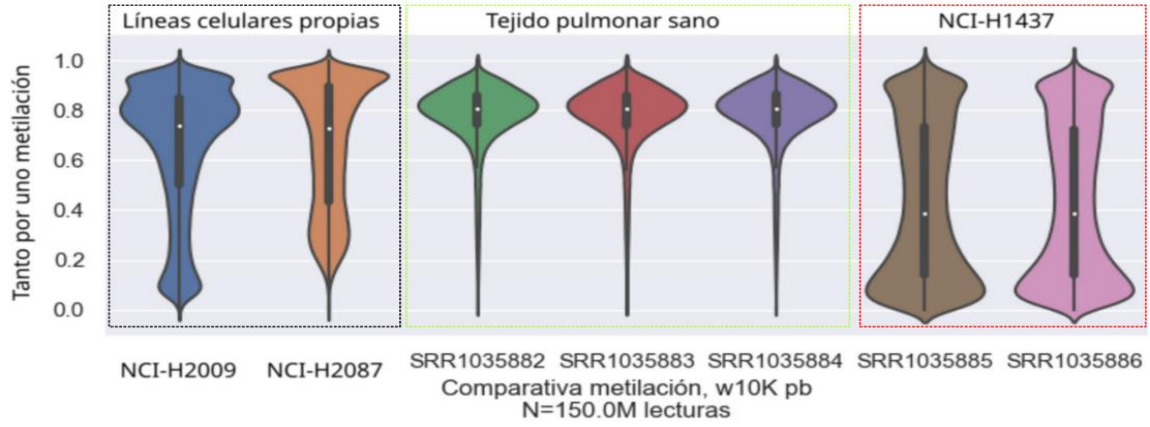


Figura 125: Comprobación de la metilación de las líneas celulares.
Comparativa de las líneas celulares NCI-H2009 y NCI-H2087 con las lineas celulares obtenidas de Vidal et al 2017 (gráfica de violín con la media por ventanas de 10Kpb)

Muestra	Tipo	Media Bismarck	Media Megalodon
NCI-H2009	Línea celular (tumor)	65.49	60.26
NCI-H2087	Línea celular (tumor)	70.4	66.38
NCI-BL2009	Línea celular (normal)		42.96
NCI-BL2087	Línea celular (normal)		48.07
H1437-Replica 1 (SRR1035885)	Línea celular (tumor)	49.34	
H1437-Replica 2 (SRR1035886)	Línea celular (tumor)	48.29	
Replica 1 (SRR1035882)	Tejido pulmonar sano	74.89	
Replica 2 (SRR1035883)	Tejido pulmonar sano	75.03	
Replica 3 (SRR1035884)	Tejido pulmonar sano	75.14	

Tabla 31: Comprobación de la metilación.
Comparativa de las líneas celulares NCI-H2009 y NCI-H2087 con las lineas celulares obtenidas de Vidal et al 2017

Es importante tener en cuenta que los análisis comparativos de líneas celulares deben ser interpretados con mucha precaución, ya que se evalúan líneas celulares obtenidas a partir de tumores de pulmón con líneas celulares obtenidas a partir de linfocitos inmortalizados del mismo paciente. Como resultado, los patrones de metilación pueden diferir significativamente entre estas dos fuentes de células y drásticamente con respecto a sus tejidos originales.

8.1.2 Hidroximetilación

La tecnología de secuenciación de nanoporos ofrece múltiples ventajas, ya que no solo permite detectar la metilcitosina en contextos CpG, sino que también es capaz de identificar esta modificación en otras citosinas y otras modificaciones de diferentes nucleótidos. Para lograr esto, se han desarrollado algoritmos y redes neuronales que tienen la capacidad de "aprender" a reconocer estas diversas modificaciones. La empresa ONT ha invertido un gran esfuerzo en entrenar estas redes neuronales, no solo para detectar la metilación en el contexto CpG, sino también en otros contextos como CHH y CHG. Además, se ha dedicado considerable atención a distinguir entre la metilación y la hidroximetilación, así como a identificar metiladeninas. Frente a los procedimientos tradicionales con pretratamiento de bisulfito que tienden a enmascarar la hidroximetilación como metilación, lo que dificulta su discernimiento, la secuenciación directa de ONT posibilita su distinción. Esto permite explorar simultáneamente el papel único que desempeñan estas diversas modificaciones en la modulación de la expresión génica.

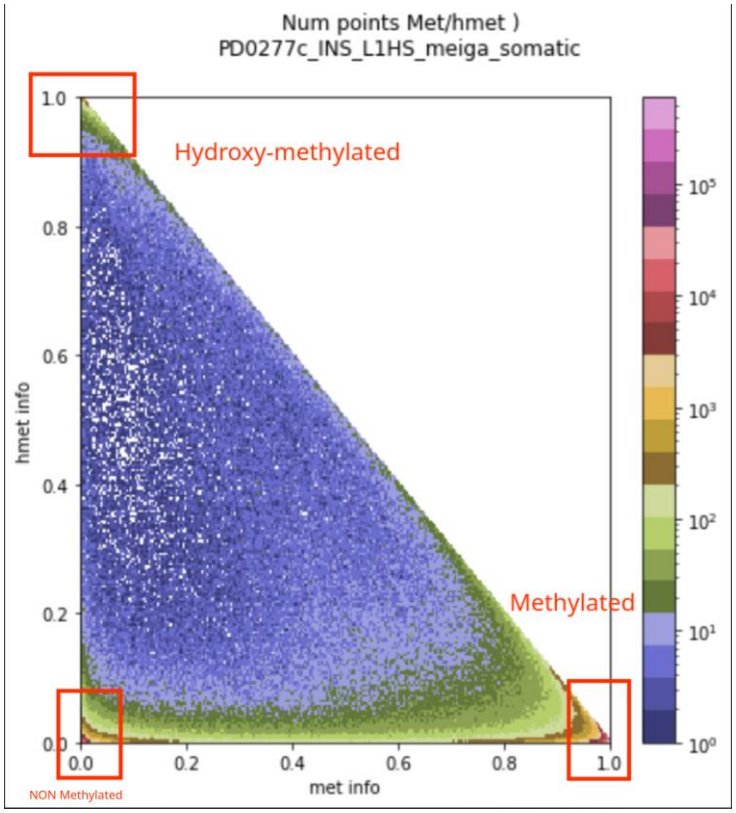


Figura 126: Mapa de metilación - hidroximetilación de la muestra PD0277c

Megalodon informa simultánea e independientemente del logaritmo en base 2 de la probabilidad de metilación y la probabilidad de hidroximetilación para cada citosina (en el contexto CpG) en el genoma de referencia. Esta independencia se visualiza en la figura 126, donde se comparan las probabilidades de ambas modificaciones para cada CpG en el genoma. Podemos observar que la mayoría de los puntos se encuentran en los vértices del triángulo de distribuciones. Estos vértices representan una alta probabilidad de metilación y baja probabilidad de hidroximetilación, una baja probabilidad de metilación y alta probabilidad de

hidroximetilación, y una baja probabilidad de ambas. En consecuencia, los primeros se asocian con el estado metilado, los segundos con el estado hidroximetilado, y los últimos con el estado no metilado.

Se llevó a cabo un análisis similar al de la metilación general de las muestras, pero esta vez enfocado en la hidroximetilación utilizando el *script* indicado en el notebook

<../notebooks/0_vision_general/20230307_vision_general__hidroxy_metilacion_binary_win_10k_04.ipynb>.

Sin embargo, no se pudo identificar ningún patrón regular ni una acumulación significativa de citosinas hidroximetiladas a lo largo del genoma, lo que sugiere que dicho proceso puede ser estocástico en su naturaleza. Además, se observó que los valores de hidroximetilación son excepcionalmente bajos. De manera general, se pudo notar que los tejidos adyacentes o la sangre presentaban valores ligeramente superiores (en el orden de unas pocas décimas porcentuales) en comparación con los tumores correspondientes (Figura 127). En base a estos hallazgos, no parece que la hidroximetilación desempeñe un papel relevante en la modulación genómica de estas muestras.

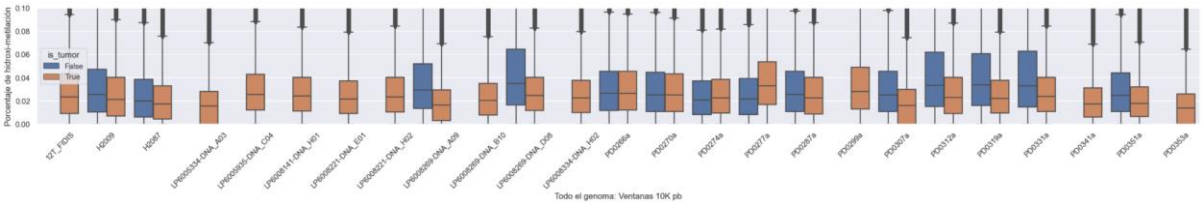


Figura 127: Gráfica de la hidroximetilación global de tumores y tejido adyacente de nuestra cohorte.

Cuando nos centramos en áreas específicas del genoma, como los LINE1, no se observa ningún patrón discernible en la hidroximetilación de las citosinas. Su distribución parece ser aleatoria, lo que ha llevado a que no se haya realizado un estudio más detallado sobre este aspecto. No obstante, se realizará un breve análisis de esta distribución junto con el análisis de la metilación de los LINE1, tanto en elementos germinales como somáticos.

8.2 METILACIÓN DE LINE1 GERMINALES Y REGIONES ADYACENTES EN GENOMAS TUMORALES Y NORMALES.

Durante más de un año, me dediqué al minucioso análisis de la metilación de los LINE1 germinales íntegros aplicando enfoques similares a los empleados por otros investigadores los cuales utilizaron lecturas cortas de Illumina. He explorado el conjunto de datos de nuestras muestras para detectar regiones con metilación diferencial, combinándolo con diferentes métodos para el análisis de variantes estructurales mediante lecturas largas, aunque no específicamente centrados en LINE1, dada la singularidad, novedad del campo de trabajo y la tecnología. Desafortunadamente, la mayoría de estos análisis no han arrojado resultados significativos, lo cual ha requerido el desarrollo de herramientas informáticas dedicadas para avanzar en la investigación. Además, uno de los desafíos más significativos que he enfrentado durante este proceso ha sido la problemática de la contaminación cruzada entre el tejido tumoral y el tejido adyacente. Con frecuencia, se ha observado la infiltración de células del tejido adyacente o del sistema inmunitario en el tejido tumoral y viceversa. Esta contaminación mutua ha planteado dificultades considerables en mis investigaciones y ha requerido una atención especial, ya que es esencial abordar adecuadamente este problema para garantizar la validez y fiabilidad de los resultados obtenidos en mi estudio.

8.2.1 Búsqueda de regiones diferencialmente metiladas.

El estudio de la metilación del ADN es un campo que lleva firmemente establecido durante años, lo que ha llevado al desarrollo de numerosas herramientas computacionales para descubrir y analizar regiones con metilación diferencial en datos de bisulfito de genoma completo (WGBS). Normalmente, estos programas se aplican a regiones muy amplias, como genes completos, cromosomas enteros o incluso genomas completos y, por lo general, fueron diseñados bajo la concepción de que la unidad de control génico no es la metilación de CpGs individuales, sino de grupos de CpGs concentradas en elementos reguladores, las islas CpG. Esto supone un problema para la identificación de regiones con metilación diferencial en los LINE1, ya que estos tienen un tamaño reducido, de alrededor de 6000 pares de bases, así como una distribución no uniforme de CpGs en su secuencia. Entre múltiples opciones optamos por utilizar el programa RADMeth desarrollado por Egor Dolzhenko y Andrew D. Smith de la Universidad del Sur de California. RADMeth es un software que permite analizar regiones genómicas con metilación diferencial mediante un análisis de regresión de metilación. Puede aplicarse tanto a regiones genómicas completas como a loci individuales en datos de secuenciación de bisulfito de genoma completo.

(http://smithlabresearch.org/wp-content/uploads/radmeth_manual.pdf (accedido sep. 2020)

Adaptar los datos de la plataforma de Nanopore para ajustarse a los archivos de entrada para este programa, entre otros, el design_matrix.txt y el radmeth_proportion_table.tsv, demostró ser un desafío considerable. Una vez solventados y debidamente regenerados, ejecutamos el programa en la muestra PD0277a (fragmento de código12) analizando diversas regiones, como las regiones corrientes arriba (20 Kpb) y corriente abajo (20 Kpb) del alelo lleno y vacío del LINE1 heterocigoto: chr1:184845616-184851641 (6025 pares de bases). Esta elección y proceso nos permitió explorar la capacidad de RADMeth para identificar regiones con metilación diferencial en los LINE1 y obtener información relevante para nuestro estudio.

```
Wand -factor case design_matrix.txt radmeth_proportion_table.tsv > cpgs.bed
adjust -bins 1:100:1 cpgs.bed > cpgs.adjusted.bed
dmrs -p 0.1 cpgs.adjusted.bed > dmrs1.bed
dmrs -p 0.5 cpgs.adjusted.bed > dmrs1.bed
adjust -bins 1:10:1 cpgs.bed > cpgs.adjusted.bed
adjust -bins 1:1:1 cpgs.bed > cpgs.adjusted.bed
adjust -bins 2:10:2 cpgs.bed > cpgs.adjusted.bed
adjust -bins 1:1000:1 cpgs.bed > cpgs.adjusted.bed
awk '$5 < 0.01 "{ print $0; $}"' cpgs.adjusted.bed > dm_cpgs.bed
dmrs -p 0.8 cpgs.adjusted.bed > dmrs1.bed
dmrs -p 0.01 cpgs.adjusted.bed > dmrs.bed
```

Código 12: Ejecución de RADMeth.

Lamentablemente, los resultados obtenidos para ese elemento y otros varios (no mostrados) fueron negativos, ya que no se encontró ninguna región diferencialmente metilada en las dos regiones de 20 Kpb pares de bases, lo que nos llevó a descartar el uso del programa en regiones más pequeñas, como dentro de los propios LINE1. Es posible que este resultado se deba a que el programa no está específicamente diseñado para el análisis de datos de metilación de lecturas largas ONT. El notebook para obtener los ficheros se encuentra en el repositorio adjunto:

<../notebooks/01_LINE1_germial/radmeth.ipynb>

8.2.2 Estimación de la pureza basada en la metilación

Tal como se comentó en el apartado (7.1.1.1 Pureza de los tumores) los tumores primarios son de naturaleza heterogénea, conteniendo células tumorales, células del estroma, del sistema inmune y del tejido adyacente. Esto implica que, inevitablemente, cierta proporción de las lecturas corresponden al tumor y otras han de ser asignadas al resto de células no tumorales. Ante la marcada diferencia en los patrones de metilación entre los tumores y sus tejidos adyacentes, tanto a nivel general como específicamente en los LINE1, surgió la necesidad de investigar la posibilidad de separar las lecturas correspondientes al tumor de aquellas pertenecientes al tejido adyacente. Teóricamente, además, este enfoque permitiría calcular la pureza de las muestras basándose en el estado de metilación de los LINE1. Con el objetivo de estimar la pureza de las muestras tumorales basándonos en el estado de metilación de los LINE1 germinales íntegros, se aplicó el algoritmo K-means en Python (*klearn*) para agrupar las lecturas de cada LINE1 en grupos (*clusters*), utilizando los patrones de metilación en las CpGs. Se realizaron pruebas con agrupaciones K-means que variaban desde 1 hasta 5 grupos. Cabe destacar que cada CpG en cada lectura puede presentar tres posibles estados de metilación: metilado (1), no metilado (0) o indeterminado (nan). Dado que el algoritmo K-means no puede operar con valores indeterminados, fue necesario realizar una normalización previa de los datos. La estrategia seguida consistió en eliminar sistemáticamente las lecturas que contenían más del 10% de citosinas con valores indeterminados de metilación. En las lecturas con menos del 10% de valores indeterminados, se asignó el estado de metilación mayoritario a las CpGs con estado no asignado. A excepción de las líneas celulares, cuyas poblaciones mostraron una asignación más probable de K=1, se encontró que la asignación óptima para el algoritmo K-means en las muestras analizadas fue de K=2. Este resultado resultaba consistente con la hipótesis de la existencia

de dos subpoblaciones celulares distintas, una tumoral y otra normal, diferenciables según sus patrones de metilación. Las lecturas asignadas al grupo mayoritario fueron consideradas de origen tumoral, mientras que las asignadas al grupo minoritario se asumieron como posibles contaminaciones de células no tumorales. Posteriormente, se calculó la media de las proporciones de lecturas asignadas a ambos grupos y se correlacionaron con los resultados de pureza obtenidos mediante Battemberg. Se realizó un test de Pearson para evaluar la asociación entre estas dos medidas. El conjunto de datos, resultados y cómo se realizaron los cálculos se pueden consultar el notebook:

< ../01_L1_germinal/20221204_Pregunta1_L1_germinales_purity.ipynb >

Se realizaron análisis de correlación entre la pureza de las muestras tumorales y la metilación de diferentes regiones del LINE1, incluyendo el LINE1 completo, la región del promotor rica en CpG (las primeras 500 pb en la región 5' UTR) y el resto del LINE1 (cuerpo). Los resultados de estas correlaciones se presentan a continuación y se pueden observar en la figura 128

Se encontró una correlación fuerte entre la pureza de las muestras tumorales y la metilación de estas regiones del LINE1. El predictor más sólido de la pureza resultó ser el calculado utilizando la región del promotor, con un coeficiente de correlación de r = 0.69 y un valor p de 0.00029. Asimismo, se observaron correlaciones significativas moderadas con el LINE1 completo (r = 0.66, p-val = 0.00054) y el cuerpo del LINE1 (r = 0.58, p-val = 0.00379).

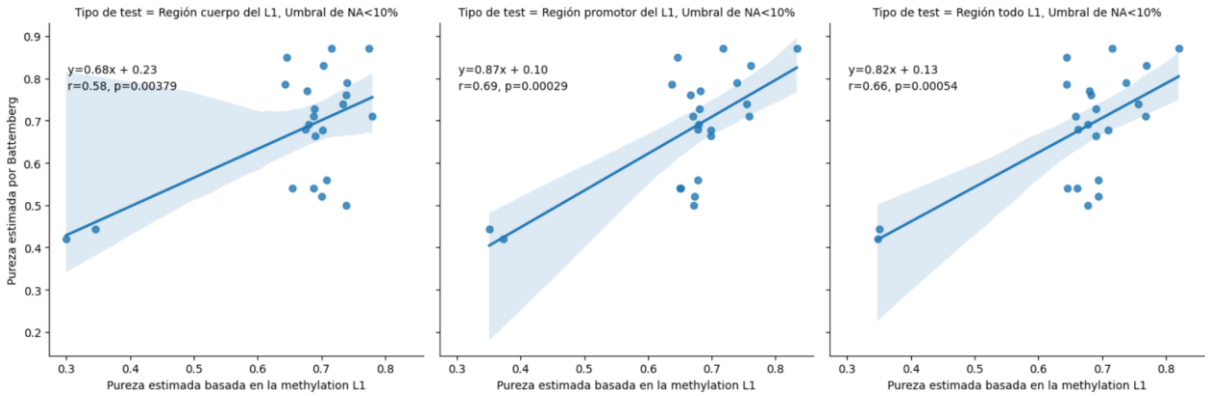


Figura 128: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1 I.

Aunque estos hallazgos sugerían que el estado de metilación en la región del promotor del LINE1 podía ofrecer una estimación confiable de la pureza de las muestras tumorales, es relevante destacar que dos de las 23 muestras analizadas (LP6008141-DNA_H01, LP6008269-DNA_A09) exhibieron una pureza muy baja. Según las estimaciones obtenidas mediante Battemberg, estas muestras presentaron una pureza inferior a 0.5, lo cual se encuentra fuera del rango considerado como confiable por dicha herramienta. Estos resultados indican la presencia de una significativa contaminación de células no tumorales en estas muestras, lo cual puede afectar la interpretación de los datos de metilación del LINE1 en estos casos específicos. Debido a ello, decidimos eliminar estos valores atípicos (*outliers*) y realizar el análisis de nuevo. En este caso hay marcadas diferencias entre el promotor y cuerpo del LINE1, (con los datos de todo el LINE1 con unos valores intermedios)

En este caso, observamos una correlación moderada y estadísticamente significativa ($p < 0.05$) entre la estimación de Battemberg y la correspondiente a la metilación en la región del promotor, mientras que la correlación con todo el LINE1 ($r = 0.41$, p-val = 0.0655) y sólo con el cuerpo ($r = 0.11$, p-val = 0.6) fueron no significativas (figura 129). Es interesante destacar que, a diferencia de nuestras expectativas, la correlación no mejora en comparación con el análisis previo. Esta menor correlación nos lleva a sospechar que podría tratarse de una fluctuación estadística. Es posible que con un conjunto de muestras más amplio podamos confirmar o refutar esta observación.

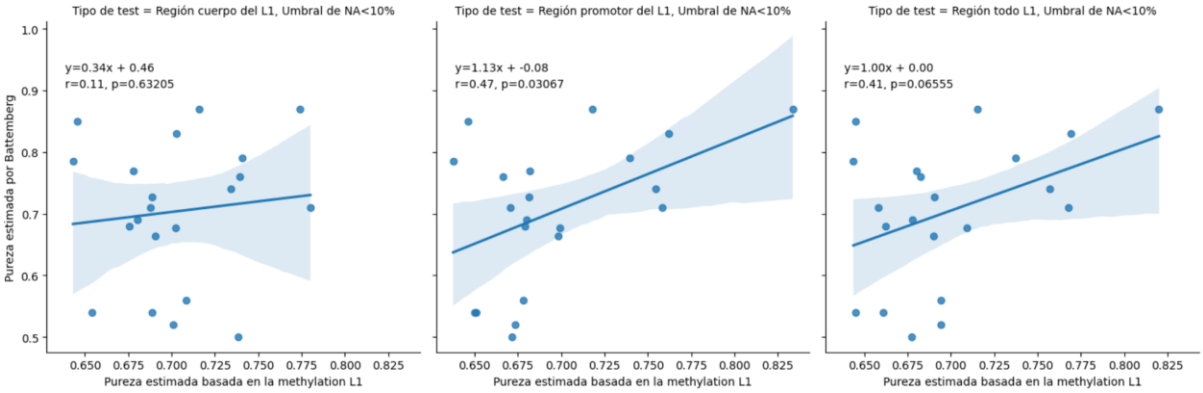


Figura 129: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1 II. (sin valores extremos (outliers))

A pesar de nuestros esfuerzos, no pudimos establecer ninguna correlación entre la diferencia de pureza estimada por Battemberg y la pureza estimada utilizando la metilación de los LINE1, en relación con la proporción de clonalidad en cada muestra calculada por Battemberg (figura 130). Estos resultados sugieren que las diferencias observadas en las estimaciones de pureza no pueden atribuirse a la clonalidad de las muestras.

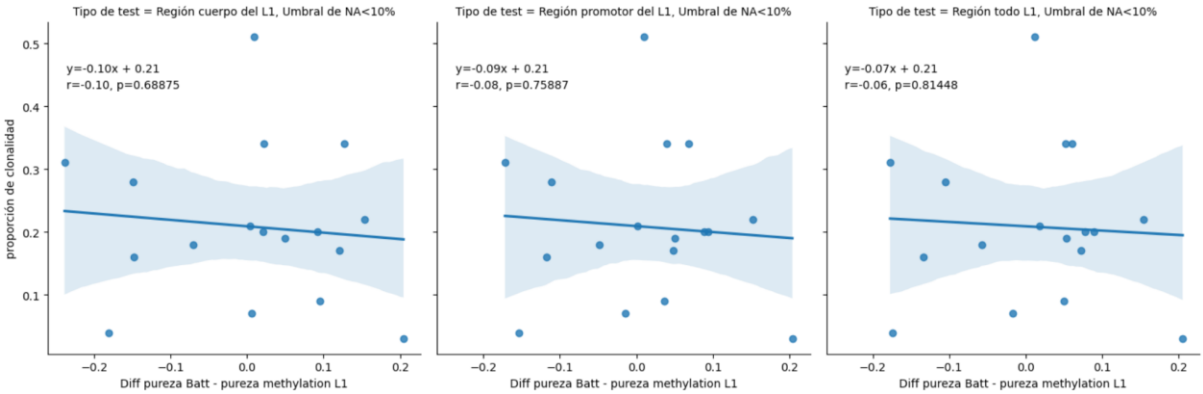


Figura 130: Correlación entre la clonalidad (estimada por Battemberg) y la pureza estimada usando el estado de metilación de los LINE1.

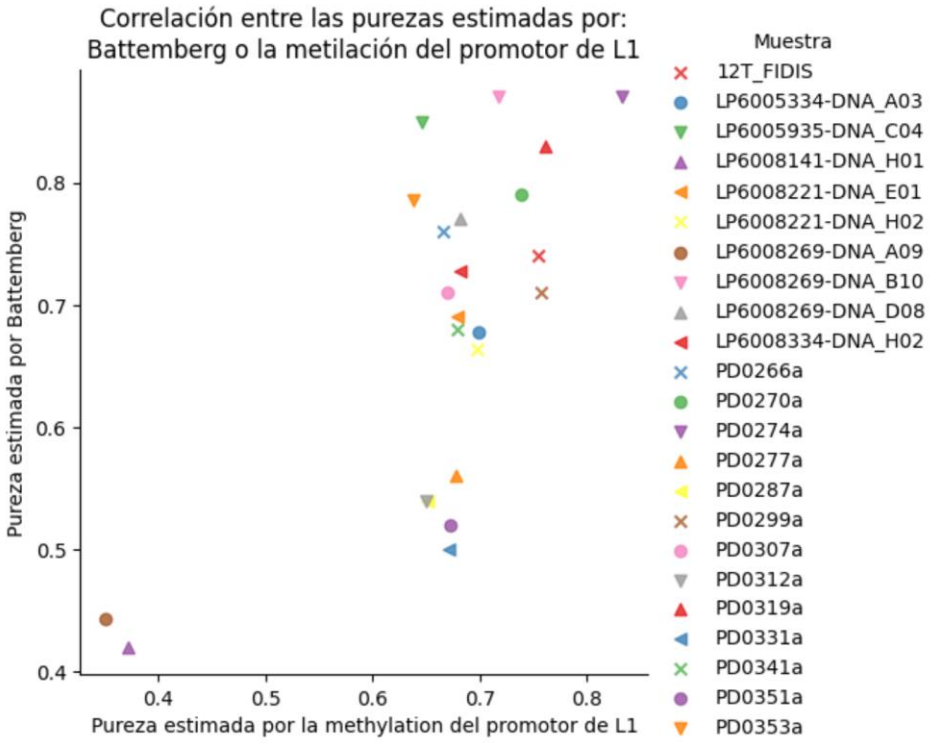


Figura 131: Correlación entre los valores de pureza estimados con Battemberg y en base a la metilación de los LINE1, por *muestra.*

Nuestros hallazgos respaldan que la utilización del estado de metilación en la región del promotor del LINE1 podría ser una herramienta confiable para estimar la pureza de las muestras tumorales. En todo caso, es importante destacar que se requieren futuras investigaciones para realizar correcciones adecuadas en la estimación de los niveles de contaminación utilizando la metilación de ONT, con el objetivo de acercarse a los estándares establecidos por Battemberg, considerado el "*Gold standard"* en este tipo de análisis. Estos avances podrían mejorar aún más la precisión de las estimaciones de pureza y fortalecer la utilidad de la metilación del LINE1 como una herramienta valiosa en la caracterización de la heterogeneidad tumoral.

8.2.3 Asignación de metil-alelos (epialelos)

En vista de las considerables discrepancias en los patrones de metilación entre el tejido normal adyacente y el tumoral mencionadas previamente, surgió la oportunidad de utilizar esta información con el propósito de distinguir las lecturas derivadas de ambos tejidos y asignarlas a sus respectivos tipos celulares. Esta tarea se puede llevar a cabo mediante un enfoque de deconvolución de lecturas basado en el estado de metilación, utilizando datos de secuenciación de obtenidos con la tecnología de ONT. Con tal propósito, procedí a agrupar las lecturas según su metilación en loci CpG, empleando el algoritmo *K-means* implementado en Python con la librería *sklearn*. Se testearon agrupaciones de 2 a 5 grupos (clústeres) de K-means. En la figura 132 se muestra a modo de ejemplo una gráfica del resultado de este análisis para una de las muestras tumorales analizadas en este trabajo, la PD0277a. Las barras verticales en los sub paneles horizontales representa cada uno de los elementos germinales

presentes en la muestra, ordenados en base a la proporción del grupo (k-mer) mayoritario. El primero de los paneles horizontales, en gama de grises, muestra la proporción de lecturas asignadas a cada uno de los Kmers (K=5). El panel en azul indica la media de metilación de cada elemento LINE1 considerando todas las lecturas, mientras que el histograma en rojo representa la media de metilación utilizando únicamente las lecturas asignadas al grupo (clúster) mayoritario, presumiblemente el tumoral. Por último, el subpanel verde representa la media de metilación para las lecturas que no pertenecen al grupo mayoritario. Esta representación gráfica nos permite visualizar de manera clara y concisa las diferencias en los niveles de metilación entre los diferentes grupos de lecturas.

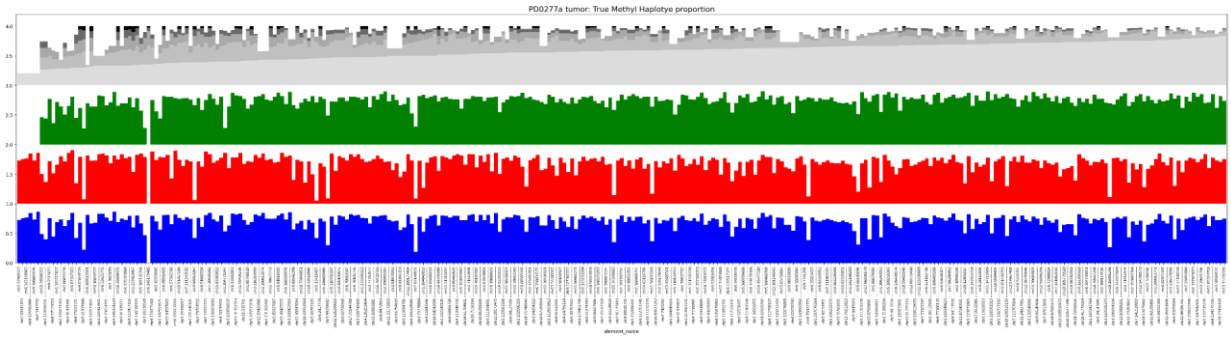


Figura 132: Ejemplo de grafica de la media de metilación por LINE1 germinal, separando las lecturas por k-means.

A pesar del potencial de esta aproximación, decidí no continuar su refinamiento por varias razones. En primer lugar, aunque en términos generales el tejido tumoral muestra una metilación notablemente inferior en comparación con su tejido adyacente, esta generalización no se aplica necesariamente de la misma manera a todos los elementos fuente ni a todas las muestras, lo que dificulta una identificación inequívoca y directa de los Kmers con los grupos predefinidos (tumor y normal). Además, el tumor tiende a presentar cierto grado de variabilidad entre los posibles clones en la muestra secuenciada, lo que puede dar la impresión de una contaminación con el tejido adyacente.

Después de discusiones internas con los demás miembros del proyecto, se llegó a la conclusión de que la estrategia óptima sería discriminar las lecturas no basándose en la información epialélica, sino en variantes puntuales (SNVs) presentes en los transposones o en las regiones adyacentes a los mismos. Estas variantes permitirían la segregación de las lecturas de ONT en tumorales y normales, y en caso de existir, en los posibles subclones mayoritarios presentes en el tumor. Una vez que se haya realizado el segregado en alelos (*phasing*) de las lecturas, se procedería a recuperar el estado de metilación de los grupos discriminados. Actualmente, el grupo se encuentra trabajando en una estrategia de llamada de variantes a partir de los datos de ONT, su filtrado y su segregación alélica, obteniendo resultados muy prometedores que permitirán concretar adecuadamente este análisis. Sin embargo, el éxito de esta estrategia dependerá de la presencia de una densidad suficiente de SNVs discriminantes y de las diferencias en los niveles de ploidía entre cada elemento para ambos tipos celulares en cada muestra.

Por razones evidentes, este análisis resulta complejo y su éxito está limitado, por lo que no se profundizó más en esta vía. No obstante, estos análisis preliminares me llevaron a

plantear el siguiente razonamiento: dado que dispongo de datos tanto del tejido adyacente de diez de los tumores, al combinar las lecturas del tumor con las del tejido sano, la contaminación correspondiente a células no tumorales debería agruparse junto con las lecturas del tejido sano. De manera sorprendente, este enfoque funcionó de manera efectiva, permitiendo la separación de las lecturas asociadas al tejido tumoral y al tejido sano, e incluso posibilitando la distinción de sus correspondientes epialelos a partir de la información de las regiones CpG de la región promotora, específicamente aquellas contenidas en los primeros 900 pares de bases de los elementos LINE1. A continuación, se presentan varios ejemplos representativos a modo de mapa de calor (*heatmap*), en el que se representan las lecturas como filas y las columnas se corresponden con los loci CpG en el promotor de un elemento fuente. El color de las celdas representa el estado de metilación de la citosina: El color crema claro indica una CpG metilada, mientras que los cuadrados negros representan una CpG no metilada. La primera anotación vertical (azul/marrón) indica la procedencia de las lecturas ora tejido adyacente (azul), ora tumor (marrón). La segunda anotación vertical (gris claro/oscuro) representa los grupos de k-means a posteriori, o epialelos. Gracias a esta información, podemos determinar si cada uno de los elementos germinales es metil-homocigoto o metil-heterocigoto (es decir, si está metilado o no cada uno de los epialelos). Por otro lado, la columna roja señala las lecturas que podrían ser contaminación. Para determinar si un epi-alelo de un elemento (en este caso, su promotor) está metilado o no, se calculó la media de todas las CpG en cada grupo. Si el valor resultante era inferior a 0.3, se consideró desmetilado; si era superior a 0.7, se consideró metilado. Se descartaron aquellos elementos con epialelos con valores de metilación comprendidos entre 0.3 y 0.7. El primer ejemplo (Figura 133) se corresponde con el LINE1 en la posición 155671310 del cromosoma 12 para las muestras PD0270 (izquierda) y PD0277 (derecha). En ambas muestras tal elemento está metilado en ambos alelos en el tejido normal adyacente (homocigoto metilado), así como homocigoto desmetilado en el tumor. La asignación al tipo celular correspondiente es muy acertada y hay una escasa contaminación de lecturas “normales” en el tumor.

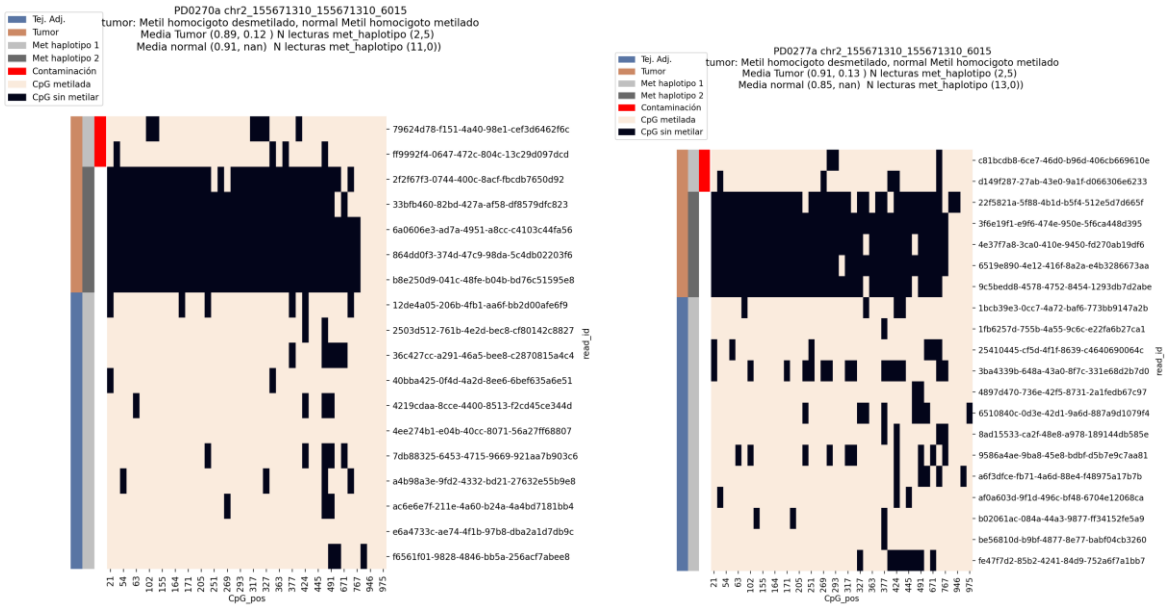


Figura 133: Metil-alelos (epialelos) del LINE1 chr2 155671310 para las muestras PD0270 y PD0277.

Por su parte, en la figura 134 se muestran dos mapas de calor del estado de metilación para el LINE1 del chr6, posición 29952435, para las muestras PD0266 y PD0331. En la muestra PD0266 este elemento es metil heterocigoto en el tumor y metil homocigoto desmetilado en el tejido normal, mientras que en la muestra PD0331 el elemento es metil homocigoto metilado en el tumor y metil heterocigoto en el tejido adyacente.

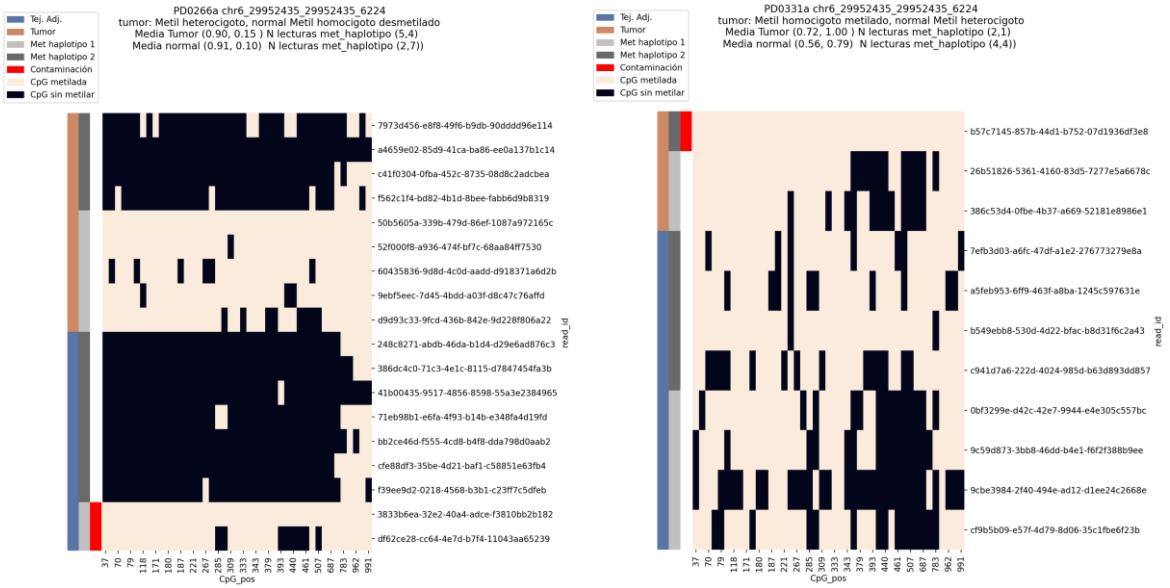


Figura 134: Metil-alelos (epialelos) del LINE1chr6 29952435 para las muestras PD0266 y PD0331.

Al considerar la información disponible de todos los elementos de todas las muestras de las cuales se dispone de tejido adyacente (en un total de 10 pacientes, excluyendo las líneas celulares), se obtiene una visión global del estado de metilación de los elementos germinales tanto en el tejido adyacente como en el tumor (Tabla 32). En el tejido adyacente predomina una mayoría de elementos altamente metilados, lo cual contrasta significativamente con la situación en el tumor, donde se produce una pérdida de metilación en más de la mitad de los elementos. Aunque existen ejemplos de epialelos en heterocigosis en adyacente y tejido tumoral, el estado mayoritario es de un único epialelo, en ambos casos.

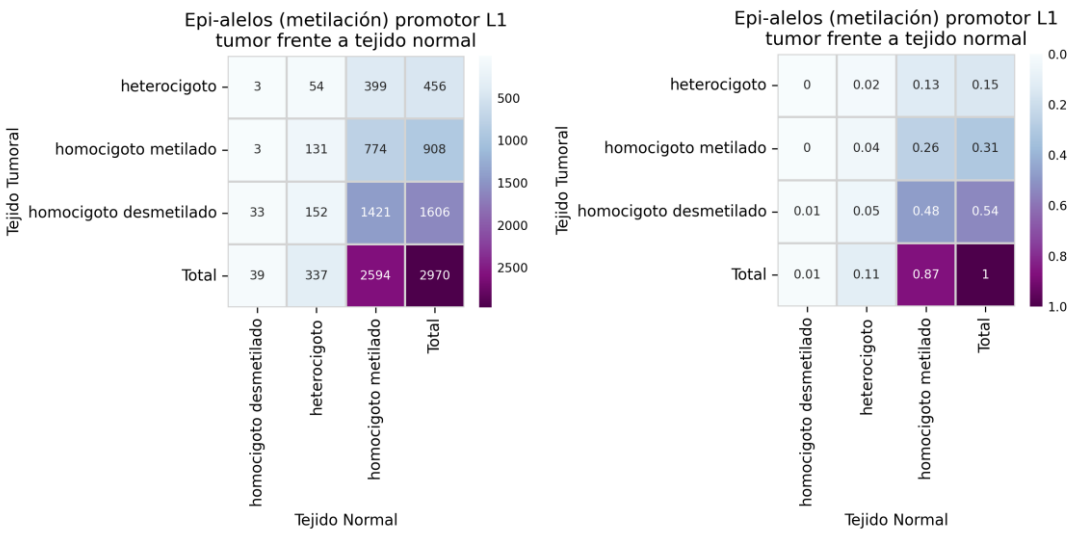


Tabla 32: Comparativa de los epialelos, entre el tejido sano y los tumores (en número de elementos y porcentaje).

Las imágenes restantes de todas las muestras y elementos, junto con el *script* utilizado para su generación, se encuentran disponibles en el repositorio adjunto en el siguiente enlace: **<.../notebooks/01_L1_germinal/20221205_Pregunta1_L1_germinales_02.ipynb>.**

8.2.3.1 Asignación de metil-alelos (epialelos) en líneas celulares.

Durante los análisis adicionales realizados en las dos parejas de líneas celulares tumorales NCI-H2009 y NCI-H2087 y sus correspondientes linfocitos NCI-BL2009 y NCI-BL2087, se observaron patrones de metilación altamente atípicos en los elementos LINE1. En concreto, no aparecen bloques o grupos de lecturas casi completamente metilados o desmetilados como en los tumores primarios analizados en este trabajo, sino que las CpG individuales de cada lectura parecen seguir un patrón aleatorio (figura 135). Estos hallazgos corroboran las observaciones previas descritas en el apartado 8.1.1.1 (metilación de las líneas celulares), donde se evidenció una inversión en los patrones de metilación genómica, con una alta metilación en las líneas tumorales y una baja metilación en las líneas derivadas de linfocitos inmortalizados. Este resultado respalda mi decisión de no incluir estas líneas celulares en la elaboración de las estadísticas en el apartado anterior.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

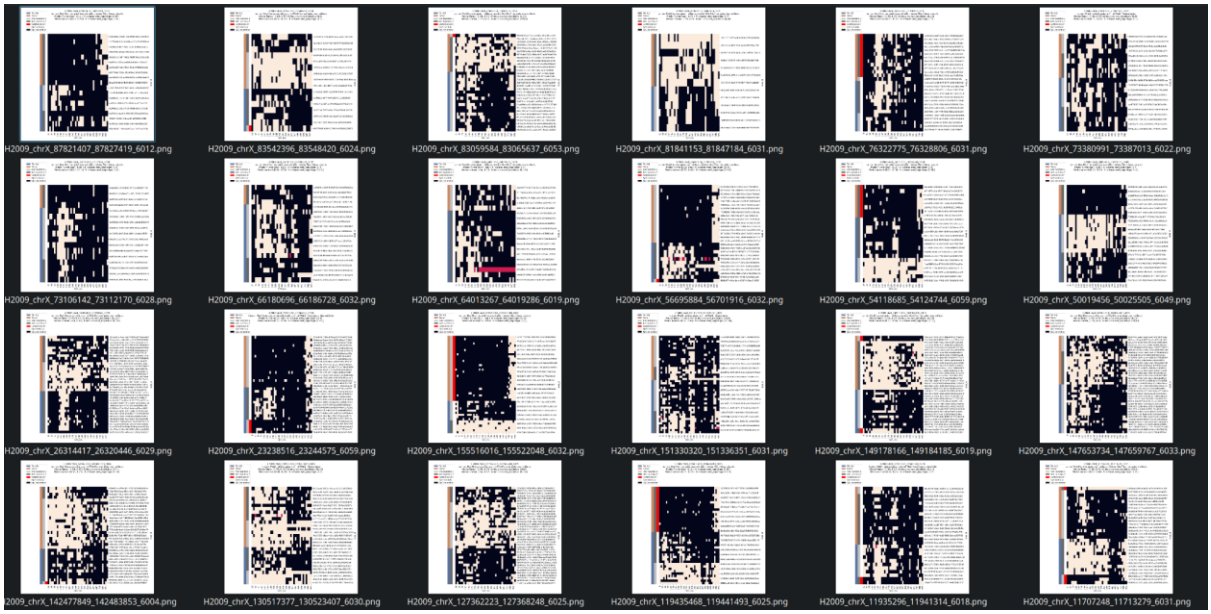


Figura 135: Ejemplo de los epialelos de una de las líneas celulares, mostrando el patrón caótico de metilación.

En las líneas derivades de linfocitos, se observa una proporción significativa de elementos no metilados, mucho mayor que en los tumores primarios. De hecho, solo un 60% de los elementos presentes ambas líneas normales son homocigotos metilados, en comparación con casi un 90% en el tejido adyacente de los tumores primarios (Tabla 33). En contraste, los epialelos heterocigotos predominan en ambas líneas tumorales, en las que representan el 46% de los elementos, seguidas de epialelos homocigotos metilados y desmetilados en proporciones similares (Tabla 33).

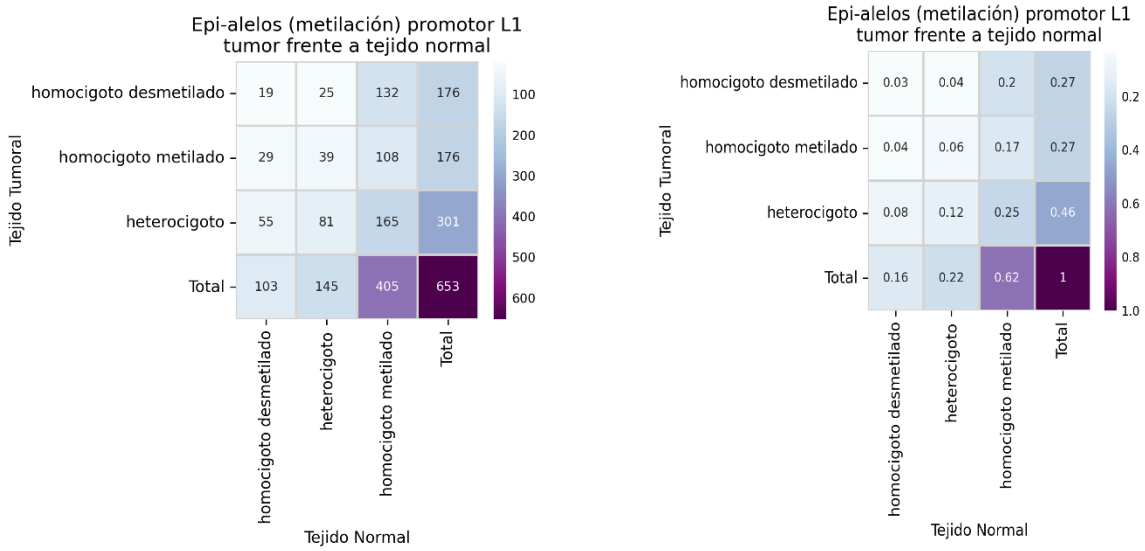


Tabla 33: Comparativa de los epialelos (líneas celulares), entre el tejido sano y los tumores (en número de elementos y porcentaje).

Frente a ello, en los tumores primarios, los epialelos homocigotos no metilados constituyen la mayoría con un 54%, seguidos por un 30% de homocigotos metilados y un 15% de heterocigotos (Tabla 32). Estos resultados evidencian profundísimas diferencias en los perfiles de metilación entre ambas líneas celulares tumorales y los tumores primarios,

destacando la importancia de considerar estas variaciones al analizar y comprender el estado de metilación en diferentes muestras de distintas procedencias y poniendo en tela de juicio la validez de, al menos, este modelo biológico para estudiar este tipo de modificaciones epigenéticas.

8.2.4 Metilación de los LINE1 germinales íntegros por regiones.

Los datos de la media y medianas de metilación de los LINE1 germinales íntegros por regiones se muestran a continuación (tabla 34 y figuras 136 y 137). El primer dato que llama la atención es la disminución de la metilación del tumor con respecto al tejido normal en las tres regiones (regiones adyacentes, cuerpo y promotor del LINE1). En los tejidos normales, las tres regiones presentan unos valores de metilación altos muy similares, siendo la de los de los cuerpos del LINE1 ligeramente mayores. En cambio en los tumores las diferencias son muy abultadas, principalmente con respecto a la región de los promotores del LINE1. En la comparación de la metilación de los tejidos normales con respecto a los tumores vemos que la mediana de los promotores de los LINE1 disminuye de 0.827 (tej. normal) a 0.357 (tumor), es decir un 56.78% de diferencia en comparación a un 12.88% de diferencia en las regiones adyacentes o un 6.7% de diferencia en los cuerpos del LINE1.

Región	Es tumor	Media de Metilación	Diferencia en la media N - T	Mediana de metilación	Diferencia en la mediana N - T	Es activo	Media de metilación	Mediana de metilación	Número de elementos
Adyacente (20K pb corriente arriba /abajo)	NO	0.757	0.132	0.769	0.099	NO	0.756	0.766	2188
						SI	0.759	0.772	1218
	SI	0.625		0.670		NO	0.610	0.658	2176
						SI	0.640	0.683	1205
Cuerpo LINE1	NO	0.852	0.110	0.857	0.059	NO	0.845	0.850	2174
						SI	0.859	0.865	1218
	SI	0.742		0.799		NO	0.719	0.782	2161
						SI	0.765	0.816	1204
Promotor LINE1	NO	0.786	0.374	0.827	0.470	NO	0.791	0.830	2184
						SI	0.782	0.825	1218
	SI	0.413		0.357		NO	0.433	0.374	2173
						SI	0.393	0.341	1204

Tabla 34: Estadísticas de la metilación de LINE1 germinales (tumor y tejido adyacente), en regiones adyacentes, promotor y cuerpo de LINE1

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

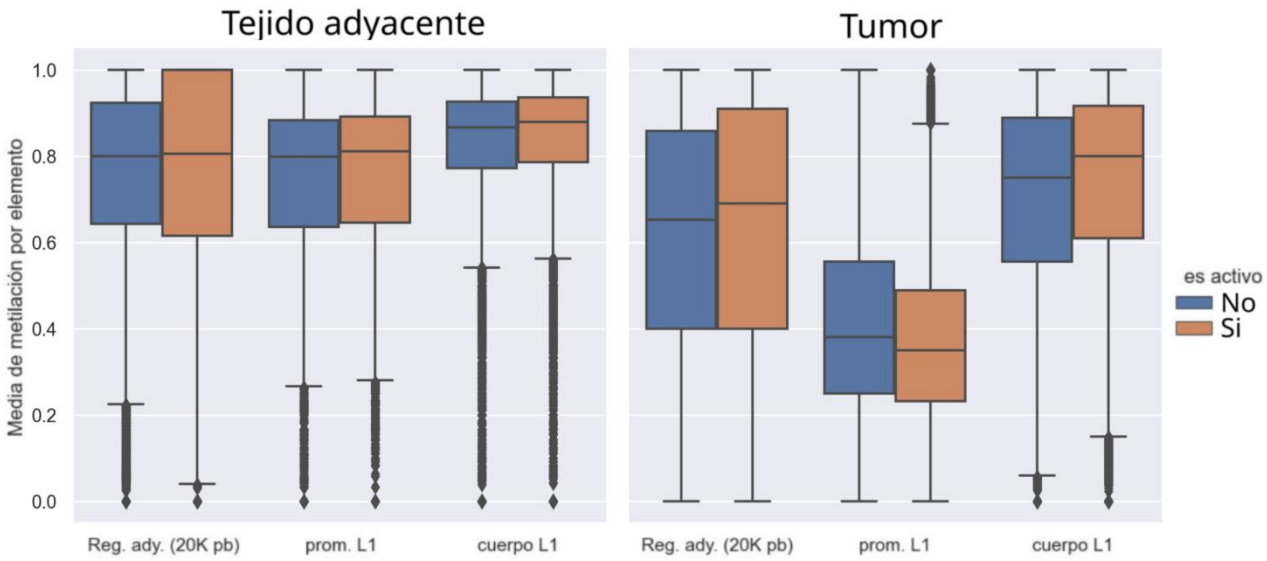


Figura 136: Comparativa tumor - normal (gráfica de bigotes) de la metilación de LINE1 germinales y regiones adyacentes.

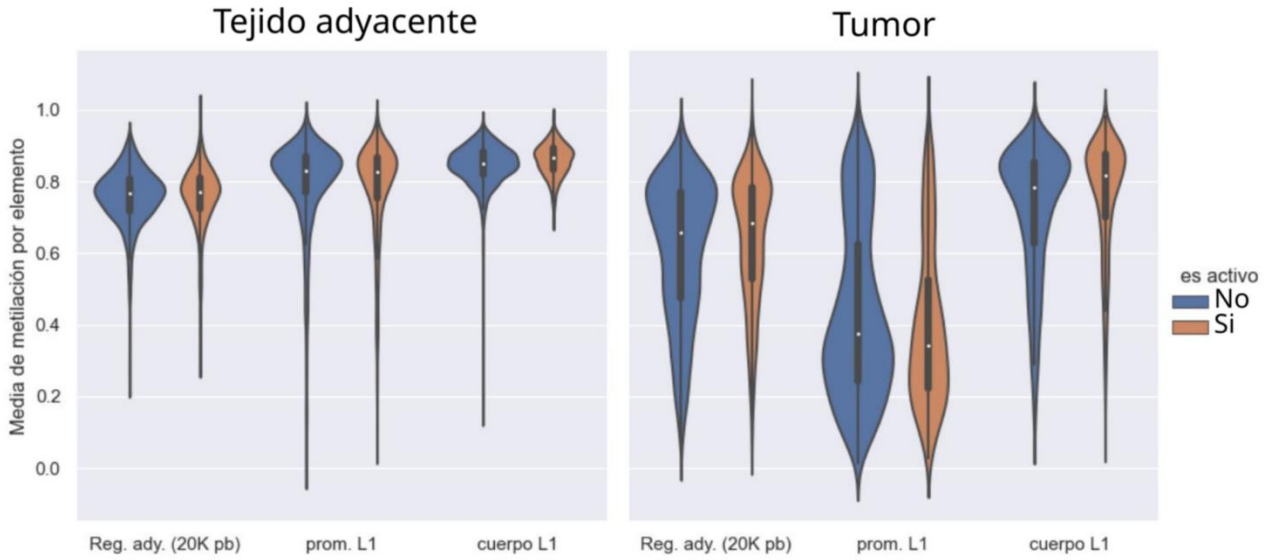


Figura 137: Comparativa tumor - normal (gráfica de violín) de la metilación de LINE1 germinales y regiones adyacentes.

Nota: Se ha desagregado la información en elementos activos y no activos, para inspeccionar visualmente la existencia de algún patrón al agrupar los LINE1 por esa categoría. Como se puede comprobar las diferencias son mínimas. Se dedica una sección entera al análisis de la actividad de los LINE1 germinales (ver apartado 8.3)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8.2.5 Análisis individual de los LINE1 germinales íntegros

Como complemento al análisis anterior, decidí investigar el estado concreto de cada uno de los LINE1 germinales, íntegros. En la figuras 138, 139, 140 se presentan los niveles de metilación de los promotores, cuerpos y regiones adyacentes (respectivamente) de los 269 elementos fuentes presentes en todas las muestras tumorales y los tejidos adyacentes analizados en este estudio. Se observa claramente que la mayoría de los promotores de LINE1 (figura 138) están hipometilados en las muestras tumorales y hipermetilados en los tejidos adyacentes. Estos perfiles de metilación son suficientes para distinguir entre los tejidos adyacentes y los tumores mediante una agrupación de muestras basada en las distancias euclídeas calculadas a partir de los patrones de metilación. Además, estas diferencias parecen ser lo suficientemente distintivas como para identificar el tejido de origen dentro de los tejidos adyacentes, aunque debido al pequeño número de muestras, no podemos afirmar con certeza esta observación. Es importante destacar que los tejidos adyacentes presentan, en su promotor, un nivel de metilación mucho mayor que sus tumores correspondientes, como se puede apreciar en la anotación vertical de densidad de metilación en cada muestra. Sin embargo, hay una excepción: la muestra de tumor colorrectal PD0351 se agrupa más cercanamente con los tejidos adyacentes que con los tumores. Esto podría indicar la presencia de un fenotipo globalmente hipermetilado característico de algunos de estos tipos de tumores, conocido como CIMP (fenotipo de metilación de las islas CpG) (Mojarad et al. 2013). También se realizó una agrupación de los elementos (columnas) basada en sus niveles de metilación en las muestras correspondientes, pero no se observó una agrupación aparente en relación con su nivel de actividad, entendiendo la actividad como el número de movilizaciones somáticas atribuibles a cada elemento en el conjunto de muestras analizadas. Como ya se ha comentado anteriormente, con este tipo de gráficas se puede comprobar como los patrones de metilación de los cuerpos de los LINE1 (figura 139) en los tumores, no presentan un patrón tan diferente al del tejido normal como los promotores (figura 138). De hecho, su patrón de metilación se asemeja más al de las regiones adyacentes (figura 140).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

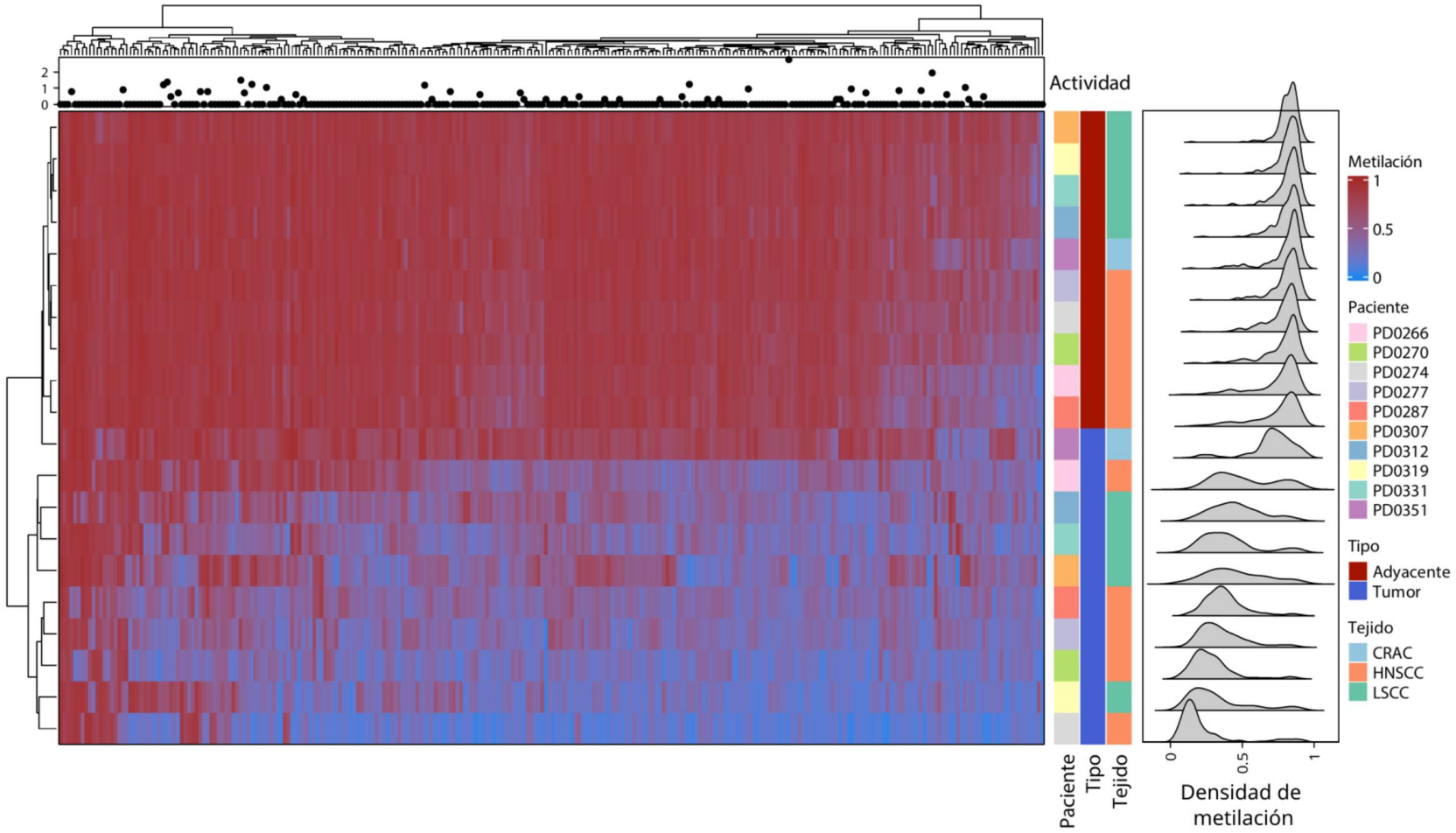


Figura 138: Mapa de calor (heatmap) de la metilación de los promotores de LINE1 germinales íntegros.
Metilación promedio en los promotores las CpG de los 269 loci monomórficos presentes en las muestras tumorales y de tejido adyacente analizadas en este estudio. Se muestra el tipo tumoral (CRAC: Cáncer colorrectal, HNSCC: Carcinoma escamoso de cabeza y cuello, LSCC: Carcinoma escamoso de pulmón) y la actividad en términos de movilizaciones en la cohorte. La agrupación de las muestras (filas) y los elementos fuente (columnas) se realizó utilizando distancias euclídeas basadas en los patrones de metilación. El panel superior muestra el logaritmo en base 10 de la actividad total cada uno de los LINE1en el conjunto de las diez muestras tumorales.

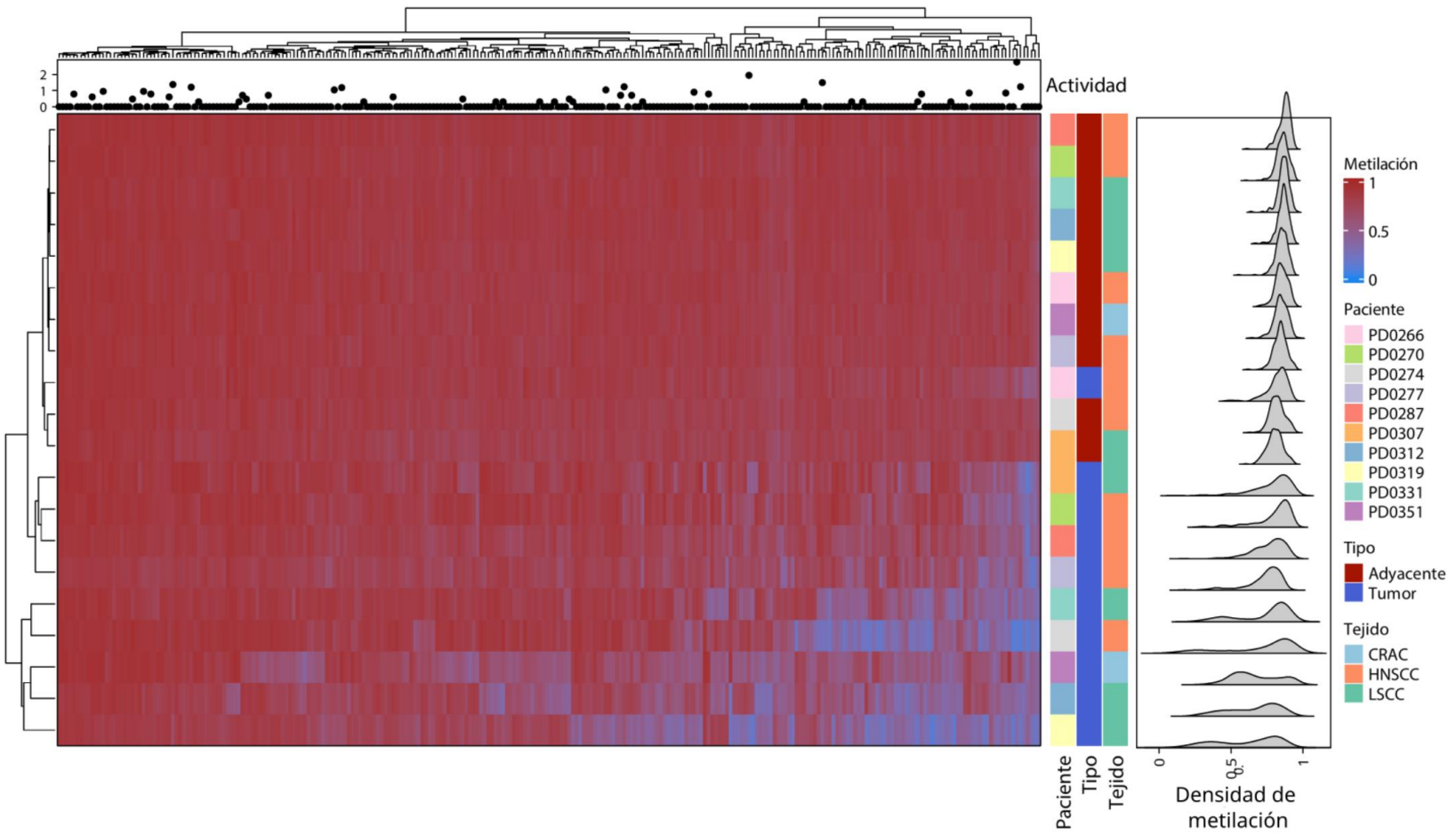


Figura 139: Mapa de calor (heatmap) de la metilación de los cuerpos de LINE1 germinales íntegros.
Metilación promedio en los promotores las CpG de los 269 loci monomórficos presentes en las muestras tumorales y de tejido adyacente analizadas en este estudio. Se muestra el tipo tumoral (CRAC: Cáncer colorrectal, HNSCC: Carcinoma escamoso de cabeza y cuello, LSCC: Carcinoma escamoso de pulmón) y la actividad en términos de movilizaciones en la cohorte. La agrupación de las muestras (filas) y los elementos fuente (columnas) se realizó utilizando distancias euclídeas basadas en los patrones de metilación. El panel superior muestra el logaritmo en base 10 de la actividad total cada uno de los LINE1en el conjunto de las diez muestras tumorales.

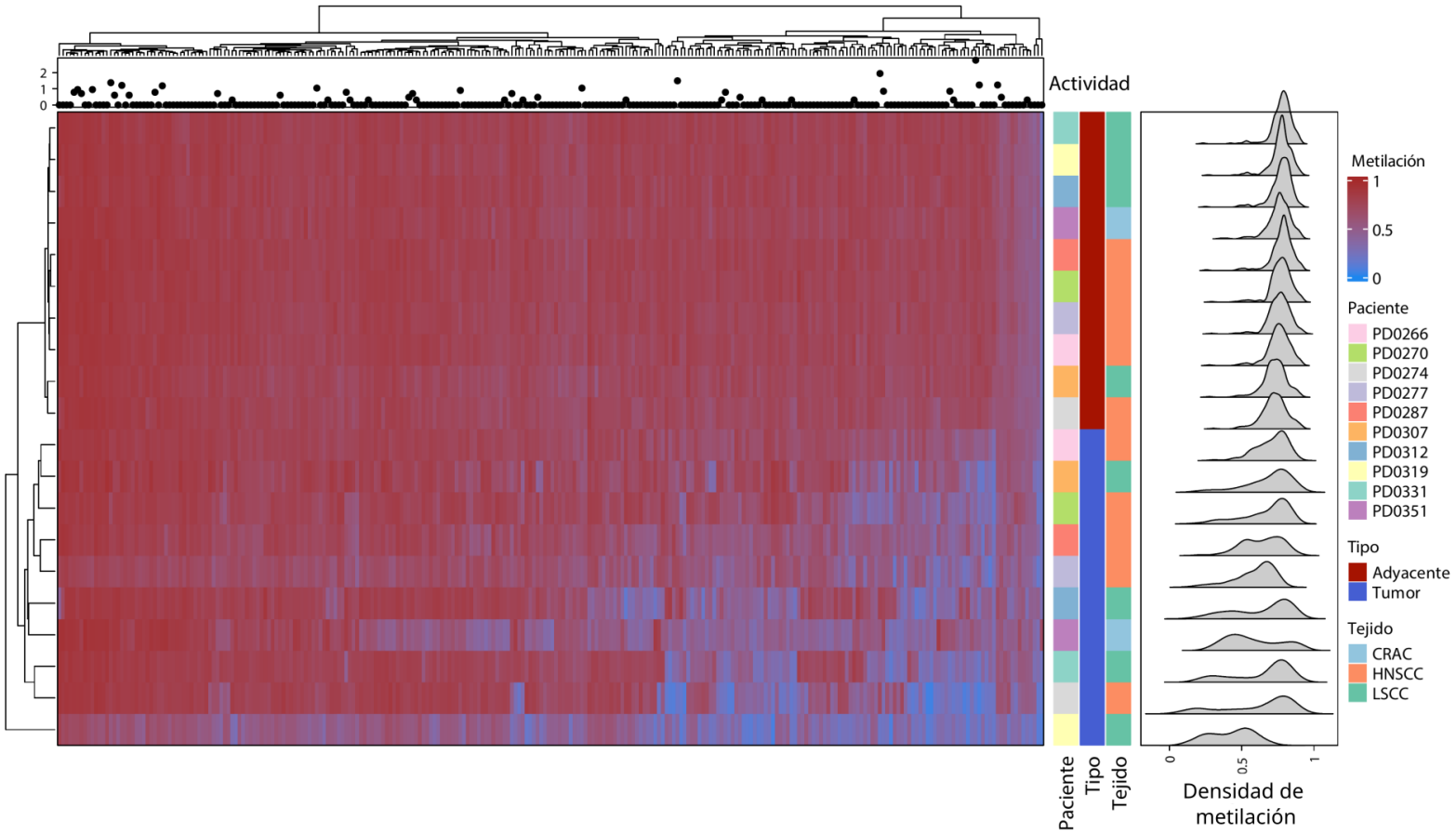


Figura 140: Mapa de calor (heatmap) de la metilación de las regiones adyacentes a los LINE1 germinales íntegros.
Metilación promedio en los promotores las CpG de los 269 loci monomórficos presentes en las muestras tumorales y de tejido adyacente analizadas en este estudio. Se muestra el tipo tumoral (CRAC: Cáncer colorrectal, HNSCC: Carcinoma escamoso de cabeza y cuello, LSCC: Carcinoma escamoso de pulmón) y la actividad en términos de movilizaciones en la cohorte. La agrupación de las muestras (filas) y los elementos fuente (columnas) se realizó utilizando distancias euclídeas basadas en los patrones de metilación. El panel superior muestra el logaritmo en base 10 de la actividad total cada uno de los LINE1en el conjunto de las diez muestras tumorales.

Las figuras 141 y 142 constituyen una representación gráfica adicional del estado de metilación de las CpGs a lo largo de la secuencia de todos los LINE1 en las muestras de tejido adyacentes y tumorales analizadas, respectivamente. En estas figuras, el eje X muestra la posición en la secuencia del LINE1, mientras que el eje Y representa la media suavizada de metilación en los dinucleótidos CpG en ventanas de 25 pb. El área correspondiente al promotor de los LINE1 se sombrea en verde y se resalta entre líneas verticales punteadas. Cada curva de color azul representa el nivel de metilación a lo largo de la secuencia para cada elemento fuente y muestra analizados en este estudio, y se ha suavizado mediante la media por ventana de 25 pares de bases consecutivas. Además, se muestra una curva roja que representa la media suavizada de la metilación a lo largo de la secuencia consenso de LINE1, y se destacan los intervalos de confianza al 95% en torno a la misma en tonos de rojo más claro. Estas representaciones gráficas proporcionan una visualización detallada de los patrones de metilación en los LINE1, permitiendo identificar las regiones específicas de la secuencia con niveles de metilación distintos entre tumores y tejido adyacente. Así, se hace patente de visu que (1) tanto en los tejidos adyacentes como en los tumores el nivel de metilación es menor en el promotor que en el cuerpo del LINE1, (2) que el nivel de metilación es, en general, menor en los tejidos tumorales que en sus tejidos adyacentes y que (3) tal diferencia es incluso más acusade en la región correspondiente al promotor. Del mismo modo, la curva promedio suavizada y los intervalos de confianza en torno a ella brindan una referencia para comparar los perfiles de metilación de los elementos individuales.

En los *notebooks*

<../notebooks/1_L1_germinal/20230517_Pregunta1_LINE1_germinales_graficas_pile_up_NORMAL_01.ipynb>

<../notebooks/1_L1_germinal/20230517_Pregunta1_LINE1_germinales_graficas_pile_up_TUMOR_01.ipynb>

se recogen los *scripts* de Python necesarios para recrear estas figuras.

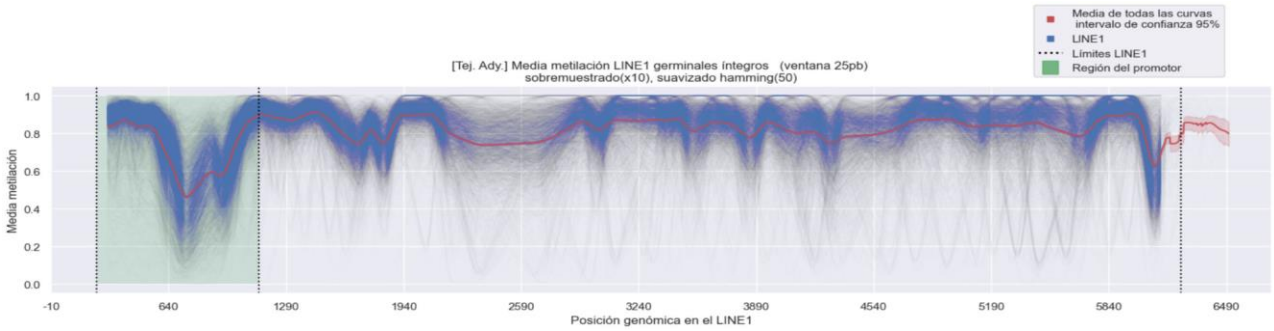


Figura 141: Gráfica suavizada de la metilación de todos los LINE1 germinales en los tejidos adyacentes

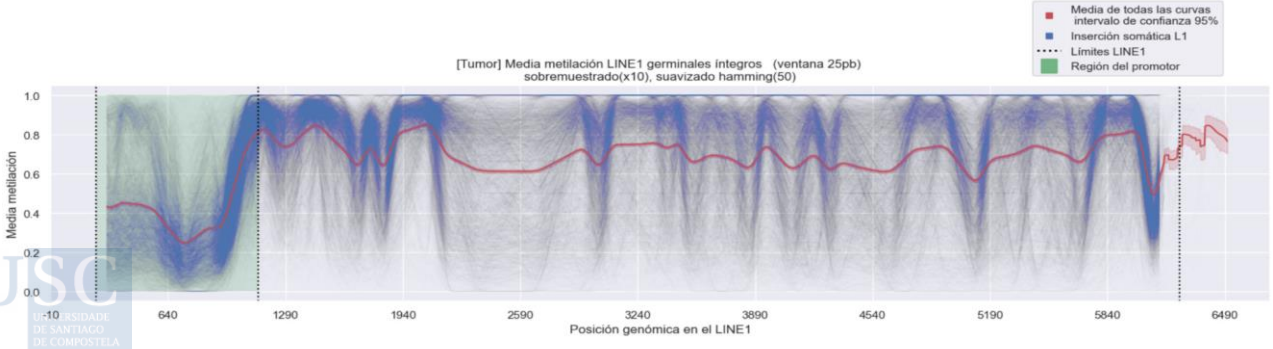


Figura 142: Gráfica suavizada de la metilación de todos los LINE1 germinales en las muestras tumorales

Nota 1: El análisis general de este apartado se realizó mediante el "notebook" de Python
<../notebooks/2_correlacion_actividad_metilacion/20230318_LINE1_activos_correlacion_metilacion_05.ipynb>

Nota 2: He realizado un análisis similar estudiando la hidroximetilación de los LINE1 en la misma cohorte. Estos muestran en toda su longitud unos valores muy bajos de hidroximetilación, generalmente menores de 0.2 y con una media de 0.01. Los patrones aleatorios que presentan no parecen aportar ninguna información, motivo por el cual se han omitido aquí su descripción pormenorizada y sus imágenes correspondientes, no obstante se pueden visualizar en el *notebook* del repositorio adjunto.

<../notebooks/2_correlacion_actividad_metilacion/20230318_LINE1_activos_correlacion_hmetilacion_05.ipynb>

8.2.6 Metilación diferencial entre el promotor y el cuerpo de los LINE1 germinales íntegros.

En general, los patrones de metilación en el cuerpo de los elementos LINE1 muestran similitudes con las regiones adyacentes, tanto corriente arriba como corriente abajo del propio elemento LINE1. En contraste, los promotores de estos elementos presentan niveles de metilación considerablemente más bajos en comparación con el cuerpo y las regiones adyacentes, lo que sugiere una regulación diferencial de la metilación en estas áreas genómicas. Estos patrones son particularmente marcados en las muestras tumorales analizadas, ya que no se observan de manera generalizada en los tejidos adyacentes al tumor. Estos resultados se hacen muy aparentes en la tabla 34, en las figuras 136 y 137 (gráficas de bigotes y de violín), 141 y 142 (gráficas suavizadas a lo largo del LINE1) y las figuras de mapa de calor 138 (promotor), 139 (cuerpo) y 140 (regiones adyacentes). Los resultados son consistentes en la mayoría de las muestras, excepto en la muestra de adenocarcinoma colorrectal PD351 donde el patrón de metilación del promotor en el tumor es muy similar al del tejido normal y es compatible con un fenotipo tumoral hipermetilado, como ya se mencionó anteriormente.

Estas observaciones se corroboraron mediante la realización de una serie de correlaciones de Pearson, tanto para los tumores como para los tejidos adyacentes. Se comparó a) la metilación media del promotor de los LINE1 y la metilación media de su cuerpo junto con la regiones adyacentes (figura 143 a) y b) la media de metilación del cuerpo y la media de metilación de la región adyacente (figura 143 b) (para cada LINE1 germinal íntegro de la cohorte de muestras). Los resultados son recogidos en la tabla 35 y muestran como correlaciones significativas. Entre el promotor del LINE1 por un lado y el cuerpo y regiones adyacentes por el otro la correlación es muy baja o inexistente, siendo ligeramente mayor en los tumores que en el tejido normal (r = 0.181 y r = 0.126 respectivamente). En cambio la correlación entre el cuerpo del LINE1 y las regiones adyacentes es muy fuerte para los tumores r = 0.830 y moderada para los tejidos adyacentes.

	Tumor			Normal		
	y	r	p-value	y	r	p-value
a	0.136x + 0.612	0.181	2.299 e-26	0.057x + 0.752	0.126	1.561 e-13
b	0.917x - 0.054	0.830	0.0	0.640x + 0.210	0.398	1.487 e-129

Tabla 35: Datos de correlación entre regiones de los LINE1 germinales (promotor, cuerpo, regiones adyacentes)

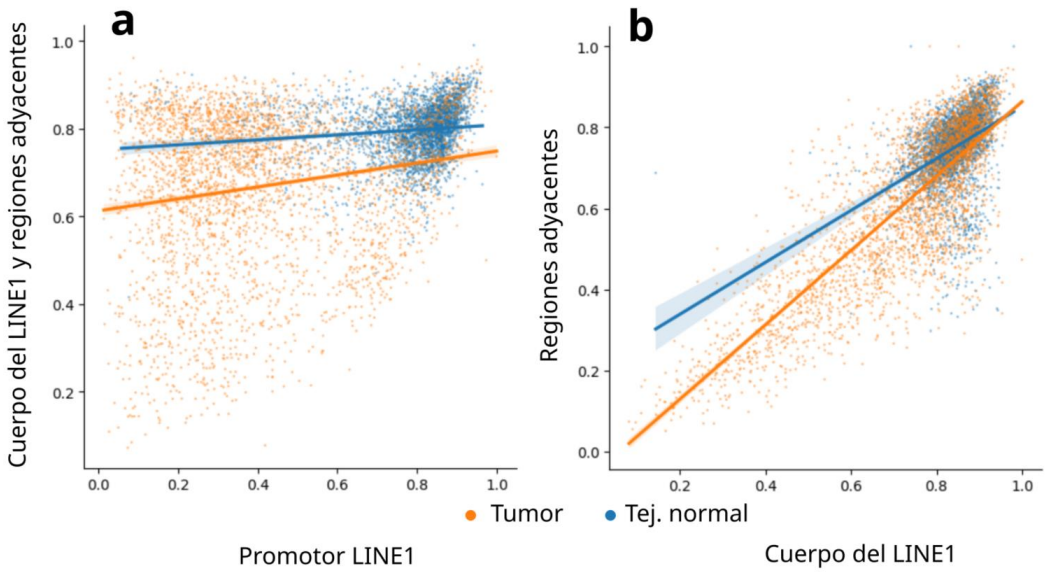


Figura 143: Correlación entre regiones de los LINE1 germinales (promotor, cuerpo, regiones adyacentes)

Para respaldar este hallazgo, se llevó a cabo un análisis de regresión lineal para examinar posibles relaciones entre los niveles de metilación en el promotor de los elementos LINE1 (prom. LINE1) y los niveles de metilación en su cuerpo (*cuerpo LINE1*), así como en las regiones genómicas adyacentes (*Reg. ady. (20K pb)*) en las muestras tumorales. Los resultados (tabla 36) de la regresión lineal revelaron asociaciones estadísticamente significativas entre estas variables. En particular, se observa una correlación inversa entre la metilación del promotor y la metilación en el cuerpo de los elementos LINE1 (-0.31605, $p < 0.001$) así como una correlación positiva entre la metilación del promotor y la metilación de las áreas genómicas adyacentes al mismo (0.51397, $p < 0.001$).

Curiosamente, al aplicar un modelo de regresión lineal sobre los datos de tejidos adyacentes a estos mismos tumores, encontramos tendencias muy similares (tabla 36), en las que, nuevamente, el estado de metilación de los promotores de LINE1 posee una correlación inversa con los de su cuerpo (-0.16852, $p < 0.001$) y positiva con las regiones adyacentes al mismo (0.31511, $p < 0.001$). En ambos casos, las variables predictoras muestran asociaciones consistentes con las observadas en las muestras tumorales.

	Modelo Tumoral			**Modelo No Tumoral**		
	Coeficiente	*p value*		*Coeficiente*	*p value*	
(Intercept)	0.33149	< 2e-16	***	0.69275	< 2e-16	***
Cuerpo LINE1	-0.31605	< 2e-16	***	-0.16852	0.000757	***
Reg. ady. (20K pb)	0.51397	< 2e-16	***	0.31511	< 2e-16	***
Adjusted R-squared	0.06			0.03		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Tabla 36: Regresión lineal entre los promotores y el cuerpo y regiones adyacentes de los LINE1 germinales.

Es importante destacar que el modelo basado en la cohorte tumoral muestra un R-cuadrado ajustado de 0.06, lo que sugiere que las variables predictoras explican el 6% de la variabilidad en los niveles de metilación en el promotor de los elementos LINE1 en estas

muestras. Por otro lado, el modelo equivalente en las muestras no tumorales tiene un R-cuadrado ajustado de 0.03, lo que indica que las mismas variables predictoras tienen un impacto menos pronunciado en la explicación de la variabilidad en los niveles de metilación en el promotor de los elementos LINE1 en tejidos no tumorales. En resumen, estos resultados sugieren que las relaciones entre los niveles de metilación en los elementos LINE1 y sus regiones adyacentes son marginalmente más influyentes en las muestras tumorales en comparación con las no tumorales. Es menester indicar que, pese a que esta aseveración es muy gruesa, ya que con total seguridad los valores de metilación serán dependientes de muestra, tipo de tejido y elemento, este resultado sugiere la presencia de mecanismos que favorecen una pérdida de metilación en los promotores de LINE1 ya en las muestras de tejido adyacente y que éstas se ven augmentadas en las muestras tumorales. Estos resultados son, no obstante, dudosos, ya que es posible se trate nuevamente de un efecto de infiltraciones del tejido adyacente "normal" sano en las muestras tumorales, o incluso del tejido tumoral en el adyacente, que podrían desvirtuar los valores reales. Sería necesario realizar un estudio más pormenorizado de estas muestras de tejido adyacente, desarrollando pipelines que permitan la discriminación de lecturas en haplotipos normales y tumorales para, posteriormente, extrapolar de ellas estadísticas "reales" de ambos tipos celulares.

En todo caso y hasta la aparición de la tecnología de lecturas largas, la comprensión de la metilación en el interior de los elementos LINE1, abarcando tanto el promotor como el cuerpo de estos elementos, era en gran medida un enigma. Los enfoques utilizados en investigaciones previas presentaban diversas limitaciones. Por un lado, los métodos basados en corte enzimático o en arrays de metilación carecían de la capacidad para ofrecer detalles sobre retrotransposones específicos y el estado de metilación a nivel de nucleótido. Por otro lado, los métodos que involucraban pretratamiento con bisulfito y secuenciación NGS se encontraban limitados debido al tamaño reducido de las lecturas, lo que dificultaba la obtención de información precisa sobre la metilación en el interior de los retrotransposones debido a problemas de mapeo.

La introducción de la tecnología de lecturas largas ha supuesto un cambio significativo en este panorama. Ahora es posible analizar el estado de metilación de elementos LINE1 concretos, y en realidad, la mayoría, si no todos, de estos elementos, a nivel de nucleótido. En el contexto de este estudio, esta tecnología ha permitido una evaluación minuciosa de la metilación de todos los dinucleótidos CpG presentes en los LINE1 germinales íntegros. Este enfoque ha brindado la oportunidad de visualizar y comprender con mayor detalle el patrón de metilación que se ha mencionado previamente, ofreciendo una perspectiva nueva y esclarecedora sobre la metilación de los LINE1 en muestras tumorales.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8.3 METILACIÓN Y ACTIVIDAD DE LOS LINE1 GERMINALES

8.3.1 Análisis general

La relación entre la actividad de los LINE1 en los tumores y su estado de metilación es un tema recurrente en el campo de la investigación oncológica, así como uno de los objetivos inicialmente planteados en esta tesis doctoral como parte del proyecto de investigación sobre retrotransposición y cáncer en nuestro grupo de investigación. Para abordar esta cuestión era necesario evaluar primero los niveles de actividad de cada elemento fuente en las muestras analizadas en este estudio. En este sentido, el número de inserciones somáticas de transposones en el genoma tumoral es un buen estimador del grado de actividad, tanto de un elemento fuente como de un modo global (Tubio et al. 2014; Rodriguez-Martin et al. 2020), habiéndose desarrollado a lo largo de los años herramientas específicas para el análisis de la retrotransposición, revisadas profusamente en (H. Lee et al. 2022). Nuestro laboratorio ha desarrollado varios algoritmos y programas para tal fin como Trafic (Tubio et al. 2014), MEIGA-SR y MEIGA-LR (pendientes de publicar), siendo yo partícipe en el desarrollo de los dos últimos. No obstante, para determinar la actividad de elementos individuales es imprescindible identificar las inserciones somáticas que producen. Tal inferencia se realizó en base a dos enfoques complementarios, el primero, en base a las transducciones de las regiones corriente abajo (*downstream*) del elemento fuente, que posibilitan la identificación inmediata de la procedencia de las inserciones somáticas.

Tal como se introdujo en el apartado (3.2.3.2 Transducciones) la transducción es un proceso por el cual la movilización de un LINE1 arrastra parte de la secuencia del genoma que se encuentra adyacente al LINE1 que se está movilizando, al que denominamos elemento fuente (Hancks y Kazazian 2016; Tubio et al. 2014). Si la zona que se arrastra se ubica corriente abajo del transposón se denominan transducciones 3', y son las más frecuentes. En caso de que la región que se arrastra es la que se encuentra corriente arriba del elemento fuente se denomina transducción 5'. Estos fragmentos caracterizan de forma unívoca el elemento fuente que produce las inserciones. Estas pueden ser *partner* si se inserta el LINE1 (total o parcialmente) o huérfanas si solo se inserta la región corriente abajo o arriba del LINE1. Para el presente trabajo se usó la información generada por MEIGA-LR, aplicado a cada una de las muestras tumorales y filtrando con el tejido adyacente aquellas inserciones que no son somáticas. Esta información fue analizada gracias a la colaboración de Sonia Zumalave y al desarrollo de su programa de detección de eventos de retrotransposición de LINE1, MEIGA-LR.

El segundo enfoque emplea conjuntos de datos (*data sets*) de nucleótidos diagnóstico (SNVs) que permite asignar la procedencia de gran parte de las inserciones somáticas de LINE1 (MEI). Para su análisis es necesario un programa que determine la probabilidad asignada al elemento de origen, llamado ***Source Inference*** *(pendiente de publicación)* y desarrollado en el seno del proyecto MEIGA de nuestro grupo de investigación por Martín Santamarina. En el primer paso de este método, se realiza un genotipado completo de los elementos fuente en las muestras estudiadas. Esto implica alinear las secuencias de todos los candidatos de elementos fuente con la secuencia consenso de LINE1 utilizando minimap2 (Li 2018) y detectar las variantes polimórficas en el alineamiento. Sin embargo, en la versión actual, solo se consideran variantes puntuales (SNVs) y no se toman en cuenta otras variantes como INDELs y variantes estructurales internas. Luego, cada variante se clasifica en una de dos categorías posibles: (a) diagnóstico específico, si la variante es única para un solo

candidato de elemento fuente, o (b) diagnóstico no específico, si la variante es compartida por varios candidatos de elementos fuente. Después de esta evaluación, las secuencias de cada inserción somática (MEI) de LINE1 (solo y partner, TD0 y TD1 respectivamente) se alinean con la secuencia consenso de LINE1 para determinar sus variantes diagnósticas empleando un alineador local que implementa el algoritmo de Smith-Waterman. A través del alineamiento por pares, se realiza una anotación iterativa del grupo de inserciones somáticas LINE1, de modo que cada inserción se caracteriza por su conjunto de variantes informativas, ya sean específicas o no específicas, que se utilizarán posteriormente para asignar las inserciones a un locus LINE1 particular. El paso final consiste en la clasificación de las inserciones somáticas LINE1. En este paso, todas las inserciones se evalúan en comparación con el grupo de candidatos de elementos fuente, que están ordenados por similitud. La asignación a un locus LINE1 específico solo se realiza si se cumplen tres criterios: (1) hay una proporción del 100 % de variantes específicas compartidas entre la inserción y el candidato principal de elemento fuente, (2) ningún otro candidato comparte una proporción del 100 % de variantes específicas con la inserción, y (3) hay una proporción ≥75 % de variantes no específicas compartidas entre la inserción y el candidato principal de elemento fuente. Estimaciones preliminares de concordancia utilizando un conjunto de datos simulados (TD0) mostraron una concordancia del 99,9 % con tasas de asignación del 47,59 %. Las tasas de asignación por subfamilia L1HS fueron del 36,8 % (Ta-1), 47,1 % (Ta-0) y 59,6 % (pre-Ta). Para el conjunto de datos reales (TD1), las estimaciones de concordancia fueron superiores al 89,9 % (más del 94,1 % para TD grandes), con tasas de asignación del 27,8 % (para TD0 y TD1).

La concatenación del resultado de ambos análisis permitió corroborar el grado de activación de elementos fuente individuales en base al número de movilizaciones somáticas que han generado para cada una de las muestras objeto de estudio, tanto normales como tumorales, datos que se encuentran disponibles en el archivo **<../5_datos/*source_inference_BARPLOT_rawdata.txt*>,** incluido en el repositorio adjunto.

Utilizando la información obtenida de los datos de metilación en los cuerpos y promotores de los elementos germinales y derivativos, pudimos establecer correlaciones relevantes entre la actividad y la metilación. De esta manera, pudimos verificar si la metilación afecta de alguna manera a la capacidad de retrotransposición de estos elementos. La gráfica incluida en la figura 144 muestra la regresión lineal entre el número de inserciones somáticas de los LINE1 con elemento fuente conocido frente al estado de metilación del promotor de esos elementos fuente para las líneas celulares (NCI-H2009, NCI-H2087) y todos los tumores analizados en este estudio (23)

El resultado con un valor de r = -0.07 (p-val = 3.7 e^{-12}) indica que no hay ninguna correlación actividad-metilación en el conjunto de datos que disponemos. Se obtienen valores similares para cada una de las muestras estudiadas de forma individual (25) (disponibles en el repositorio adjunto). Los resultados también son negativos, aunque se seleccionen sólo los elementos con actividad (Número de inserciones >0, >=5, >=15, >=25) (Figura 145).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

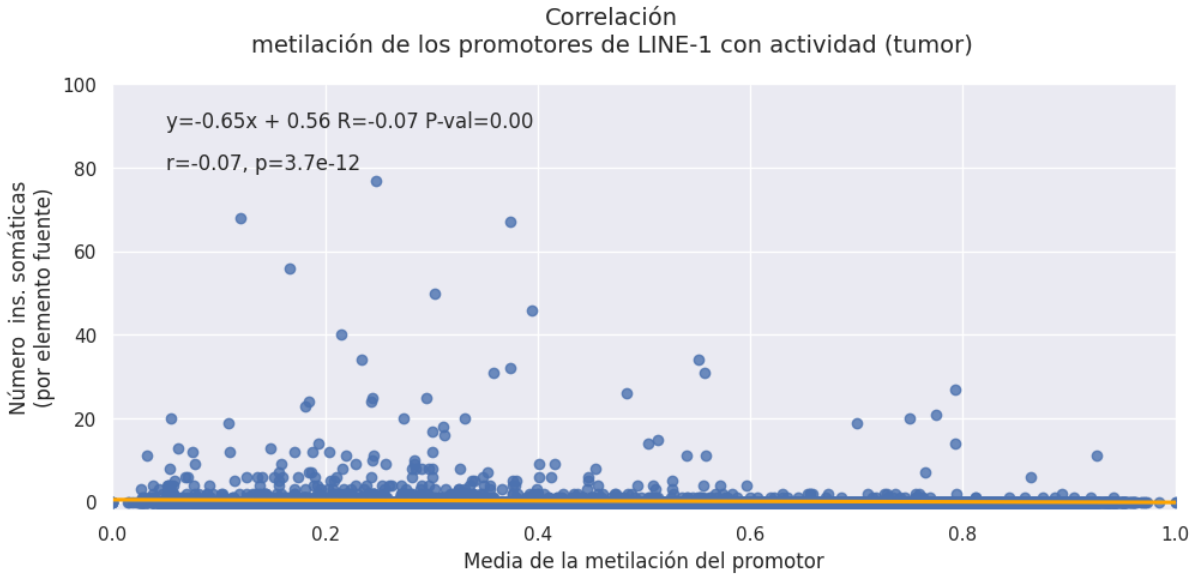


Figura 144: Correlación actividad – metilación de LINE1 germinales I
Correlación actividad LINE1 germinales íntegros con la metilación del promotor.

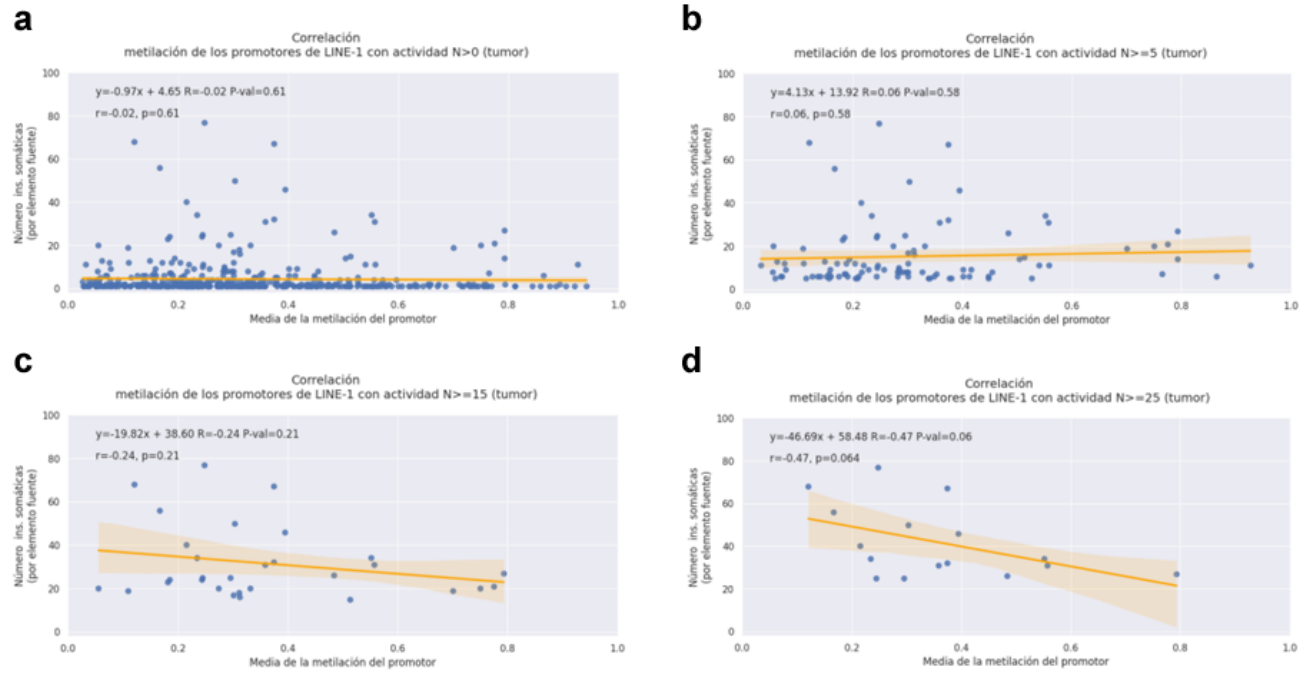


Figura 145: Correlación actividad – metilación de LINE1 germinales II.
Correlación actividad de LINE1 germinales íntegros con la metilación del promotor para elementos activos con más de una (a), 5 (b), 15 (c) y 25 (d) inserciones somáticas.

Se ha investigado si la correlación hallada en el artículo del grupo de J C Tubio en el 2014 era específica de un subconjunto de elementos, por ejemplo los que se conoce que producen gran cantidad de transducciones, como los de los estudios de Pancancer (PCAWG) (figura 146a), los de BROUHA (Brouha et al. 2003) (figura 146b). Por último, se analizó si existía alguna correlación con la edad de los elementos fuente, no detectándose ninguna para los rangos de edades analizados (figura 147).

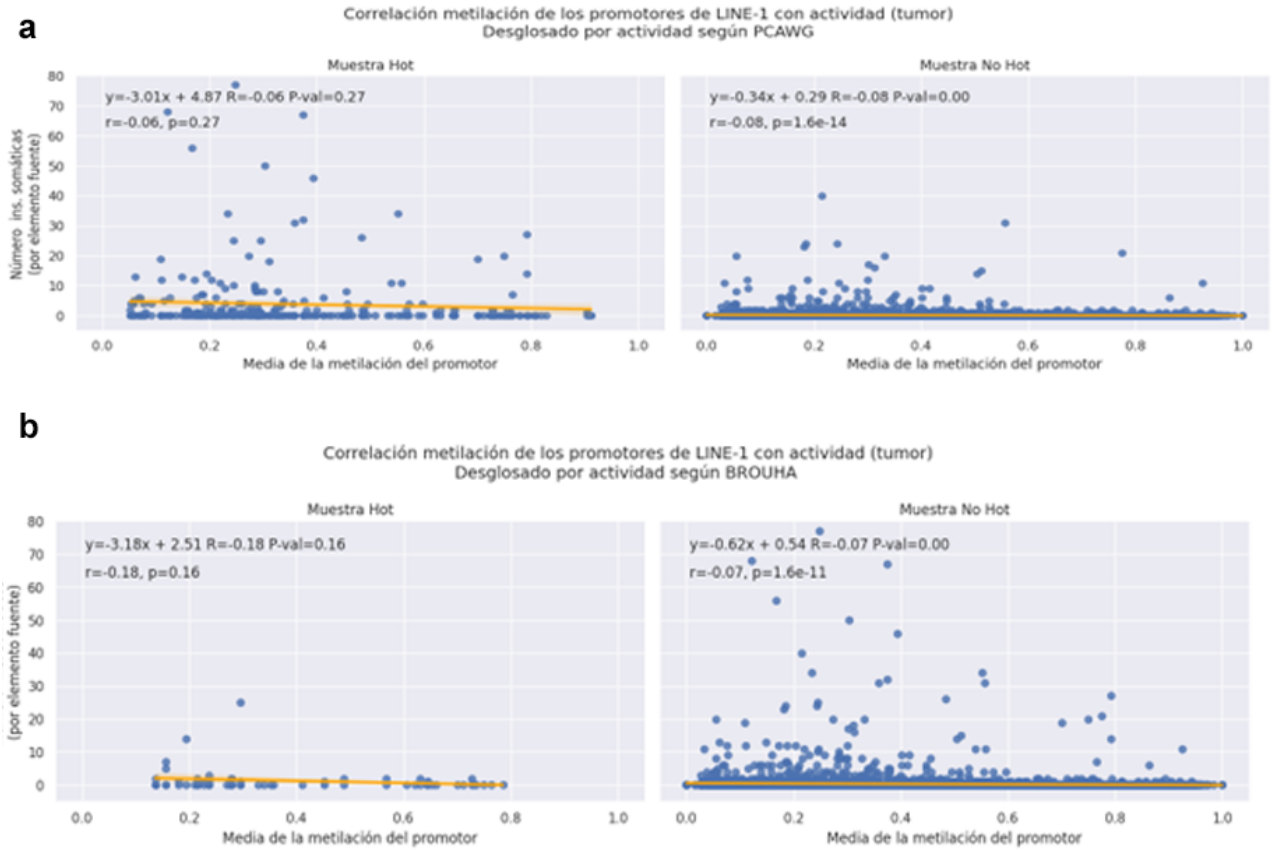


Figura 146: Correlación actividad – metilación de LINE1 germinales III.
a) Correlación actividad de LINE1 germinales, íntegros con la metilación del promotor para elementos hot (izquierda) y no hot (derecha) según Pancancer (PCAWG). b) Correlación actividad de LINE1 germinales, íntegros con la metilación del promotor para elementos hot (izquierda) y no hot (derecha) según (Brohua et al 2003).

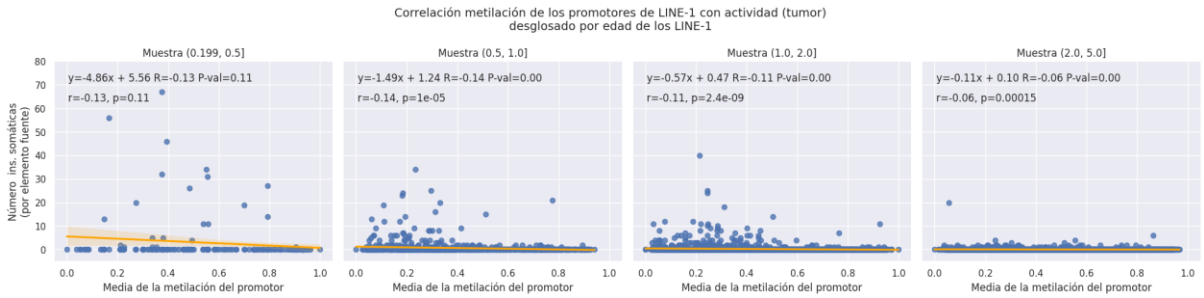


Figura 147: Correlación actividad – metilación de LINE1 germinales IV
Correlación actividad de LINE1 germinales, íntegros con la metilación del promotor para elementos desglosados por edad (rangos en millones de años)

Siguiendo el análisis realizado por Lanciano et al. (2023), se llevó a cabo un intento de verificar si existía alguna correlación al agrupar los elementos según el estado de la cromatina en la que se encuentran los elementos fuente. Para ello, se habría preferido disponer de datos de secuenciación Chip-seq, ATAC-seq o DNASE-seq de nuestra propia cohorte de datos. Sin embargo, al no contar con estos datos, se recurrió a la información pública disponible.

USC DE COMPOSTELA

Tanto el proyecto ENCODE **(https://www.encodeproject.org/)** como el UCSC Genome Browser **(https://genome.ucsc.edu/)** proporcionan información sobre el estado de la cromatina y las marcas de histonas según los tipos celulares. En el artículo "Universal annotation of the human genome through integration of over a thousand epigenomic datasets", (Vu y Ernst 2022) se aborda el análisis, agregación y asignación de esta información a las posiciones genómicas correspondientes. Para ello, se utilizó la herramienta ChromHMM

(https://github.com/ernstlab/full_stack_ChromHMM_annotations)

la cual integra datos obtenidos previamente y almacenados en bases de datos públicas (1032 conjuntos de datos). Esta agregación generó una serie de regiones de consenso que son comunes a múltiples tipos celulares. La tabla utilizada en el *script* de correlación se obtuvo de **https://public.hoffman2.idre.ucla.edu/ernst/UUKP7/hg38lift_genome_100_segments.bed.gz** (Acceso realizado el 15.04.2023) y contiene estas regiones de consenso correspondientes al genoma de referencia HG38. Esta tabla está compuesta por 100 categorías diferentes que se pueden agrupar en 16 grupos generales de estados de la cromatina cuyo resumen se muestra en la Tabla 37

https://github.com/ernstlab/full_stack_ChromHMM_annotations/blob/main/chromHMM_utilities/state_annotations.csv (acceso realizado el 15.04.2023).

En ella se describen los distintos estados epigenéticos en un compendio de distintos linajes celulares. Cada estado cuenta con un nombre abreviado y una descripción más detallada, que incluye las marcas epigenéticas asociadas, la presencia o ausencia de ciertos elementos genómicos, la expresión génica y otros rasgos. Los estados engloban diferentes tipos de quiescencia, estados de la cromatina/heterocromatina, acetilaciones y posibles potenciadores (enhancers), así como marcas de transcripción, agrupados según las siguientes categorías.

GapArtf	Abreviatura
Quesciente	Quies
Heterocromatina	HET
Acetilaciones	Acet
Promotores (enhancers)	EnhA
Potenciadores débiles	EnhWk
Potenciadores y transcritos	TxEx
Transcritos, exones	Tx
Transcripción débil	TxWk
Promotores	Tes
TSS	TSS
Artefactos, gaps, otros	GapArtf
znf	znf
DNase	DNase

Tabla 37: Abreviaturas de las regiones de la cromatina usadas

Tampoco se ha hallado una correlación significativa tal como puede comprobarse en las gráficas siguientes, donde se muestra, a modo de ejemplo, la correlación para las regiones quiescentes (Quies) (figura 148). Así mismo las correlaciones del resto de las regiones anotadas según el estado de la cromatina descritas en la tabla 37 son igualmente negativas. Todas ellas se encuentran en el Notebook adjunto:

<../notebooks/2_correlacion_actividad_metilacion/20230318_LINE1_activos_correlacion_metilacion_05.ipynb>

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

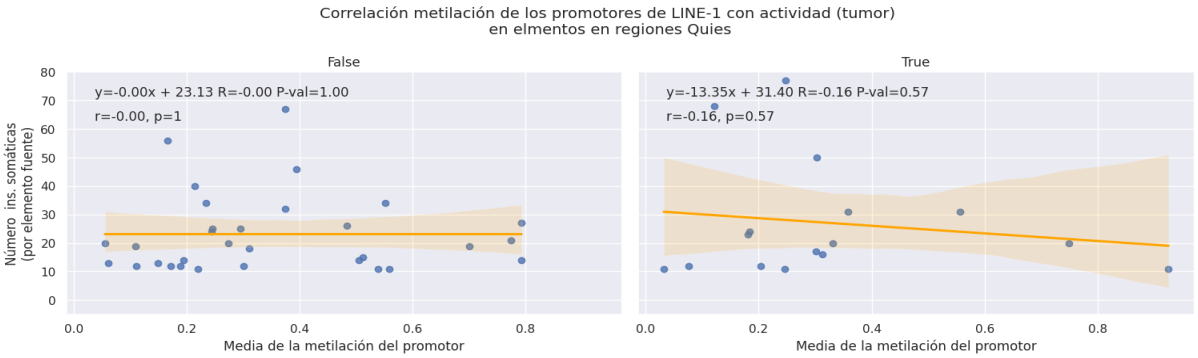


Figura 148: Correlación actividad – metilación de LINE1 germinales V

Correlación actividad de LINE1 germinales íntegros con la metilación del promotor para elementos en regiones no quiescentes de la cromatina (izquierda) y quiescentes (derecha)

8.3.2 Discusión

En respuesta al objetivo 4.1.2, el presente estudio no ha encontrado ninguna correlación entre estado de metilación y actividad en los LINE1 en la cohorte de muestras estudiada. Estos resultados están parcialmente en desacuerdo con los previamente obtenidos por (Tubio et al. 2014), en los que se encontraba que para cada uno de los elementos fuente (6) en las muestras analizadas (16), el metil-haplotipo más frecuente estaba completamente desmetilado en los LINE1 con actividad, siendo por lo tanto la correlación metilación-actividad significativa. Este desacuerdo podría estar originado porque en nuestro caso estamos teniendo en cuenta el estado de metilación de cerca de 700 LINE1 "íntegros" en vez de 6 (al ser los LINE1 polimórficos en la población de media cada muestra tiene alrededor de 300), además debido a que se tienen en cuenta todas las CpG del promotor (alrededor de 40 en vez de 6), no es posible encontrar un metil-haplotipo claramente mayoritario, por lo que se ha usado el promedio de la metilación de todas las CpG del promotor en todas las lecturas. También hay que tener en cuenta que, los tumores analizados no son puros, por lo que estamos viendo el efecto combinado de la metilación de las células tumorales y parte del tejido normal. A pesar de esto se esperaría ver una señal por encima del ruido que realmente no somos capaces de distinguir. Adicionalmente, es importante considerar que la secuenciación de las muestras se produjo en un momento puntual del desarrollo tumoral, por lo que solo tenemos una foto estática de un momento concreto en la evolución del tumor (lo que se refleja en los datos). Es sabido que la actividad de los LINE1 puede ser intermitente, sufriendo ciclos de actividad intercalados con periodos en los que no se produce retrotransposición (Rodríguez-Martín, 2020). En cualquier caso, también esperábamos que acumulando los datos de todos los LINE1 ("íntegros") y de todas las muestras apareciese una correlación que no se aprecia. Esto parece confirmar que existen otros mecanismos de control de la retrotransposición además de la metilación de los promotores. Actualmente nuestro grupo de investigación está realizando un análisis de estos LINE1 "íntegros" de forma individualizada para cada uno de los haplotipos parentales, en cada una de las muestras analizadas. En un trabajo posterior sería igualmente muy interesante realizar los análisis descritos en este apartado con esa información desglosada por haplotipo.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8.4 METILACIÓN DE LAS INSERCIONES SOMÁTICAS DE LINE1 EN EL GENOMA TUMORAL

Con el objetivo de ampliar nuestro conocimiento sobre la metilación en inserciones somáticas de LINE1 en nuestra cohorte, ejecutamos el programa Methyl-Meiga utilizando como entrada las inserciones somáticas proporcionados por MEIGA-LR, que además aporta la reconstrucción exhaustiva de estas inserciones mediante el análisis de lecturas largas.

Esta aproximación nos permitió obtener una determinación inequívoca del estado de metilación de dichas inserciones. En la iteración presente, se han analizado las inserciones somáticas de LINE1 en los tumores obtenidas de aplicar el programa MEIGA-LR 3.0.0 a la cohorte de tumores de los cuales tenemos tejido adyacente para filtrar los LINE1 germinales, así como de las dos líneas celulares derivadas de cáncer de pulmón. MEIGA-LR también genera datos de las inserciones somáticas en el tejido normal, pero en vista del exiguo número estas y al bajo soporte de lecturas de las mismas en los datos de secuenciación masiva, decidimos obviar este análisis, que además se encuadra fuera de los objetivos de esta Tesis.

Los análisis correspondientes a este apartado se realizaron con el programa informático

<../notebooks/*03_LINE1_somaticos/20230317_descripcion_insercciones_somaticas_00.ipynb*>, el cual se encarga de extraer la información necesaria para poder lanzar Methyl-Meiga en modo somático, entendiendo por modo somático, el que permite identificar las lecturas correspondientes a inserciones de LINE1 presentes específicamente en una muestra tumoral y filtrar los que pueden atribuirse al tejido adyacente (normal). El resultado es la información del estado de metilación tanto de las inserciones somáticas como de las regiones adyacentes a las mismas. Asimismo, este programa permite evaluar el estado de metilación de los “alelos vacíos”, esto es, el alelo en el cual no se ha insertado la nueva copia de LINE1. Asimismo, este programa genera tablas de resultados y gráficas de apoyo para la visualización y evaluación de los resultados. En el repositorio adjunto a la presente Tesis de doctorado se muestra la relación completa de los resultados (tablas y gráficas) obtenidos de este modo, que fueron el punto de partida para los posteriores análisis con una serie de *notebooks* de Python.

A continuación, se presenta la figura 149 como un ejemplo ilustrativo de una gráfica generada por Methyl-Meiga, mostrando las lecturas correspondientes a ambos alelos: el alelo con la inserción y el alelo vacío. Cada lectura se visualiza en tres regiones: la región corriente arriba de la inserción (posiciones genómicas 0-20.000), el LINE1 insertado (en este caso de una longitud de 6015 pares de bases y que siempre comienza en la posición 20.000 debido a la forma en que se genera el genoma de referencia artificial) y la región corriente debajo de la inserción. En la figura, se resaltan las citosinas (CpG) no metiladas en color azul, mientras que las metiladas se representan en color rojo. Además, las CpGs hidroximetiladas se identifican con un símbolo de "+". Mediante esta visualización detallada, obtenemos información precisa sobre la metilación en las diferentes regiones genómicas asociadas a la inserción, lo cual nos brinda una mayor comprensión de los patrones de metilación en relación con los elementos LINE1.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

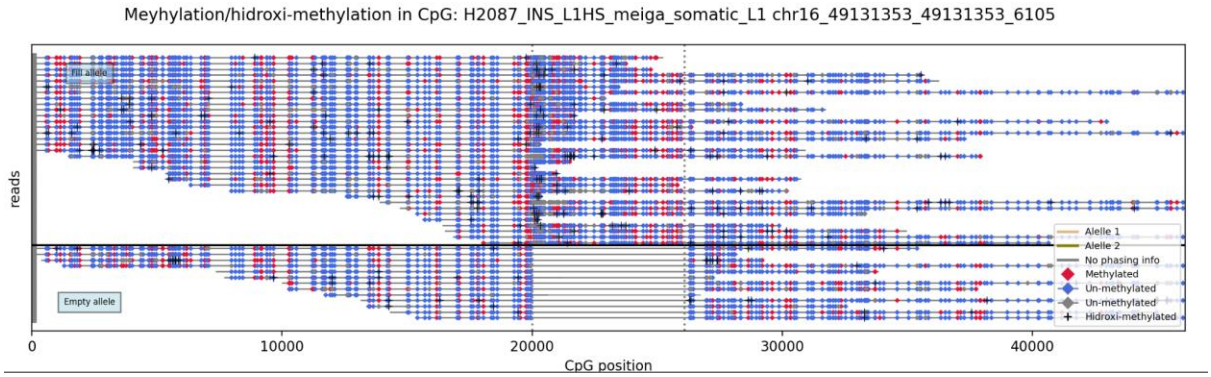


Figura 149: Grafica resultante de ejecutar Methyl-Meiga sobre inserciones somáticas de LINE1

La tabla 38 a continuación presenta los valores de media y mediana correspondientes a las inserciones de los elementos LINE1 en el tejido tumoral de todas las muestras analizadas, para las que teníamos tejido "normal", incluyendo tanto los tumores como las líneas celulares. La media nos indica el valor promedio de estas inserciones, mientras que la mediana nos brinda un valor que se encuentra en el punto medio de la distribución, lo que nos ayuda a entender mejor la tendencia central de estos eventos. Se ha establecido un criterio para definir los promotores y cuerpos de las inserciones de LINE1 en base a su longitud. En las inserciones que superan los 5500 pares de bases, se considera como promotor a las primeras 900 bases de la región 5'UTR, mientras que el resto de la inserción se clasifica como cuerpo. En el caso de las inserciones de menos de 5500 pares de bases, se considera que toda la secuencia corresponde al cuerpo. Esta distinción se realiza debido a que las inserciones somáticas de LINE1 tienden a estar truncadas en el extremo 5' UTR, lo que resulta en la pérdida efectiva del promotor (Hancks y Kazazian 2016; Tubio et al. 2014) . La gran frecuencia de truncamiento hacia 5' de las copias somáticas de LINE1 es muy patente en nuestro conjunto de datos, ya que, de un total de 5621 inserciones, sólo 71 (1.26%) fueron íntegras.

Región	Media metilación	Mediana metilación	Número de elementos
Reg. adyacente (20K pb)	0.386	0.335	5625
Cuerpo LINE1	0.365	0.308	5621
Promotor LINE1	0.315	0.086	71

Tabla 38: Estadísticas de metilación de las inserciones somáticas

Al analizar los niveles de metilación, se observa que la media de metilación en el cuerpo es muy similar a la de las regiones adyacentes, pero ligeramente diferente a la del promotor. Esta diferencia se hace más evidente al considerar la mediana de metilación, donde los valores del promotor difieren significativamente de los del cuerpo y de las regiones adyacentes, como se muestra en la tabla 38 y en el gráfico de violín y bigotes (figura 150). En el gráfico de bigotes (figura 150 derecha) se han incorporado elementos visuales adicionales para resaltar estas diferencias. Se incluye la media (representada como un punto blanco) y la desviación estándar (representada como muescas o *notches*). Los indicadores nos señalan que

existe una diferencia significativa entre los niveles de metilación del promotor, el cuerpo y las regiones adyacentes. Este análisis revela que la media de metilación en el cuerpo de las inserciones somáticas de LINE1 es comparable a la de las regiones adyacentes al punto de inserción. Sin embargo, se observa que el promotor de estas inserciones, cuando está presente, muestra un patrón distintivo de hipometilación.

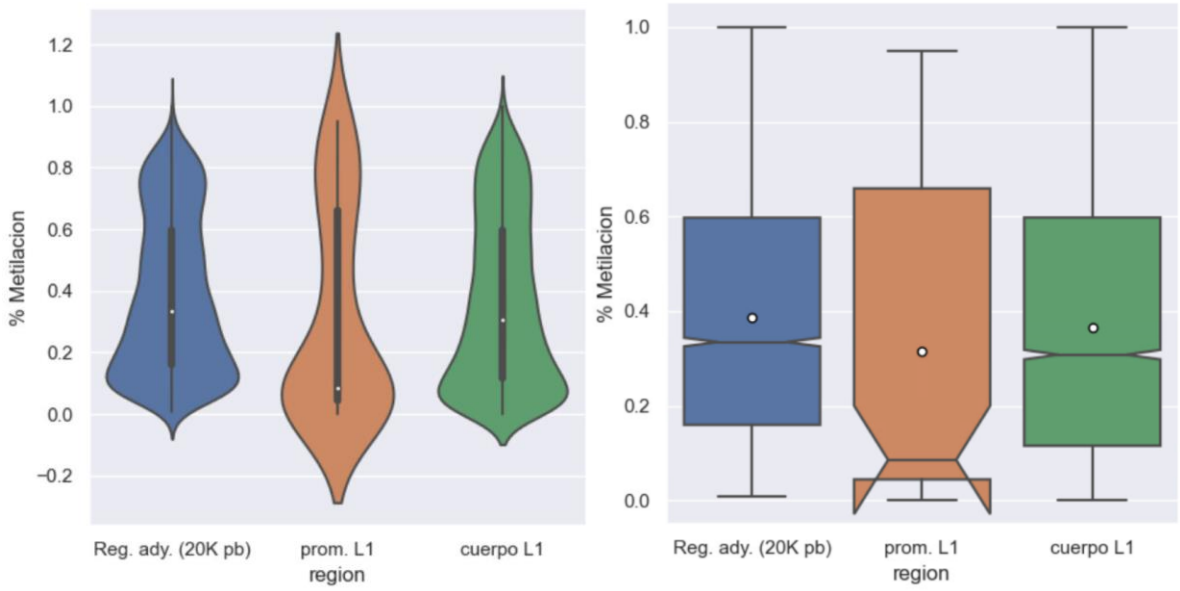


Figura 150: Metilación de inserciones somáticas.
Gráfica de violín (izquierda) y de bigotes (derecha) del estado de metilación de las inserciones somáticas de LINE. (En azul las regiones adyacentes, en naranja los promotores de las inserciones de más de 5500 pb, y el cuerpo del LINE1 en verde)

Las figuras 151 y 152 aportan una visión más concreta del estado de metilación de las copias somáticas de LINE1 en las muestras de tumores analizadas (aquellas de las cueles disponemos de tejido adyacente) para, respectivamente, copias truncadas y copias no truncadas. Para llevar a cabo este análisis, se utilizó el paquete Biopython 1.81 **(https://biopython.org/),** el cual implementa diferentes algoritmos de alineamiento, incluyendo Needleman-Wunsch, Smith-Waterman, Gotoh y Waterman-Smith-Beyer, tanto globales como locales. Se obtuvieron las secuencias consenso de todas las inserciones somáticas proporcionadas por MEIGA-LR y se alinearon con la secuencia consenso del elemento L1HS En nuestro caso, se asignaron puntuaciones de 1, -1, -1 y -1 a las coincidencias (*match*), las no coincidencias (*mismatch*), la apertura de un espacio (gap) y la extensión de un espacio (*gap extension*), respectivamente. Se calculó la media, mediana y moda de la metilación de todas las CpG en ventanas no solapantes de 25 pares de bases a lo largo de estos alineamientos, que fueron remuestreados a una resolución 10 veces mayor, interpolando los valores intermedios de forma lineal y se aplicó un suavizado utilizando ventanas de Hamming con un tamaño de 50 valores. Los *scripts* usados para generar las gráficas, así como las siguientes gráficas y las de las medias y las modas de todos los elementos agrupados y por cada muestra se encuentran en los dos *notebooks* a continuación.

<../notebooks/3_L1_somatico/20230513_Pregunta3_inserciones_somaticas_MENOR_5500_graficas_05.ipynb>

<../notebooks/3_L1_somatico/20230512_Pregunta3_inserciones_somaticas_mas_6000pb_graficas_02.ipynb>

Tal y como se aprecia en la figura 151, hay una acumulación de inserciones en 3' UTR como era de esperar debido al mecanismo de inserción de los LINE1, que no siempre llega a insertar la secuencia completa, produciéndose truncamientos en 5' UTR. La media de la metilación en este tipo de inserciones varía entre 0.2 y 0.5, en línea con la de las regiones adyacentes a la inserción. Esto contrasta con las inserciones completas (figura 152) cuya media de metilación en el cuerpo es mayor que la de las inserciones truncadas, en tanto que las regiones correspondientes a sus promotores es muy inferior a la del resto de sus cuerpos y con respecto a la de la región genómica.

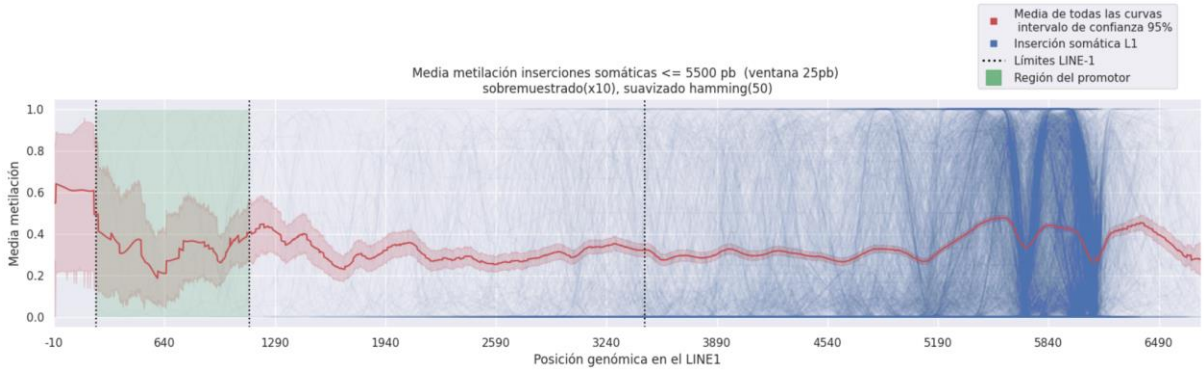


Figura 151 Gráfica suavizada de la metilación de las inserciones somáticas de LINE1 menores de 5500 pares de bases en la cohorte de tumores.

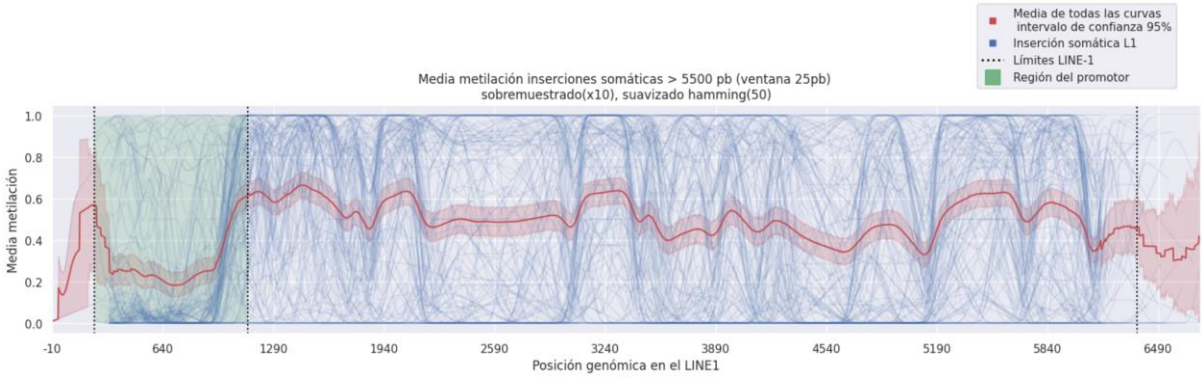


Figura 152: Gráfica suavizada de la metilación de las inserciones somáticas de LINE1 mayores de 5500 pares de bases en la cohorte de tumores.

8.4.1 Patrones de metilación de inserciones somáticas de LINE1 por lectura

A diferencia de los análisis de elementos germinales donde es muy difícil separar las lecturas correspondientes al tumor y a la contaminación. En los análisis de inserciones somáticas solo utilizamos las lecturas que contiene la inserción, por lo que son lecturas solo del tumor, no teniéndonos que preocupar por las lecturas correspondientes a la contaminación. Esto posibilita la realización gráficas de las medias por lectura sin ningún tipo de ruido (como ocurre en análisis similares en LINE1 germinales (datos no mostrados)). Este tipo de gráficas nos permiten constatar de forma ortogonal, algo que ya se había observado en los análisis realizados previamente, de que las regiones del promotor de aquellas inserciones que lo contienen tienen una metilación diferente que las regiones adyacentes (generalmente menor). Para la figura 153 se realizó, por cada lectura de cada inserción, la diferencia entre la media de la metilación de las regiones adyacentes menos la media de metilación de la inserción, haciendo una curva de densidad (KDE) por grupos de tamaño. Se puede comprobar

como para tamaños de inserción menores de 5500pb la curva se aproxima mucho a una campana de Gauss centrada en el 0, indicándonos que las diferencias entre la metilación de las regiones adyacentes y la de los LINE1 tienden a ser similares. Por el contrario, en las inserciones de más de 5500 pb, la curva presenta una distribución bimodal, con uno de los nodos centrado en el 0 como el resto de las inserciones y otro centrado alrededor de -0.3 indicándonos que algunas lecturas no siguen el patrón de las inserciones de menos de 5500 pb.

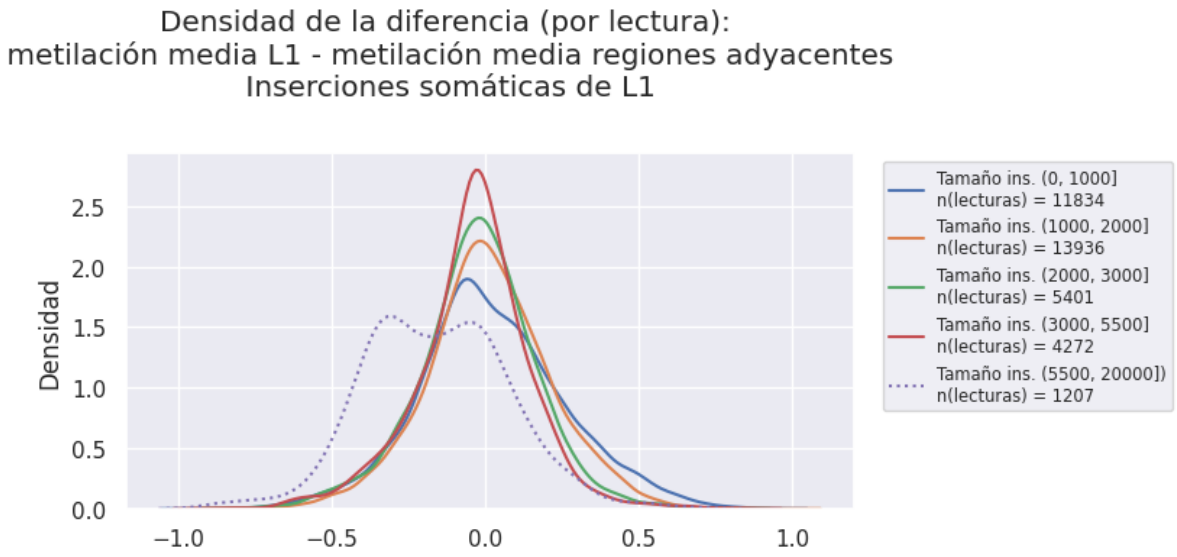


Figura 153: Gráfica de densidad (KDE) de la metilación por lectura de las inserciones somáticas desglosadas por tamaño de la inserción.

Para precisar mejor el patón observado, ser realizó otra gráfica (figura 154) comparando la densidad (KDE) de la diferencia entre la media de metilación de las regiones adyacentes menos la media de la metilación de la inserción, por lectura de todas las inserciones de menos de 5500 pb, y en otras dos curvas la diferencia entre la media de metilación de las regiones adyacentes y la de la inserción por lectura, de por un lado el cuerpo del LINE1 insertado y por otro el promotor. En este caso se ve claramente que el cuerpo tiene una distribución similar a una curva normal centrada en el 0 como las inserciones de menos de 5500 pb. En cambio, la metilación de los promotores sigue una curva bimodal, con uno de los nodos alrededor -0.06. Esto confirma que la metilación del promotor de las inserciones somáticas del line presentan un patrón de metilación diferente al del cuerpo y al de las regiones adyacentes (generalmente menor). Por el contrario, el cuerpo suele presentar un patrón de metilación similar al de las regiones adyacentes.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

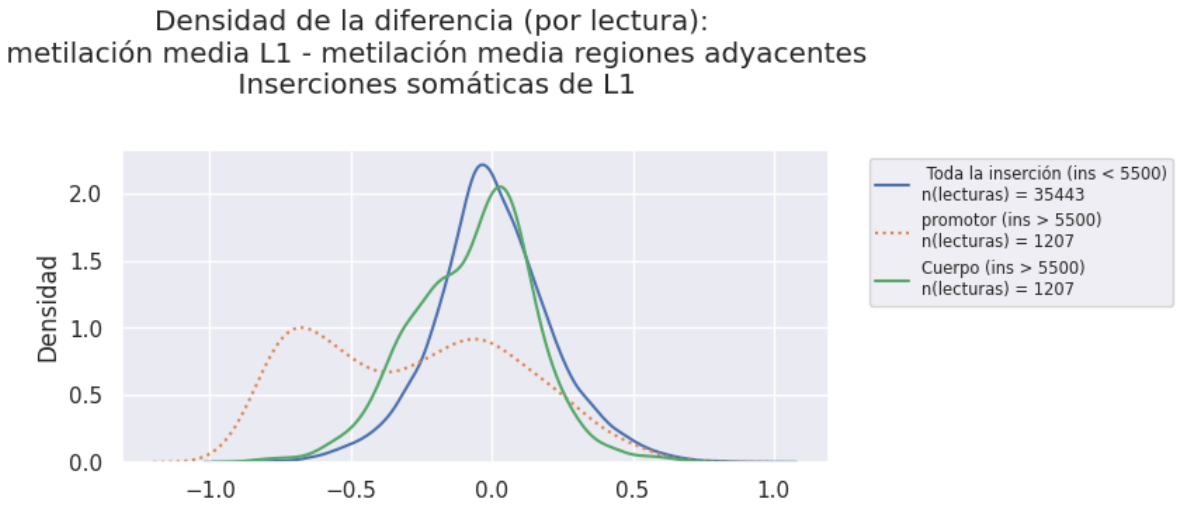


Figura 154: Grafica de densidad (KDE) de la metilación por lectura de las inserciones somáticas.
Inserciones de menos de 5500pb, frente al cuerpo y promotor de aquellas con tamaño mayor de 5500 pb (N representa el número de lecturas de cada grupo).

Para terminar de caracterizar las inserciones se realizó una gráfica de la densidad media de metilación por lectura e inserción de las regiones promotoras y el cuerpo de las inserciones de 5500pb (figura 155). En ella se puede observar que los promotores de las inserciones de LINE1 están mayoritariamente desmetilados, mientras que el cuerpo la distribución es más plana, con un nodo en la región de alta metilación.

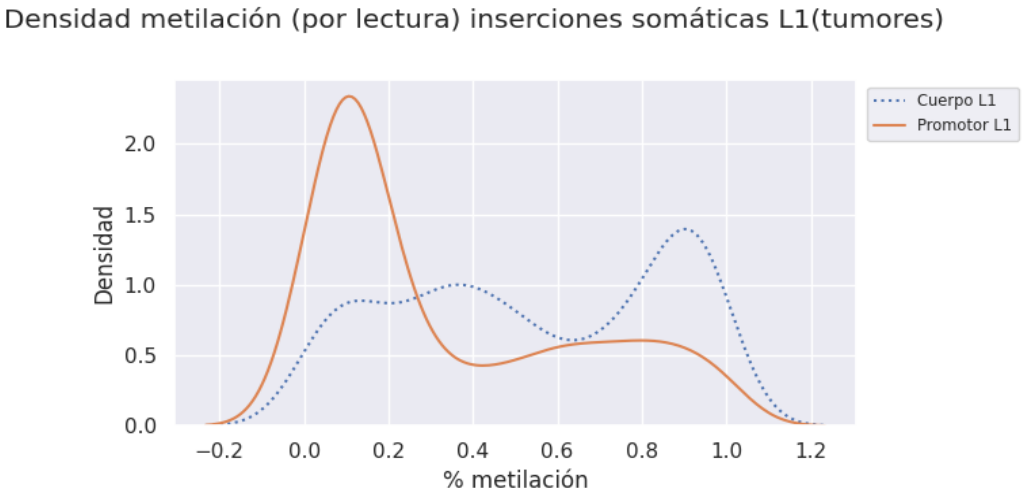


Figura 155: Gráfica de la media de metilación por lectura de las regiones promotoras y el cuerpo de las inserciones mayores de 5500 pb

Nota: graficas generadas con el siguiente notebook.

<..*notebooks*/3_L1_somatico/20230420_patrones_metilacion_L1_somaticos_densidad_media_L1_reg_adyacentes_por_lectura_00.ipynb>

8.4.2 Discusión

Estos hallazgos demuestran que la metilación en el promotor de las inserciones somáticas completas de LINE1 tiende a estar significativamente reducida en comparación con las regiones adyacentes. Tal marcada hipometilación del promotor podría tener implicaciones funcionales en la regulación de la actividad de LINE1 y en la expresión génica en el contexto tumoral. Considerando el proceso de retrotransposición de LINE1 y las características epigenéticas asociadas, este fenómeno difícilmente puede explicarse como una "herencia" directa de los perfiles del estado de metilación del elemento fuente. Durante la retrotransposición de LINE1 (ver apartado 3.2.3), las copias derivadas se generan a partir de un intermediario de ARN que se libera al citoplasma para su procesamiento, traducción y ensamblaje de forma previa a su traslado al núcleo donde se produce la inserción somática correspondiente. Dado que este intermediario de ARN no porta de información epigenética alguna, las posibilidades de una transferencia directa de la metilación de las copias fuente a las copias derivadas son extremadamente remotas. En su lugar postulamos la posibilidad de que existan mecanismos de protección a la remetilación *de novo* en la región promotora de las inserciones somáticas de LINE1. Algunos podrían estar relacionados con cambios en la conformación estructural o la accesibilidad de la cromatina. Otros mecanismos podrían estar mediados por factores proteicos que impidan de manera específica la acción de las enzimas metiltransferasas, evitando la metilación de la región promotora de las nuevas copias.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8.5 LAS COPIAS HIJAS DE LOS LINE1 ACTIVOS EN EL SER HUMANO, NO HEREDAN EL PATRÓN DE METILACIÓN DE LOS ELEMENTOS FUENTE DE LOS CUALES PROVIENEN

Debido a la naturaleza de las inserciones sonaticas de los LINE1 y su mecanismo de inserción, la mayoría están truncadas y o contienen inversiones en el extremo 5', además solo el 20% por ciento de ellas son transducciones. Por ello las posibilidades de contestar a la pregunta 4.1.4 *"Las copias hijas de los LINE1 activos en el ser humano, ¿heredan el patrón de metilación de los elementos fuente de los cuales provienen cuando se movilizan en los tumores?"* se antojaba a priori muy complicado con una corte de muestras tan pequeña. Afortunadamente sí que hay inserciones somáticas en nuestra cohorte que cumplen los criterios necesarios de ser transducciones (necesario para conocer el elemento fuente de origen) y mayores de 6000 pares de bases (para poder estudiar el promotor). Tras una inspección inicial con el IGV de los datos proporcionados por Methyl-Meiga de los posibles candidatos (24) se constató que efectivamente la mayoría (23) contienen el promotor intacto y la zona transducida, lo que nos permite conocer el elemento fuente del cual provienen.

Tal como se muestra en la siguiente figura, para realizar el análisis se compararon las inserciones somáticas íntegras de los LINE1 en los tumores con el elemento fuente del que provienen.

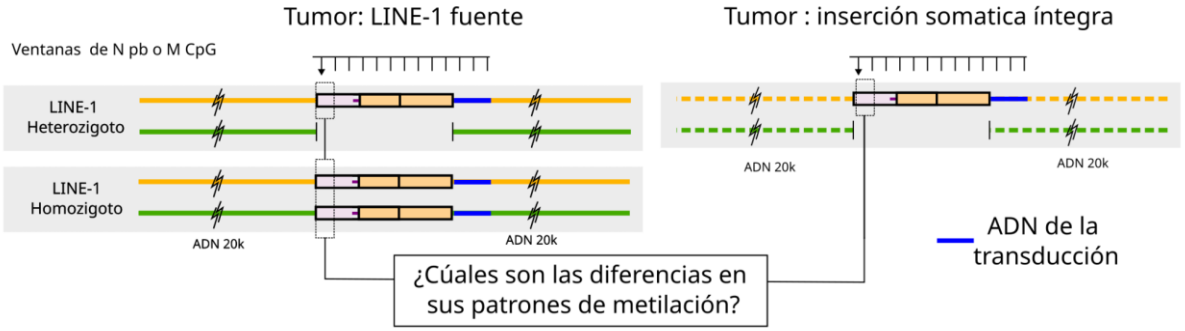


Figura 156: Diagrama de cómo se realizó el análisis de los patrones de metilación de las copias hijas íntegras de LINE1.

Para realizar este análisis se partió de dos conjuntos de datos

a) Datos de LINE1 fuente: Resultado de ejecutar Methyl-Meiga en modo germinal sobre toda la cohorte de tumores usando el fichero **<../3_LINE1_fuente/FL-L1_for_Me-Me_Consensus_GRCh38_v1.tsv>** de los LINE1 íntegros germinales descrito en el apartado 7.2.3.3, este fichero incluye los LINE1 que se conoce que presentan actividad.
b) Datos de inserciones somáticas: Resultado de ejecutar Methyl-Meiga en modo somático sobre toda la cohorte de tumores usando para ello dos aproximaciones
   i) Generando el fichero de entrada para Methyl-Meiga a partir de los ficheros de inserciones somáticas proporcionados por MEIGA-LR y filtrados con datos del tejido adyacente, tal como se describe en el apartado (7.2.6.1). Para generar los cromosomas artificiales correspondientes a cada inserción de LINE1 se usó la secuencia consenso proporcionada por MEIGA-LR.
   ii) Generando el fichero de entrada para Methyl-Meiga a partir de los ficheros de inserciones somáticas proporcionados por MEIGA-LR filtrado con datos de xTea a partir de datos de Illumina (ver apartado 7.2.6.2). Para generar los cromosomas

artificiales correspondientes a cada inserción de LINE1 se usó la secuencia consenso L1HS.

Un problema de este análisis en particular es la dificultad de comparar la secuencia de la inserción (que puede estar truncada y además contiene una transducción) con la del elemento fuente si estos no están bien alineados. Por ejemplo en la inserción somática de la figura 157 no resulta sencillo indicar si las CpG están en las mismas posiciones que las CpG equivalentes del elemento fuente.

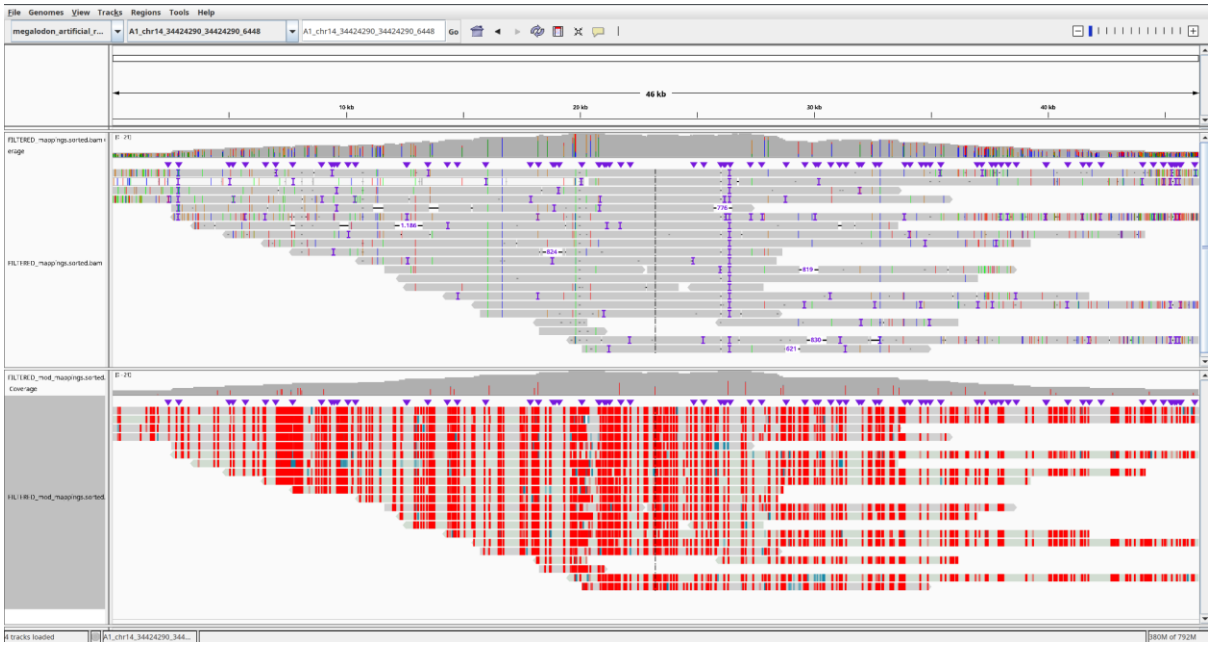


Figura 157: Visualización en el IGV de un ejemplo de una inserción somática integra usando la secuencia proporcionada por MEIGA-LR (muestra NCI-H2087).

Se idearon varios procedimientos para tratar de hacer alineamientos a partir del CIGAR del bam, pero se desestimaron debido a su complejidad. En última instancia los alineamientos se realizaron directamente en los *scripts* de Python. Donde usando funciones de alineamiento del paquete Biopython, se consiguieron alinear sin saltos (gap) las secuencias de la inserción somática con la secuencia del elemento fuente del que provienen, permitiendo así su comparación. (La técnica de alineamiento se describe con cierto detalle en el apartado 8.4 y se utilizó para generar los alineamientos de las inserciones de las figuras 151 y 152)

Para realizar las gráficas comparando las medianas de la metilación entre la transducción y el elemento fuente del cual provienen se creó el programa (notebook de Python)

<../notebooks/04_copias_hijas/20230510_copias_hijas_LINE-1_activos_gráficas.ipynb>

Este programa parte de los datos de Methyl-Meiga de los elementos germinales de las muestras tumorales y tejido adyacente, y de los datos de las inserciones somáticas en las muestras tumorales. El *script* crea dos tipos de gráficas: Unas contienen las medianas de la metilación por ventanas de 25 pares de bases , en el denominado espacio genómico (figura 158)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

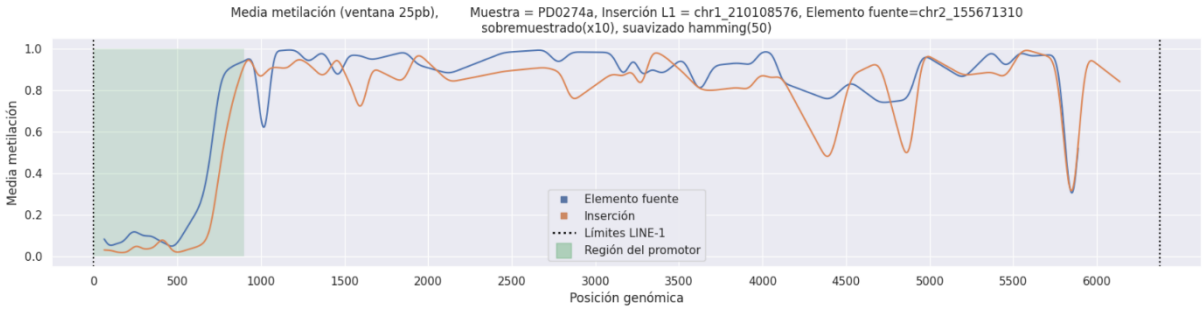


Figura 158: Ejemplo de gráfica comparativa con patrones similares de una inserción somática y el elemento fuente del que deriva (espacio genómico)

y otras contienen la mediana de la metilación en ventanas de un número determinado de CpG (2) sin importar a qué distancia se encuentre estás, es el denominado espacio CpG (figuras 159 y 160).

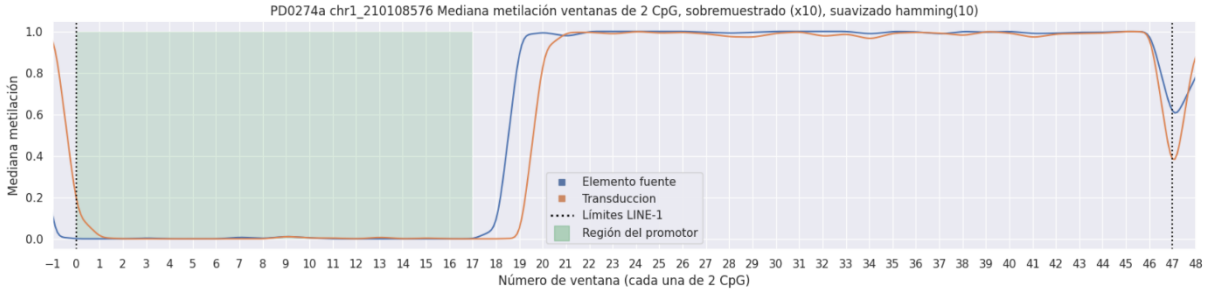


Figura 159: Ejemplo de gráfica comparativa con patrones similares de una inserción somática y el elemento fuente del que deriva (espacio CpG)

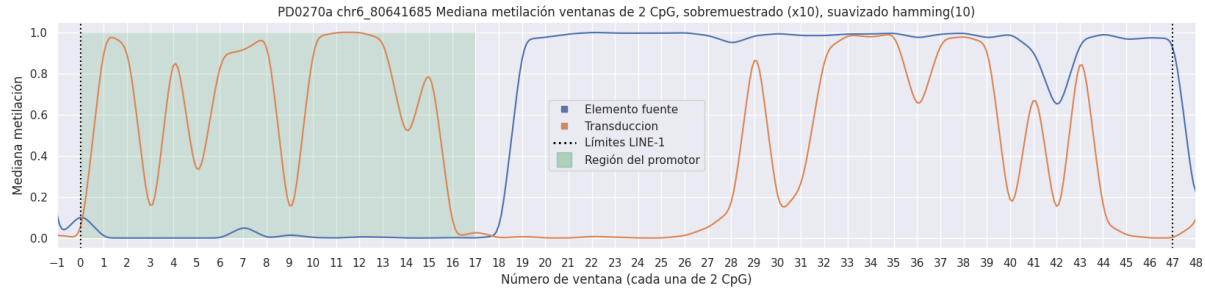


Figura 160: Ejemplo de gráfica comparativa con patrones distintos de una inserción somática y el elemento fuente del que deriva (espacio CpG)

Esta información por ventana se remuestreó (x10) y se le aplicó la media de Hamming en ventanas rodantes de 50 valores en el espacio genómico y ventanas de 10 valores en el espacio CpG. Estas gráficas son el resultado de múltiples pruebas alterando los parámetros anteriormente mencionados, el tipo de suavizado (hamming, triangular, media pura…) y son un compromiso entre el número de CpG por ventana y el tamaño de la ventana rodante. Esto nos ha permitido eliminar ruido sin distorsionar la señal de tal forma que nos permite distinguir los patrones subyacentes. También se realizaron las mismas pruebas con las medias de la metilación, pero las gráficas son mucho más ruidosas por lo que se descartaron para el análisis. (El resto de las gráficas no son mostradas por brevedad se pueden observar en el propio notebook).

Como se puede comprobar en algunos elementos, el patrón de metilación de la inserción sigue con bastante precisión el patrón de metilación del elemento fuente (figuras

158 y 159 correspondientes a la misma inserción). Pero no en todas las inserciones se observa ese patrón (figura 160), siendo los casos bastante heterogéneos. Todo ello nos lleva a deducir que las copias hijas (inserciones) **no heredan los patrones de metilación de los elementos de los cueles proceden**, aunque debido al reducido número de elementos es una afirmación que debe corroborarse en una cohorte más numerosa.

Para crear la tabla de medias, medianas y su significancia estadística se usó el programa (notebook Python)

<../notebooks/04_copias_hijas/6.7_copias_hijas_LINE1_activos_Tabla_diferencias.ipynb>

parte de los mismos datos que el programa anterior y las estadísticas se realizaron en el espacio genómico en ventanas de 10, 25, 50, 500, 10,000 (esta última corresponde a todo el LINE1) pares de bases en las regiones de corriente arriba (-20.000 - 0), LINE1 (0 - tamaño del LINE1), corriente abajo (fin del LINE1, fin del LINE1 + 20.000)

El objetivo de usar ventanas y no comparar directamente las CpG es que a pesar de que estamos comparando el mismo elemento (copia hija y elemento fuente) es muy difícil conseguir alinearlas a nivel de CpG. Además, al ser grupos de CpG se elimina ruido y se aumenta la significancia de las comparaciones.

Los resultados de las diferencias entre las medias y medianas de la metilación entre las inserciones somáticas y los elementos fuentes del que provienen (todo el elemento y la región del promotor) se muestran en la siguiente tabla 39. Como se puede observar, por inspección *de visu* no se puede apreciar ningún patrón entre la metilación de las copias hijas con respecto a las copias de las cuales derivan, más allá de que comparten una tendencia de desmetilación general en el promotor y una tendencia del cuerpo de los LINE1 a estar mucho más metilados, característico de la metilación de los LINE1 en los tumores. Estos resultados están en la misma línea de los obtenidos analizando las gráficas comparativas de las inserciones somáticas hijas, con el elemento fuente del que provienen (figuras 163, 164,165), de **que las copias hijas no heredan los patrones de metilación del elemento del cual proceden**. Tal como discutíamos en el apartado 8.4.2 es difícil imaginar un mecanismo de “heredabilidad” de la metilación, algo que concuerda con los resultados aquí presentados y que se analiza en un contexto global en el apartado de discusión general (9.3).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Muestra	Elemento fuente		Inserción somática				MEDIA								MEDIANA							
							Elementos fuente			Inserción somática			Diff. Elem. fuente - inserción		Elementos fuente			Inserción somática			Diff. Elem. fuente - inserción	
	Crom.	Inicio	Crom.	Inicio	Tamaño	Orien.	Prom.	Cuer.	Diff. Cuer. - prom.	Prom.	Cuer.	Diff. Cuerpo - prom	Prom.	Cue,	Prom.	Cuer.	Diff. Cuer. - prom	Prom.	Cuer.	Diff. Cuer. - prom	Prom.	Cuer.
H2009	chr1	86674895	chr2	55182701	6157	+	No existe															
H2009	chr2	155671310	chr15	20990269	6343	-	0.497	0.836	0.339	0.326	0.768	0.442	0.171	0.068	0.492	0.996	0.504	0.475	0.988	0.514	0.018	0.008
H2087	chr6	13190783	chr14	34424290	6442	+	0.446	0.603	0.157	0.422	0.904	0.483	0.025	-0.301	0.088	0.978	0.890	0.414	0.996	0.582	-0.325	-0.018
H2087	chr6	13190783	chr2	39939368	6519	+	0.309	0.917	0.608	0.070	0.058	-0.012	0.240	0.859	0.459	0.996	0.537	0.000	0.000	0.000	0.459	0.996
H2087	chr6	13190783	chr2	222483709	6196	+	0.309	0.917	0.608	0.849	0.667	-0.181	-0.540	0.250	0.459	0.996	0.537	0.992	0.988	-0.004	-0.533	0.008
H2087	chr6	13190783	chr5	155188432	6596	-	0.446	0.603	0.157	0.550	0.933	0.383	-0.104	-0.330	0.088	0.978	0.890	0.502	0.996	0.494	-0.414	-0.018
H2087	chr6	13190783	chrX	52783101	6532	+	0.446	0.603	0.157	0.695	0.624	-0.071	-0.249	-0.021	0.088	0.978	0.890	0.969	0.973	0.004	-0.880	0.006
H2087	chr10	85355506	chr7	54329028	6116	-	0.446	0.603	0.157	0.097	0.089	-0.008	0.350	0.515	0.088	0.978	0.890	0.009	0.004	-0.005	0.079	0.975
H2087	chr22	28663283	chr16	13324947	6030	+	0.525	0.605	0.080	0.132	0.088	-0.044	0.393	0.517	0.496	0.980	0.484	0.002	0.000	-0.002	0.494	0.980
H2087	chr22	28663283	chr2	11879618	6681	-	0.446	0.603	0.157	0.117	0.101	-0.016	0.330	0.503	0.088	0.978	0.890	0.000	0.000	0.000	0.088	0.978
	Total líneas celulares						0.430	0.699	0.269	0.362	0.470	0.108	0.068	0.229	0.261	0.985	0.724	0.374	0.549	0.176	-0.113	0.435
PD0266a	chr2	155671310	chr12	17991140	6394	+	0.682	0.804	0.122	0.207	0.595	0.388	0.476	0.209	0.924	0.990	0.067	0.012	0.949	0.937	0.912	0.041
PD0266a	chr6	29952435	chrX	58146950	6647	+	0.641	0.861	0.219	0.108	0.127	0.019	0.534	0.734	0.900	0.995	0.095	0.006	0.005	-	0.894	0.990

																			0.001		
PD0270a	chr2	155671310	chr6	80641685	6111	+	0.456	0.864	0.408	0.501	0.582	0.081	-0.044	0.282	0.486	0.991	0.505	0.535	0.955	0.420	-0.049	0.036
PD0270a	chr2	155671310	chrX	85482808	6130	-	0.456	0.864	0.408	0.789	0.733	-0.056	-0.333	0.131	0.486	0.991	0.505	0.992	0.970	-0.023	-0.506	0.022
PD0274a	chr2	155671310	chr1	169831730	6399	+	0.388	0.904	0.515	0.160	0.225	0.065	0.228	0.679	0.488	0.996	0.508	0.002	0.000	-0.002	0.486	0.996
PD0274a	chr2	155671310	chr1	210108576	6379	-	0.388	0.904	0.515	0.860	0.891	0.030	-0.472	0.013	0.488	0.996	0.508	0.994	0.996	0.002	-0.506	0.000
PD0274a	chr2	155671310	chr10	33722438	6291	+	0.388	0.904	0.515	0.279	0.843	0.564	0.109	0.061	0.488	0.996	0.508	0.418	0.988	0.571	0.071	0.008
PD0274a	chr2	155671310	chr4	138682983	6233	+	0.388	0.904	0.515	0.304	0.774	0.470	0.084	0.129	0.488	0.996	0.508	0.459	0.990	0.531	0.029	0.006
PD0274a	chr2	155671310	chr6	111719636	6367	-	0.388	0.904	0.515	0.278	0.851	0.574	0.111	0.052	0.488	0.996	0.508	0.435	0.996	0.561	0.053	0.000
PD0274a	chr2	155671310	chrX	142617656	6313	+	0.388	0.904	0.515	0.119	0.372	0.253	0.269	0.531	0.488	0.996	0.508	0.008	0.031	0.024	0.480	0.965
PD0277a	chr2	155671310	chr11	95649314	6655	-	0.368	0.725	0.357	0.465	0.498	0.033	-0.097	0.226	0.472	0.984	0.513	0.507	0.614	0.107	-0.035	0.371
PD0277a	chr20	23426108	chr10	115241074	6133	+	0.445	0.816	0.371	0.893	0.799	-0.094	-0.448	0.017	0.493	0.995	0.502	0.982	0.984	0.002	-0.489	0.011
PD0312a	chrX	11707248	chr7	121384840	7421	+	0.381	0.562	0.180	0.081	0.683	0.602	0.300	-0.121	0.369	0.910	0.541	0.000	0.971	0.971	0.369	-0.061
PD0312a	chrX	11707248	chr7	126484656	6955	-	0.381	0.562	0.180	0.025	0.033	0.008	0.357	0.529	0.369	0.910	0.541	0.000	0.000	0.000	0.369	0.910
	Total tumores primarios						0.439	0.820	0.381	0.362	0.572	0.210	0.077	0.248	0.531	0.982	0.451	0.382	0.675	0.293	0.148	0.307
	Total todas las muestras						0.434	0.760	0.325	0.362	0.521	0.159	0.072	0.238	0.396	0.983	0.587	0.378	0.612	0.234	0.018	0.371

Tabla 39: Estadísticas de la metilación de las copias hijas y los elementos fuentes de los cuales derivan

Abreviaturas: Orien. = Orientación de la inserción, Prom. = Promotor LINE1 (900 primeros pares de bases en 5' UTR), Cuer. = Cuerpo del LINE 1, Diff. = Diferencia

8.6 ANÁLISIS ADICIONALES

8.6.1 Patrones de hidroximetilación de las inserciones somáticas

Realicé el análisis de la hidroximetilación de las inserciones somáticas para ampliar nuestro entendimiento sobre este fenómeno epigenético. Al igual que en los análisis previos de todo el genoma (apartado 8.1.2) y de los LINE1 germinales (no mostrados), los resultados revelaron que los niveles de hidroximetilación son consistentemente bajos en las muestras estudiadas.

Observamos que los valores de hidroximetilación son generalmente inferiores a 0.2, con una media cercana a 0.01, tanto en las inserciones de menos de 5.500 pares de bases (Figura 161) como en las de más de 5.500 pb (Figura 162). Estos resultados indican que la hidroximetilación de las inserciones somáticas de LINE1 es un fenómeno poco frecuente y presenta una distribución heterogénea en nuestras muestras de estudio. Estos hallazgos sugieren que el proceso de hidroximetilación en estas inserciones puede seguir un patrón estocástico, donde la aparición de grupos hidroximetilo por oxidación del grupo metilo de las metilcitosinas ocurre de manera aleatoria y no está determinada por señales específicas o regiones de interés. Estos resultados respaldan la noción de que la hidroximetilación puede desempeñar un papel secundario en la regulación epigenética de las inserciones somáticas de LINE1.

Nota : Los *notebooks* utilizados para generar estas gráficas son los siguientes:

<../notebooks/3_L1_somatico/20230518_Pregunta3_HMET_inserciones_somaticas_MENOR_5500_graficas_06.ipynb>

<..*notebooks*/3_L1_somatico/20230518_Pregunta3_HMET_inserciones_somaticas_mas_6000pb_graficas_03.ipynb>

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

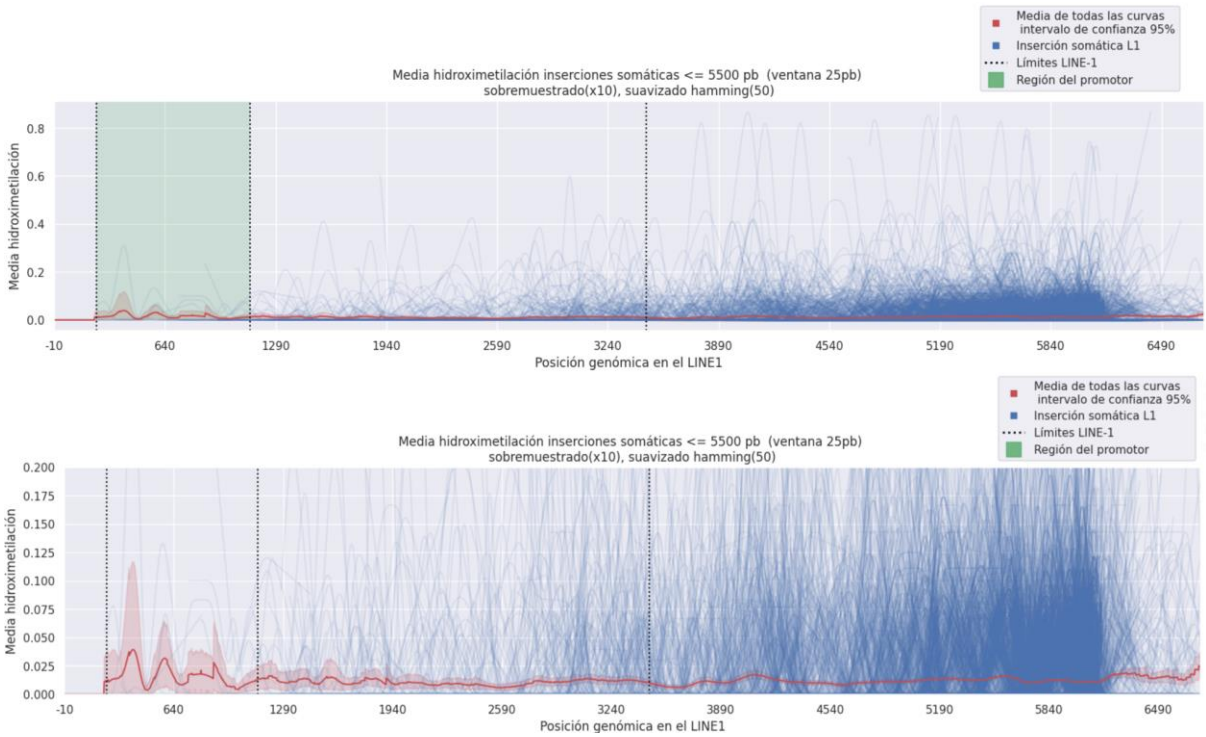


Figura 161: Gráfica suavizada, de la hidroximetilación de las inserciones somáticas de LINE1 menores de 5500 pares de bases en la cohorte de tumores.
Escala eje Y: 0-1 (panel superior), detalle, escala eje Y: 0-0.2 (panel inferior)

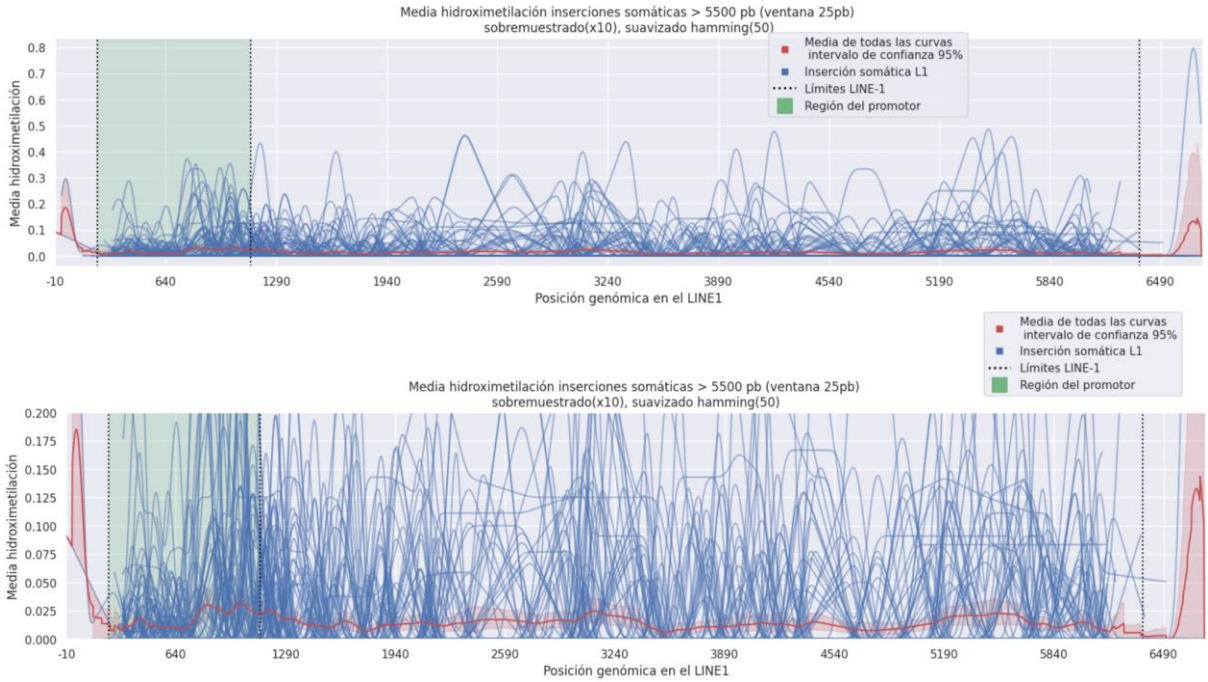


Figura 162: Gráfica suavizada, de la hidroximetilación de las inserciones somáticas de LINE1 mayores de 5500 pares de bases en la cohorte de tumores.
Escala eje Y: 0-1 (panel superior), detalle, escala eje Y: 0-0.2 (panel inferior)

8.6.2 Epivariación proximal, falso hallazgo

Es conocido en la literatura que hay una correlación alta entre el estado de metilación del promotor de los LINE1 y las primeras CpG de la región adyacente. En el artículo (Lanciano et al. 2023) analizan si la metilación del LINE1 influencia el estado de metilación de las regiones adyacentes donde se encuentra el LINE1. El estudio realizado en líneas celulares incluye líneas celulares tumorales. Para ello se centran solo en los LINE1 heterocigotos, ya que permite comprobar cómo es el estado de metilación del ADN con y sin el LINE1. De los cuatro casos posibles para este estudio sólo interesan los casos en que el estado de metilación de las regiones adyacentes al LINE1 son distintas al estado de metilación del promotor (casos marcados como **informativo** en la figura 163). Corroboran que efectivamente que el LINE1 afecta a la metilación de la región adyacente hasta 300 pb, siendo el efecto menos marcado a medida que aumenta la distancia de la CpG del inicio del promotor del LINE1.

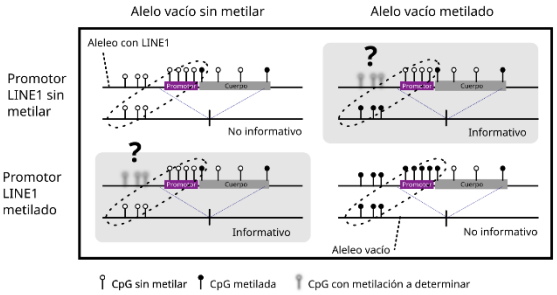


Figura 163: Comparación de epialelos informativos y no informativos de LINE1.
Adaptado de (Lanciano et al. 2023)

Siguiendo un razonamiento similar decidimos comparar el estado de metilación de las regiones adyacentes de los dos alelos tanto de las inserciones somáticas (que por definición son todas heterocigotas, pues solo se insertan en uno de los alelos), como de los LINE1 germinales, íntegros heterocigotos.

El análisis consistió en analizar relación de la media de metilación de las regiones adyacentes de las inserciones somáticas de LINE1 en varios tumores de nuestra cohorte en ventanas de 5 o 20 CpG, comparando el alelo que contenía el LINE1 (alelo lleno) con el que no lo tenían (alelo vacío): *Fold_change* = ***met(alelo lleno) / met(alelo vacío)***. Se realizó un test *"U de Mann–Whitney"* corregido por Bonferroni para ver marcar aquellas ventanas cuya mediana era significativamente diferente.

(https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html accedido sep. 2022)

(https://www.statsmodels.org/stable/generated/statsmodels.stats.multitest.multipletests.html accedido sep. 2022)

El resultado se muestra en la siguiente gráfica de volcán (figura 164) donde cada punto representa la información de una ventana de 20 CpG (los colores representan los distintos tipos tumorales). Como se puede comprobar en el panel de la izquierda, que corresponde a las inserciones somáticas, hay muchas más ventanas donde la metilación del alelo que contiene el LINE1 es menor que la metilación de la misma ventana en el alelo vacío (rama izquierda). En cambio, en el panel de la derecha correspondiente a los LINE1 germinales, vemos que aproximadamente es igual el número de ventanas significativas donde la metilación del alelo lleno es mayor que la del alelo vacío, que viceversa. Durante cierto tiempo el resultado nos pareció sorprendente, e inicialmente lo achacamos a la influencia de los LINE1.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

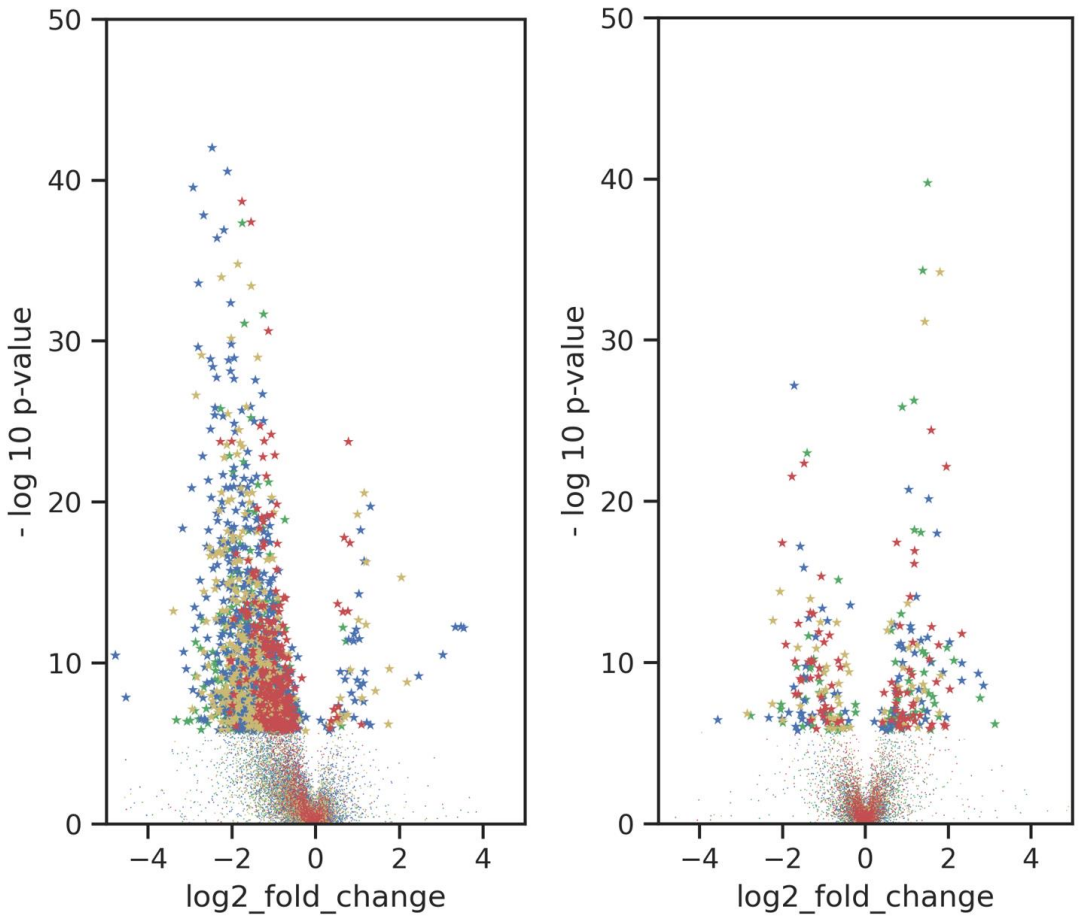


Figura 164: Gráfica de volcán, del log2 fold change de la metilación del alelo lleno / alelo vacío de inserciones somáticas (izquierda) y de inserciones germinales (derecha)

En un análisis más detallado y contemplando posibles problemas del experimento, mi compañera Sonia Zumalave descubrió que era un problema de la contaminación de las muestras tumorales con el tejido adyacente, que normalmente suele estar más metilado. El razonamiento es el siguiente, en las inserciones somáticas de LINE1 las lecturas que contiene la inserción son solo lecturas del tumor (por que el tejido adyacente no tiene la inserción), por el contrario, las lecturas que no contienen la inserción pueden ser tanto del tumor como del tejido adyacente, al estar este normalmente más metilado hace que el denominador se mayor y por lo tanto el cociente menor que cero, haciendo el logaritmo negativo. En cambio, en los LINE1 germinales heterocigotos puede haber contaminación tanto en las lecturas con el LINE1 (alelo lleno) como en las lecturas sin el LINE1 (alelo vacío) haciendo que el cociente sea más neutro.

Para confirmar el resultado se realizaron los mismos análisis en las líneas celulares donde no existe contaminación. Estas son mostradas en las figuras mostradas a continuación, en ellas se puede comprobar que en las inserciones somáticas muestran un patrón equilibrado entre ambas ramas del gráfico de volcán para las inserciones somáticas de LINE1 en dos líneas celulares

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

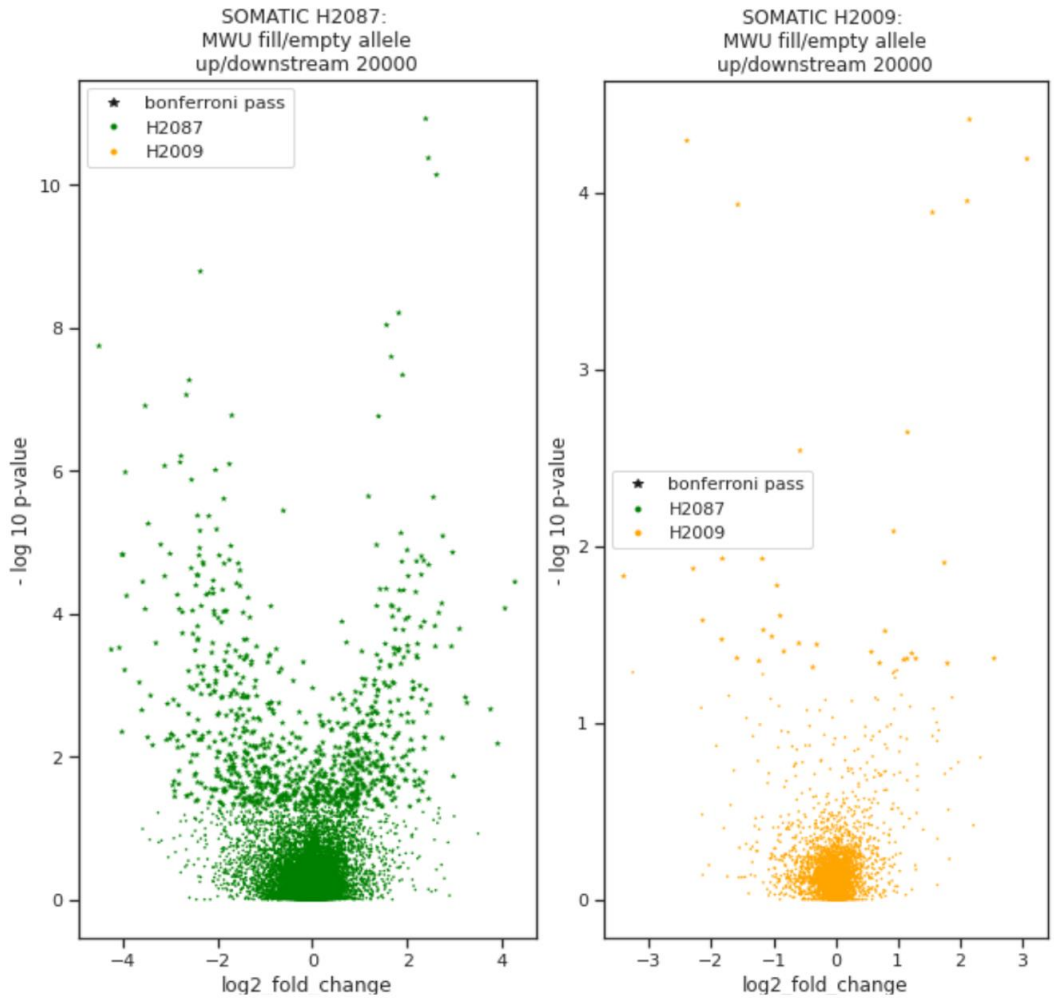


Figura 165: Gráfica de volcán del log2 fold change de la metilación del alelo lleno / alelo vacío de inserciones somáticas en las dos líneas celulares (NCI-H2087 y NCI-H2009)

He decidido incluir este análisis a pesar de no tener un resultado positivo, por dos motivos que me ha parecido muy ilustrativos. A) Muestra de forma muy clara la dificultad que implica el trabajar con muestras contaminadas, haciendo muy difícil un análisis como el de (Lanciano et al. 2023) los cuales usaron líneas celulares soslayando el problema de la contaminación. B) Normalmente no se publican los resultados negativos y considero que es un error pues aportan información muy valiosa que evitaría que otros investigadores invirtiesen tiempo y esfuerzo en una vía infructuosa.

Nota: los *scripts* usados para sacar estas gráficas y otras se encuentran en

<../notebooks/01_LINE1_germinal/stats_volcano_plot_ALL_germinal_win_CpG_20_man_withney_V19_01.ipynb>

<../notebooks/01_LINE1_germinal/stats_volcano_plot_ALL_germinal_win_CpG_5_man_withney_V19_02_multisample.ipynb>

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9. DISCUSIÓN GENERAL

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Discusión” Lápiz sobre papel A4 (May 2017)

9.1 PUESTA A PUNTO DE UNA UNIDAD DE SECUENCIACIÓN DE LECTURAS LARGAS

A lo largo de este manuscrito he intentado dar una visión general del trabajo de doctorado realizado por mi durante estos últimos apasionantes años. Quizá la parte más difícil haya sido sintetizar todos esos conocimientos adquiridos y la más dolorosa el tener que omitir parte del material conseguido con mucho esfuerzo y tesón, ya que haría inabordable el presente manuscrito. En mi defensa he de destacar que la decisión de eliminar ciertos contenidos no disminuye el valor y la relevancia de los esfuerzos realizados. Más bien, se buscó presentar una estructura y un enfoque que permitieran una comprensión más profunda y eficiente de los aspectos fundamentales del estudio, facilitando así el entendimiento y asimilación de los conocimientos por parte de los lectores interesados.

Con la perspectiva que da el conocimiento, es evidente que ciertas decisiones y acciones tomadas durante esta etapa de investigación podrían haber sido abordadas de manera distinta en la actualidad. Al reflexionar sobre ello, algunas tareas hubieran merecido una dedicación más profunda y minuciosa, mientras que otras, ya a toro pasado, habrían merecido ser dejadas de lado, o tal vez nunca haberlas explorado desde un inicio. En este sentido, creo oportuno aprovechar este capítulo de discusión para llevar a cabo un análisis desde esta perspectiva, incluyendo incluso aquellas partes que no se incorporaron al manuscrito final, con el objetivo de enriquecer la tesis doctoral en su totalidad, sin exceder innecesariamente su extensión actual.

Posiblemente, uno de los grandes hitos y un legado derivado del trabajo en esta tesis doctoral (Objetivo 4.2.1) sea el desarrollo desde cero de una **unidad de secuenciación masiva de ácidos nucleicos** de lecturas largas empleando ONT (*facility o core facility)*, lo cual representó un desafío monumental que se inició en el año 2018 tras la mudanza del grupo de investigación a la Universidad de Santiago de Compostela. En ese momento, carecíamos casi por completo de experiencia previa en esta área, lo que implicó enfrentar una curva de aprendizaje muy pronunciada. Fue necesario abordar casi cada aspecto del proceso, pasando por la implementación y adaptación de los protocolos de laboratorio y la adquisición y construcción de aparatos de biología molecular, hasta la conclusión de un sólido soporte informático que respaldara toda la infraestructura requerida.

Un apartado a reseñar, inherente a toda nueva tecnología, es la constante y rápida evolución de los procedimientos, protocolos, y en este caso también de los programas informáticos. De todo ello es un paradigmático ejemplo la evolución de la tecnología de nanoporos de Oxford Nanopore Technologies, pues en un corto periodo de tiempo (4-5 años) hemos asistido a cambios significativos en la preparación de las librerías, en el funcionamiento y rendimiento de las celdas de secuenciación (*Flowcells*), en los programas informáticos para gestionar los secuenciadores, y en el tratamiento y formato de los ficheros de salida. Esto nos obligó a una constante adaptación, reaprendizaje y en algunos casos a empezar de cero con los desarrollos. Resulta irónico que durante todo este proceso de exploración en busca de soluciones a los diversos desafíos que surgieron, nos encontramos con que algunas de estas soluciones ya habían sido desarrolladas en paralelo por la propia casa comercial y otros investigadores independientes. Por un lado, siento cierto orgulloso al comprender que estábamos en la misma línea de pensamiento que otros profesionales en el

campo, lo que valida la pertinencia de nuestras investigaciones y la relevancia de nuestras ideas. Por otro lado, esta experiencia también conlleva la sensación de haber invertido tiempo en caminos que ya estaban siendo trazados por otros.

La implementación de esta tecnología en el laboratorio parte de la necesidad, ya que el estudio de la secuencia completa de las inserciones de transposones y otros elementos en el genoma no hubiese sido posible con los datos de secuenciación de lecturas cortas (NGS). Con esa tecnología esta tarea habría sido altamente limitada e insuficiente, ya que la información que nos proporcionan estas lecturas no permitiría una resolución adecuada de eventos de inserción complejos y extensos. Por el contrario, la tecnología de secuenciación de lecturas largas, que hemos adaptado e implementado con éxito, se convirtió en un recurso imprescindible para lograr una visión más completa y detallada de estos elementos genéticos en nuestras investigaciones. Además, esta tecnología era accesible y, en aquel momento incluso más que ahora, más económica que la otra única alternativa de lecturas largas, PacBio. Además, esta *facility* demostró ser fundamental para el ensamblaje de genomas de referencia en organismos no modelo, como, por ejemplo, el berberecho. La secuenciación de lecturas largas facilitó la resolución de regiones altamente repetitivas y complejas de su genoma, lo que a su vez permitió una reconstrucción más completa y precisa. Este logro fue de gran relevancia para nuestro grupo, ya que brindó una valiosa herramienta para nuestro proyecto en cánceres transmisibles en esta especie.

El trabajo de laboratorio para la puesta a punto y adaptación de la tecnología de ONT fue una de las tareas más gratificantes y absorbentes de la realización de esta Tesis. Tal esfuerzo nos permitió abrir el camino, seguido también por otros investigadores (Schalamun et al. 2019), que finalmente permitió la implantación de la "*core facility* de secuenciación de lecturas largas". Los primeros experimentos nos permitieron descubrir la fascinante tecnología de la secuenciación por nanoporos, así como los muy exigentes requerimientos en cuanto a calidad, integridad y pureza del ADN a secuenciar. Además, durante el desarrollo de esta investigación, nos encontramos con que muchos de los pasos incluidos en los protocolos proporcionados por la casa comercial parecían ser una caja negra ("*black box*") en la que debíamos confiar ciegamente. La falta de transparencia en algunos de estos procesos nos llevó a enfrentar dudas en la implementación de los mismos, así como escaso o nulo margen de mejora si no es a través de prueba y error. El objetivo primario de mi trabajo en esas primeras etapas de mi Tesis se convirtió entonces en conseguir ADN en mayor cantidad y aceptable pureza, así como en rangos de tamaño suficientes para obtener lecturas largas con las que disponer de librerías de calidad que rindiesen una gran cantidad de datos de secuenciación masiva.

El primero de los factores limitantes en la consecución de una librería de calidad es la cantidad de ADN recuperado. Sin llegar a los requerimientos de varios microgramos de ADN para secuenciación con lecturas largas de PacBio, ONT es relativamente costoso, requiriendo un input de 1 a 1,5 microgramos de ADN para iniciar una librería. Los métodos comerciales de aislamiento de ADN a partir de muestras de tejido sólido eran, en ocasiones, insatisfactorios a la hora de obtener suficiente cantidad de ADN, lo que nos llevó a adaptar protocolos clásicos de extracción por método de cloroformo-isoamil-alcohol y al desarrollo de un protocolo de amplificación genómica mediante PCR aleatoria con una phi-polimerasa

(RepliG). Los métodos de extracción de ADN mediante cloroformo-isoamil-alcohol tenían como contraparte la obtención de muestras relativamente más contaminadas, tanto por restos de cloroformo como por contaminantes derivados de la muestra u otros buffers de extracción. A la postre, ello afectaba negativamente a los pasos de adaptación de extremos en la elaboración de librerías, lo que resultaba en una muestra significativa de fragmentos de ADN no adaptados en la librería final, con un bajo o, más bien, bajísimo rendimiento de la celda. Ello nos llevó a modificar incluso más estos protocolos de extracción, añadiendo pasos adicionales de precipitación de polisacáridos y de limpieza del *pellet* de ADN previas a la resuspensión final del ADN. La amplificación del ADN de partida con oligos aleatorios utilizando *RepliG* o GenomiPhi se planteaba pues como una alternativa interesante para paliar las carencias de ADN de extracciones de muestras insuficientes. Este protocolo, no obstante, planteaba un desafío adicional, ya que las estructuras ramificadas derivadas del proceso de amplificación por la phi-polimerasa obturaban de forma prematura los poros, resultando en el agotamiento dramático de las *flowcells* en el lapso de un par de horas. En el curso de nuestras investigaciones nos topamos con ciertos tipos de ADN nativo que provocaban este mismo efecto de obturación de poros, ya sea por algún tipo de proteína que el proceso estándar de extracción y limpieza no era capaz de eliminar o por ultraestructuras terciarias de ADN que, físicamente, no pueden atravesar los nanoporos de las celdas de ONT. Nuestra aproximación para solventar estas dificultades se centró en el tratamiento del ADN con una enzima endonucleasa T7 seguido de una digestión con proteinasa-K y subsiguiente purificación con bolitas (*beads*) magnéticas (experimentos y resultados no mostrados en este manuscrito por brevedad). Esta fórmula permitía recuperar muestras a priori no aptas para librerías de lecturas largas que, si bien se mostraron más propensas a acumular artefactos puntuales de SNPs y contaminaciones de ADN exógeno y que, en parte, sesgaban la muestra hacia ADN repetitivo, resultaron más que suficientes para la detección de múltiples variaciones estructurales en una gran diversidad de muestras.

Evidentemente, el cometido último y ventaja principal de la secuenciación de lecturas largas con ONT frente a la secuenciación de lecturas cortas de Illumina, mucho más precisa en la detección de variantes puntuales, es, como su nombre indica, la posibilidad de disponer de lecturas de ADN de mayor longitud, del orden de múltiples kilo bases. De hecho, sucesivas publicaciones, *whitepapers* y publicaciones en blogs de ONT se vanagloriaron en diversas ocasiones de la obtención de lecturas cada vez más largas, en una suerte de propaganda comercial similar a la carrera espacial durante la Guerra Fría. A la memoria me viene, por ejemplo, el manuscrito de (Payne et al. 2019), quienes recuperaron una lectura de 2,272,580 bases. Indudablemente, ONT había ganado la carrera por la longitud a PacBio, que, desde entonces, se ha centrado en la obtención de librerías consistentemente largas (de 20 a 40 Kb) con mayor precisión en la llamada de variantes puntuales que ONT. Independientemente de las batallas entre superpotencias, en nuestro laboratorio teníamos aspiraciones más mundanas y nos conformábamos con obtener consistentemente lecturas lo suficientemente largas como para resolver inserciones de LINE1 en su longitud completa y las regiones adyacentes a las mismas, esto es, del orden de 10 Kb en su rango de tamaños inferior. Huelga decir que nuestros primeros intentos fueron bastante más discretos, con librerías con un tamaño medio en torno a las 3 Kb, lejos, muy lejos de la deseada cifra de 10 Kb.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

En nuestra gesta por obtener librerías con ADN de mayor tamaño empleamos dos enfoques complementarios y no excluyentes: mejorar los protocolos de extracción de ADN y filtrar fragmentos pequeños de las muestras de ADN previamente extraídas. La implementación rigurosa de prácticas óptimas en las extracciones de ADN resultó ser un factor crucial para mejorar la integridad del ADN final. Para lograr este objetivo, exploramos diversas técnicas y estrategias. Evitar la agitación mecánica, emplear puntas de pipeta con boca ancha para prevenir la cavitación y ruptura del ADN, y mantener en todo momento la cadena de frío de las muestras, incluso durante la extracción, fueron prácticas fundamentales que incorporamos en nuestro proceso. Además, suplementamos los buffers con dosis "*cuasi enfermizas*", en palabras de otros usuarios del CIMUS, de inhibidores de la actividad de ADNasas como el Beta-Mercaptoetanol. Estas prácticas mejoraron considerablemente la integridad de las muestras y, a día de hoy, el conocimiento adquirido en nuestros experimentos se continúa aplicando y enseñando a los nuevos usuarios en el laboratorio.

No veo práctico ni asumible escribir en estas líneas todos los protocolos y aparatos que se probaron en el lapso de un par de años para eliminar fragmentos de ADN de pequeño tamaño en muestras de ADN ya extraídas, ya sea por técnicas electroforéticas comerciales o manuales o por aproximaciones basadas en partículas magnéticas y/o polietilenglicol y centrifugación para el aislamiento/precipitación de ADN de alto peso molecular. Con cariño recuerdo el que quizás sea el error más grave (que no el único) durante esta parte de la tesis, tratando de obtener fragmentos largos e íntegros de ADN extrayéndolos de un gel de agarosa (ver apartado 6.1.1.7) al no haberme dado cuenta que los fragmentos de ADN que deseaba obtener estaban fuera del rango de extracción del kit empleado. Como se indicó en el apartado 5.1.10 el método adoptado finalmente y que todavía se aplica en el laboratorio para la selección de cadenas largas fue precipitación de ADN de alto peso molecular por centrifugación utilizando un buffer comercial con polietilenglicol de la casa *Circulomics*. Resulta gratificante ver que no estábamos descaminados al seguir nosotros también esa vía (datos no mostrados), aunque en retrospectiva probablemente no tuviésemos capacidad para obtener unos resultados tan buenos como los que se consigue con el kit comercial.

En paralelo a la familiarización con los protocolos de laboratorio (*wetlab)*, asumí funciones de informático del grupo de investigación para el manejo, procesamiento y control de los datos de secuenciación masiva generados en nuestros aparatos. Esto incluía instalar, actualizar y testear programas para el control de los aparatos (*MinKnow*), para el *basecalling (Guppy)* de los datos generados sin procesar (*raw data)* y su ulterior procesamiento para realizar los pertinentes análisis *a posteriori*, incluyendo mapeo, ordenación *(sorting)* e indexado. Destacaría primeramente el abultado número de versiones del software de control de los *MinION*, no siempre adecuadamente funcionales, cuyas dependencias requieren ahora mismo de *hardware* muy especializado y puntero que tuvimos que actualizar en repetidas ocasiones, sufriendo en el proceso la carestía de elementos informáticos en los últimos años. A resaltar también el abultado número y peso de los archivos que este tipo de secuenciación masiva genera y los múltiples problemas de almacenamiento con los que tuvimos que lidiar en el curso de nuestras investigaciones, lo que nos llevó a adquirir un sistema de almacenamiento en cintas y múltiples discos duros en una NAS, así como a constantes llamadas y negociaciones con el personal del CESGA para obtener espacio de almacenamiento remoto. Además, uno de los problemas iniciales que enfrentamos fue que las

versiones más primitivas del software generaban archivos de datos sin procesar (*raw data*, *.fast5*) y *.fastq* monosecuenciales, lo que significaba que al final de un experimento de secuenciación masiva, se generaban tantos archivos como lecturas, en ocasiones alcanzando cifras de varios millones. Esta situación supuso un desafío logístico considerable, y en un intento desesperado de liberar espacio en disco para continuar con la secuenciación de otras muestras. Afortunadamente, versiones posteriores del *MinKnow* solucionaron este defecto, facilitando significativamente el manejo de los datos.

Como resultado de la casuística que enfrentamos, el grupo de investigación se vio en la necesidad de desarrollar instrumentos de laboratorio a medida para la elaboración de librerías de secuenciación masiva. Uno de estos instrumentos fue un mezclador rotativo (ver apartado 6.1.2) que construí personalmente, ya que las alternativas disponibles giraban a velocidades demasiado altas para nuestras necesidades específicas. En la fase de experimentación para obtener lecturas largas también construí una pequeña cubeta de electroforesis (6.1.3), que a la postre no llegamos a utilizar en los protocolos estandarizados. Asimismo, desarrollamos un soporte magnético para las purificaciones de ADN utilizando esferas magnéticas, lo que mejoró significativamente la eficiencia de este proceso. Así mismo fue necesario el diseño y construcción de una cámara refrigeradora (6.1.4). La ubicación de nuestro laboratorio y la política de no encender el aire acondicionado fuera del horario laboral significaba que la temperatura dependía exclusivamente de las condiciones atmosféricas. Durante épocas estivales u olas de calor, esto tenía un impacto considerable en la calidad de la secuenciación, ya que los *MinION* por sí mismos carecían de un sistema de disipación de calor suficientemente efectivo que permitiera mitigar este efecto. A temperaturas elevadas, la velocidad de translocación de las secuencias de ADN se incrementaba, lo que a su vez reducía la calidad de la secuencia obtenida. Esto se traducía en un mayor número de lecturas que no alcanzaban los mínimos de calidad requeridos y en una menor fiabilidad y calidad de mapeo en etapas posteriores del análisis. Por tanto, era fundamental contar con un sistema de refrigeración donde pudiéramos introducir los aparatos de secuenciación para asegurar un rendimiento óptimo. Desafortunadamente cuando la cámara estaba lista, las políticas de uso del aire acondicionado en los laboratorios cambiaron, haciendo innecesario su uso. Al mantener un ambiente controlado y adecuadamente refrigerado, pudimos mitigar el impacto negativo de las temperaturas elevadas en los resultados de la secuenciación. Esto no solo ha optimizado la calidad de nuestras librerías de secuenciación masiva, sino que también ha permitido una mayor estabilidad y consistencia en nuestros análisis.

En general, el legado que esta tesis deja al grupo es una *facility* de secuenciación masiva de lecturas largas con un equipo humano plenamente entrenado, unos protocolos de laboratorio completamente fiables y reproducibles, una infraestructura de hardware y software funcional, simple de mantener y actualizar en el tiempo. El esfuerzo realizado en esta sección de mi tesis doctoral demuestra el compromiso que el grupo tuvo con la excelencia científica y la búsqueda constante de soluciones innovadoras a la par de creativas. A través de la colaboración interdisciplinaria y el enfoque crítico, hemos superado desafíos y alcanzados resultados indispensables para nuestras investigaciones y que, además, nutren a otros investigadores de dentro y fuera de nuestro grupo.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

9.2 METHYL-MEIGA ES UN SOFTWARE ESPECÍFICO PARA LA DETECCIÓN DE LA METILACIÓN DE LINE1 SOMÁTICOS Y GERMINALES.

En esta sección de discusión general, ahondaré en las implicaciones y relevancia de las *pipelines* que he desarrollado.

A pesar de la existencia de herramientas específicas para el *basecalling* de metilación a partir de datos de secuenciación de ONT, no había ninguna dedicada exclusivamente al análisis de los patrones de metilación de los LINE1, que, debido a su naturaleza repetitiva, requirieron de un tratamiento especial y el desarrollo de herramientas propias. Para abordar esta necesidad, desarrollé el programa Methyl-Meiga, cuyos detalles se describen en el apartado 6.2.2. Esta herramienta se basa en los datos de metilación proporcionados por Megalodon (ONT) lo que nos apremió a realizar ajustes continuos para mantener nuestra herramienta actualizada y en sintonía con sus frecuentes (y a veces, disruptivas) actualizaciones.

Habitualmente, la evaluación exhaustiva de un nuevo programa es una práctica común para determinar su idoneidad, especificidad y sensibilidad en comparación con otras herramientas con la misma función. En todo caso, y hasta donde alcanza nuestro conocimiento no existen herramientas concretas con las que podamos comparar nuestra *pipeline.* En algunas investigaciones similares i.e. (Tubio et al. 2014; Sarkar, Lanciano, y Cristofari 2022; Lanciano et al. 2023) se recurrió al análisis *ad hoc* de los datos de metilación de los retrotransposones (incluidos los LINE1) con *scripts* diseñados específicamente para esos análisis y no con una herramienta específica como Methyl-Meiga. En algunos de los artículos mencionados, además, se limitan a indicar las directrices seguidas en el análisis de estos elementos, sin que el código utilizado para los análisis esté disponible públicamente. Hemos tenido en cuenta esta información a la hora de considerar y adoptar las estrategias más adecuadas en el desarrollo de Methyl-Meiga. Además, se han realizado las pertinentes comprobaciones de funcionamiento y validez de los resultados en base a la coherencia interna de los mismo y la concordancia con estudios previos de la metilación en retrotransposones, así como la revisión visual de los datos (con el *integrative genoma browser IGV*) y de las propias gráficas generadas por Methyl-Meiga. En adición a Methyl-Meiga se desarrolló un conjunto de pequeños y medianos *notebooks* de Python (mencionados en los apartados correspondientes) para analizar y extraer información relevante de los datos de metilación de los LINE1 y abordar las preguntas biológicas planteadas en esta investigación. Esta parte del trabajo, en términos de codificación, representó la faceta más innovadora, ya que, si bien existían desarrollos aplicados en áreas de conocimiento similares, estos se basaban en diferentes tipos de datos y se implementaban en otros lenguajes de programación. En el caso particular de estudiar la metilación de los LINE1 en cáncer utilizando datos de lecturas de secuenciación de molécula única (ONT), esta combinación era única e inexplorada al inicio de la tesis. Pese a que a lo largo de los últimos años otros grupos de investigación, como los de G. Faulkner o el de G. Cristofari han desarrollado sus propias investigaciones en temas similares (Ewing et al. 2020; Sarkar, Lanciano, y Cristofari 2022; Lanciano et al. 2023), hasta donde alcanza nuestro conocimiento, su trabajo no está enfocado de forma tan explícita a la metilación de LINE1 en tumores.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

El software que he diseñado, Metil-Meiga, se ha demostrado adecuado y altamente útil para determinar la metilación en CpGs de elementos germinales e inserciones somáticas de LINE1. Como mencioné anteriormente, el estudio de inserciones somáticas de transposones con lecturas largas es aún un campo de investigación en desarrollo, con una escasez de programas específicos para su estudio a partir de lecturas largas. En nuestro grupo de investigación, estamos inmersos en el desarrollo de un programa específico para la búsqueda de inserciones somáticas de transposones y sus reorganizaciones, conocido como MEIGA-LR (pendiente de publicación), en cuyo desarrollo he tenido la oportunidad de participar. Esto nos ha permitido obtener una lista precisa de las inserciones somáticas de LINE1 en nuestra cohorte de tumores, junto con la secuencia específica de cada una de ellas y los puntos de inserción en el genoma de referencia.

Es importante destacar que las inserciones somáticas de LINE1, al igual que los LINE1 germinales ausentes en el genoma de referencia, no son detectadas por los alineadores tradicionales (por ejemplo, Minimap2) y de los *basecallers* de metilación (por ejemplo, Megalodon), que habitualmente detectan únicamente metilación en CpGs presentes en el genoma de referencia. Esto presentó un desafío significativo, y fue uno de los motivos principales que me impulsaron a desarrollar Metil-Meiga. Para esta Tesis se usaron datos de los LINE germinales íntegros tanto del genoma de referencia (Hg38) como otros de conocida actividad, además de los datos generados por MEIGA-LR, lo incluye las secuencias consenso de las inserciones somáticas. Con estos datos Methyl-meiga genera una suerte de nuevo genoma de referencia específico de cada muestra. A continuación, el software (usando internamente Megalodon) recupera las lecturas de ONT de cada muestra, las remapea sobre sus correspondientes referencias en el nuevo genoma de referencia y extrae el estado de metilación promedio de cada citosina, así como estadísticas de las regiones corriente arriba y abajo (*up* y *downstream*) y de su promotor, cuerpo y elemento completo. La salida *(output)* generada es tabulada y proporcionada a varios libros (*notebooks*) de Python que realizan las pertinentes tablas y gráficos. Gracias a este programa, pudimos analizar con precisión los niveles de metilación de la secuencia completa de LINE1 germinales y de inserciones somáticas específicas de muestra proporcionadas por MEIGA-LR en toda la cohorte de datos analizada, ya fueran de muestras tumorales como de tejido adyacente.

Gracias a Metil-Meiga, hemos obtenido resultados detallados y fiables sobre la metilación en CpGs de las inserciones somáticas de LINE1, lo que ha enriquecido nuestro conocimiento sobre la regulación epigenética en estos elementos móviles en el contexto del cáncer. La precisión y eficiencia de este software son valiosas herramientas para el estudio de la dinámica de metilación en retrotransposones y su posible implicación en la progresión tumoral y otros procesos biológicos. En resumen, esta herramienta y los módulos adjuntos a la misma proveen al usuario (y a nuestro grupo) de una herramienta versátil y relativamente manejable que puedan emplear para el análisis de los niveles de metilación de LINE1 en muestras secuenciadas mediante ONT.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

9.3 RESULTADOS BIOLÓGICOS

En el marco de esta tesis doctoral, se llevó a cabo un minucioso análisis de la metilación de un total de 23 muestras de tejidos tumorales con alta tasa de retrotransposición, provenientes de cuatro tipos de cáncer: cabeza y cuello, esófago, pulmón y colon, junto con sus respectivos tejidos adyacentes (12 de ellos) y de dos líneas celulares de carcinoma de pulmón. Parte de ese análisis se enfocó en el procesamiento de esta cohorte de datos con el objetivo de identificar patrones de metilación tanto generales como específicos de los retrotransposones. Cabe destacar que este estudio fue realizado en paralelo con un análisis exhaustivo para determinar los patrones de movilización de retrotransposones en parte de las muestras, realizados por nuestro grupo de investigación, de cuyos resultados en gran medida se nutre esta tesis doctoral.

Quizá una de las características diferenciadoras de este trabajo de investigación en el análisis de la metilación de LINE1 en tumores primarios frente a trabajos previos es que éstos no contaban con una cohorte de datos tan adecuada como la que se ha empleado en el presente estudio. Si bien no excesivamente abultada en cuanto a número, ya que está formada únicamente por 23 tumores primarios, algunos con sus correspondientes tejidos adyacentes, y dos líneas celulares (consultar apartado 7.1), todas ellas muestran tasas de retrotransposición excepcionalmente elevadas, como resultado de una cuidadosa selección de una colección de más de 150 muestras tumorales (consultar apartado 7.1.1). A destacar entre ellas casos notables como la PD0270a y la PD277a, que contienen 1239 y 691 inserciones somáticas de LINE1, respectivamente. Asimismo, las líneas celulares NCI-H2087 y NCI-H2009 son igualmente excepcionales, en las cuales encontramos 479 y 443 inserciones somáticas, respectivamente, y han sido empleadas en investigaciones previas relacionadas con la retrotransposición (Tubio et al. 2014; Rodriguez-Martin et al. 2020). Estudios previos se limitaban a un número considerablemente menor de muestras, ya fuesen líneas celulares (Tubio et al. 2014; Lanciano et al. 2023) o unas pocas muestras de tumores primarios, como por ejemplo, un tumor de hígado en el reciente artículo de (Ewing et al. 2020). Otros estudios con cohortes de muestras más extensivas i.e. (Rodriguez-Martin et al. 2020), basados en datos de secuenciación de lecturas cortas (NGS), se centraban más en la descripción de patrones generales de transposición entre tipos tumorales que, si bien son tratados importantes de cara a analizar tendencias descriptivas y generalistas de transposición somática en cáncer, no ahondaban ni en los procesos que las controlan ni en los mecanismos epigenéticos que las regulan.

La hipometilación global del genoma (Vidal et al. 2017; A. Feinberg 2014) y, de manera específica, de los LINE1, son rasgos característicos que se observan en muchos tipos de tumores. De hecho, estudios previos han sugerido que la metilación en los elementos LINE1 podría servir como un indicador relevante de la progresión tumoral, y su potencial como biomarcador para el diagnóstico y pronóstico de la enfermedad ha sido destacado en (Shen et al. 2018; Barchitta et al. 2014). Los resultados derivados del análisis de los datos generales de la metilación de las CpG en nuestra cohorte (ver apartado 8.1.1) concuerdan plenamente con estas afirmaciones, mostrando una disminución del 10% en la metilación en los tumores en comparación con los tejidos normales. No obstante, se destaca una notable excepción en el caso de las líneas celulares, donde sorprendentemente la metilación es menor

USC DE COMPOSTELA

en el tejido de comparación, que sería equivalente al tejido adyacente en los tumores primarios. Es válido argumentar que se trata de tejidos distintos; mientras unos son tumores de pulmón, los otros son linfocitos, lo que podría implicar diferencias en la metilación de estos tejidos. Sin embargo, es importante considerar que la principal disparidad podría derivarse del uso del tejido de comparación proveniente de una línea celular de linfocitos inmortalizados (ver apartado 8.1.1.1). Estas observaciones subrayan la relevancia de la metilación de los elementos LINE1 en la progresión tumoral y enfatizan la importancia de evaluar cuidadosamente los tejidos y líneas celulares utilizados como comparación en los estudios epigenéticos. Los resultados obtenidos en esta investigación contribuyen al conocimiento en el campo de la epigenética y respaldan la utilidad potencial de la metilación de los elementos LINE1 como un biomarcador valioso en el diagnóstico y seguimiento de enfermedades relacionadas con el cáncer.

Al analizar la metilación en tumores primarios mediante secuenciación masiva (*bulk sequencing*), como en el caso presente, es esencial considerar la heterogeneidad intratumoral y la presencia de células no tumorales, ya que estos son problemas importantes que pueden afectar la interpretación de los resultados. Los bloques de tejido utilizados para extraer el ADN de los tumores generalmente contienen una mezcla de células tumorales, tejido adyacente no tumoral, células del estroma y del sistema inmunitario, lo que añade diversidad a la muestra y complica el análisis de metilación. Además, los tumores presentan una alta tasa de mutación, lo que resulta en la coexistencia de múltiples clones dentro de un solo tumor (Stratton, Campbell, y Futreal 2009; Nik-Zainal et al. 2012; 2016). Estos clones pueden exhibir diferentes patrones de metilación, lo que aumenta la complejidad del estudio y puede enmascarar patrones de metilación específicos relacionados con procesos biológicos relevantes. Para abordar estos desafíos, se requieren métodos de análisis que permitan distinguir y separar las señales de metilación específicas de las células tumorales de las demás células presentes en la muestra. En ocasiones, es necesario emplear enfoques más avanzados y sofisticados, como la secuenciación a nivel de célula única (Eberwine et al. 2014; Navin 2015; Wang y Navin 2015) o el uso de técnicas multirregión (Gerlinger et al. 2012) o de purificación de células tumorales, para obtener resultados más precisos y significativos en el estudio de la metilación en tumores. Todos estos inconvenientes de los tumores primarios se trasladan a los datos, donde parte de la información corresponde a tipos celulares distintos al del tumor, dificultando sobremanera los análisis (ver apartado 8.2.3). Es por ello por lo que resulta indispensable caracterizar la pureza y clonalidad de los tumores primarios, siendo el "Gold standard" el marcado por el programa Battemberg.

En el análisis de inserciones somáticas de LINE1 mediante secuenciación masiva, no se considera un problema la presencia de heterogeneidad intratumoral o células no tumorales, ya que estos eventos son específicos del tumor y no están presentes en la muestra de partida. Por lo tanto, es factible distinguir entre lecturas tumorales y no tumorales con relativa facilidad, especialmente empleando lecturas largas que permiten abarcar la totalidad de la inserción. No obstante, en la determinación de los niveles de metilación de los elementos fuente, que son virtualmente idénticos entre tejidos normales y tumorales, se requiere un análisis más minucioso. Específicamente, se necesitaría disponer de variantes somáticas puntuales que permitan discriminar entre tejido normal y tejido tumoral, y así agrupar las lecturas largas en base a haplotipos tumorales y normales. Aunque teóricamente realizable,

este proceso de análisis fino es tedioso y debería realizarse de manera específica para cada elemento y muestra, lo que dificulta su aplicación generalizada sin que además haya garantías de éxito, ya que depende de que tales SNVs discriminantes existan en primer lugar. Por lo tanto, a efectos de la presente tesis doctoral hemos decidido no aplicar este enfoque.

En vista de las notables diferencias en la metilación de los promotores de LINE1 entre los tumores y los tejidos adyacentes, surgió la posibilidad de utilizar esta información para inferir la pureza de las muestras tumorales, como se detalla en el apartado 8.2.2. Los resultados respaldan la idea de que el estado de metilación en la región del promotor de LINE1 puede ser un indicador bastante confiable para estimar la pureza de las muestras tumorales, incluso en aquellos casos de tumores que cursen con fenotipos anormalmente hipermetilados, como es el caso de la muestra de tumor colorrectal PD0351 analizada en este trabajo. Sin embargo, es importante destacar que se requieren investigaciones adicionales para realizar correcciones adecuadas en la estimación de los niveles de contaminación utilizando la metilación de ONT, con el objetivo de acercarse a los estándares establecidos por Battemberg u otros métodos de referencia. Estos avances potenciales podrían mejorar aún más la precisión de las estimaciones de pureza y fortalecer la utilidad de la metilación de LINE1 como una herramienta valiosa en la caracterización de la heterogeneidad tumoral.

Después de evaluar la posibilidad de utilizar la metilación de los LINE1 para inferir la pureza de los tumores, se comenzó a explorar la idea de aplicar los mismos principios para abordar el problema de la contaminación de tipos celulares normales. La propuesta consistía en agrupar las lecturas del tumor y del tejido adyacente mediante algoritmos de aprendizaje no supervisado, como *neighbour joining* o *hierarchical clustering*, y posteriormente descartar las lecturas del tumor que se agruparan con las del tejido adyacente. Sin embargo, tras los primeros intentos, se constató que esta tarea requeriría un desarrollo que resultaba imposible de llevar a cabo dentro del marco temporal de esta Tesis. A pesar de ello, esta vía sigue siendo sumamente interesante y prometedora para mejorar la exactitud de los análisis que involucran la contaminación del tumor con tejido adyacente o infiltración de células del sistema inmunitario. La implementación exitosa de esta estrategia podría permitir una mejor distinción entre las señales genuinas del tumor y las que provienen de la contaminación, mejorando así la interpretación de los resultados y aumentando la fiabilidad de los hallazgos.

Existe una gran cantidad de estudios a cerca de la metilación de los LINE1 germinales (algunos de ellos mencionados en el apartado 3.4.4) pero hasta fechas muy recientes, y debido a las limitaciones de la tecnología, eran estimaciones “genéricas” en el sentido de que analizaban todos los LINE1 en su conjunto sin poder distinguirlos individualmente, y en la mayor parte de los casos promediando los niveles de metilación de todo el LINE1. Recientes estudios como el de (Lanciano et al. 2023) hacen uso de una novedosa tecnología que les permite estudiar la metilación de las primeras CpG de gran parte los LINE1, lo que les permite evaluar el estado epigenético de sus promotores. El avance tecnológico en secuenciación de molécula única, como la tecnología de ONT y PacBio, ha abierto una nueva era en el estudio de la metilación de los LINE1, permitiendo analizar cada uno de estos elementos individualmente con una resolución a nivel de nucleótido a lo largo de toda su secuencia sin los problemas de mapeabilidad que muestran las técnicas de secuenciación de lectura corta. Esta capacidad nos ha permitido estudiar los epialelos (metil-haplotipos) de

todas las muestras tumorales, como se detalla en el apartado 8.2.3. Estos hallazgos y los del apartado 8.2.4 confirman los patrones previamente establecidos en la literatura (ver apartado 3.4.4), que indican que la metilación de los LINE1 es menor en los tumores en comparación con el tejido normal del cual proceden. Por consiguiente, esta nueva tecnología ha proporcionado una perspectiva sin precedentes en la metilación de los LINE1, lo que nos ha permitido identificar de manera más precisa y detallada los patrones epigenéticos asociados con estos elementos en el contexto del cáncer.

Respecto a los LINE1 germinales, íntegros analizados en este estudio, en el tejido adyacente predomina una mayoría de elementos altamente metilados (homocigoto metilado 87%, heterocigoto 11%, homocigoto desmetilado 1%) frente a los tumores donde se produce una pérdida de metilación en más de la mitad de los elementos (Homocigoto desmetilado 54%, 15% heterocigoto, homocigoto metilado 31%). Esto junto con los resultados mostrados en el apartado 8.2.4 muestran claramente un patrón de desmetilación de los LINE1 concordante con la literatura. No obstante, un descubrimiento, probablemente el más significativo a nivel biológico del presente trabajo, **es el claro contraste entre la metilación de los promotores de los LINE1 (los primeros 900 pares de bases) con los cuerpos de los mismos** (discutido en el apartado 8.2.6). Siendo la metilación de los promotores mucho más baja en el tumor que en el tejido normal. Mientras que el cuerpo presenta un patrón de metilación similar al de los loci corriente arriba y corriente abajo adyacente a los LINE1, tanto en los tumores como en el tejido adyacente. Estas regiones suelen estar altamente metiladas. Este efecto ha permanecido oculto hasta el momento pues los estudios de metilación de los LINE1, como se ha comentado anteriormente, consideraban ambas zonas en conjunto, o solo las primeras CpG del promotor. Como el número de CpG es comparable entre el promotor y el cuerpo, a pesar de ser este 5 veces más grande, la variación de metilación era significativa incluso siendo los valores una media entre las dos regiones.

La observación anterior del contraste en la metilación entre los promotores y el cuerpo de los LINE1 no se limita únicamente a los elementos germinales, sino que también se ha observado en inserciones somáticas íntegras (apartado 8.4, y 8.4.2), algunas de las cuales conocemos el elemento fuente del que derivan (apartado 8.5). En estas copias, nuevamente se destaca una marcada diferencia en la metilación entre las regiones, mostrando los promotores una metilación significativamente más baja en comparación con el cuerpo y las regiones adyacentes. Este patrón diferencial de metilación entre los promotores y el cuerpo del LINE1 tanto en los LINE1 germinales como en las inserciones, ha sido un hallazgo revelador y sorprendente que, sin lecturas largas, no se hubiese detectado. Hasta ahora, la mayoría de los estudios de metilación en los LINE1 han considerado ambas regiones en conjunto o se han centrado únicamente en las primeras CpG del promotor. Como resultado, este efecto distintivo de metilación ha permanecido oculto, lo que destaca la importancia de abordar un análisis más detallado y completo de la metilación en estas regiones.

Desconocemos por completo el mecanismo preciso que favorece la pérdida y/o el mantenimiento del estado de metilación de los promotores de las nuevas copias de LINE1 en los tumores. Dado que la retrotranscriptasa sintetiza e inserta el ADN complementario desmetilado durante la retrotransposición, se esperaría que las enzimas metilasas *de novo* (como las DNMT3s) (Gujar, Weisenberger, y Liang 2019) metilen todas las CpGs de este

nuevo ADN. Sin embargo, en la mayoría de los casos, las CpGs del promotor de los LINE1 (en tumores) permanecen desmetiladas o son desmetiladas posteriormente, lo que representa un enigma para la regulación de la metilación en estas nuevas copias. Es importante señalar que los promotores de los LINE1 están ampliamente metilados en el tejido sano y, en algunos casos, también en los tumores, lo que sugiere que las metilasas tienen la capacidad de metilar los promotores de los LINE1. Algunos autores han planteado la hipótesis de que las células tumorales humanas, al estar altamente diferenciadas, podrían carecer de la actividad metiltransferasa *de novo* necesaria (DNMT3s) para llevar a cabo la metilación de los LINE1 recién insertados, lo que explicaría su aparente falta de implicación en el silenciamiento de los nuevos insertos (Garcia-Perez et al. 2010). Sin embargo, nuestros datos contradicen esta idea, ya que hemos observado que una parte de los LINE1 sí está metilada, incluyendo los promotores de algunos elementos fuente e inserciones somáticas. Esta observación plantea la necesidad de realizar experimentos biológicos adicionales para dilucidar qué mecanismos están involucrados en estos procesos. Las tecnologías de secuenciación de molécula única, así como los métodos desarrollados en esta Tesis, podrían ser fundamentales para analizar los datos resultantes de esos experimentos y proporcionar una mayor comprensión de los mecanismos de regulación de la metilación en los LINE1 en el contexto del cáncer.

El análisis exhaustivo de todos los LINE1 germinales íntegros, con la capacidad teórica de movilizarse, nos permitió explorar la posible correlación entre su estado de metilación y su actividad. Los resultados obtenidos revelaron una ausencia de correlación (o en todo caso una correlación muy débil) entre ambas variables (ver apartado 8.3). Aunque es cierto que los elementos con mayor actividad tienden a estar fuertemente desmetilados, también observamos que la mayoría de los elementos desmetilados no presentan actividad. Esta discrepancia con respecto a los hallazgos de (Tubio et al. 2014), quienes encontraron una correlación significativa entre el metil-haplotipo más frecuente y la actividad, puede deberse a las diferencias en la metodología utilizada. En el estudio previo, se analizaron específicamente las primeras CpG de algunos elementos conocidos por su actividad, sin considerar la gran mayoría de LINE1 que poseen potencial capacidad de movilización y también exhiben niveles de metilación muy bajos. Además, es importante destacar que en ese trabajo se emplearon líneas celulares como modelo de estudio, mientras que en nuestra investigación hemos demostrado que las líneas celulares pueden no ser el sistema más adecuado para el estudio de la metilación (ver apartados 8.1.1.1 y 8.2.3.1).

Es importante destacar que, si bien hemos realizado un análisis exhaustivo de los LINE1 germinales y su correlación con la metilación, el estimador utilizado para establecer el nivel de actividad de estos elementos podría no es el más adecuado en determinados contextos. Uno de los principales inconvenientes radica en que este estimador no refleja la actividad en el momento presente, lo que significa que algunos elementos podrían haber experimentado un pico de actividad en el pasado, pero actualmente no están activos. Como resultado, los niveles de metilación actuales podrían no coincidir con los niveles que tenían en el momento de su actividad, lo que dificulta una interpretación directa de la correlación entre el nivel de metilación de los promotores en el momento presente y el número de copias hija asignables a cada elemento germinal. Además, durante nuestros análisis, no hemos considerado otro factor importante que puede influir en la aparente actividad de los elementos germinales. Algunos de estos elementos fuente pueden estar sujetos a amplificación focal o

ganancia neta de copias debido a la amplificación de brazo cromosómico o incluso el cromosoma completo debido a variaciones estructurales del cariotipo. Esta amplificación de copias puede sesgar la aparente sobre activación de ciertos elementos, ya que en realidad podrían tener niveles de activación basal similares a aquellos que existen en copia única. Sin embargo, debido a la acumulación sumatoria de múltiples copias, pueden aparentar una mayor actividad. Este sesgo puede afectar la interpretación de los resultados y requerir un análisis más detallado y específico para desentrañar la verdadera relación entre la metilación y la actividad de estos elementos. Otros estimadores de actividad, discutiblemente más válidos, como pueden ser el número de transcritos de ARN correspondientes a los LINE1, también han demostrado correlación nula con respecto a los niveles de metilación de los promotores de LINE1 (Lanciano et al. 2023). Ello ha llevado a que estos investigadores postulasen que deben existir otros mecanismos de regulación, por ejemplo, el estado de condensación de la cromatina, marcas epigenéticas en las histonas, o ARN interferentes cuyo análisis sea posiblemente abordado por nuestro grupo en un futuro próximo.

La pregunta 4.6 sobre la heredabilidad de los patrones de metilación de las copias hijas, abordada en el apartado 8.5 y discutida parcialmente en 8.4.2, está estrechamente relacionada con la perdida y/o mantenimiento de la metilación descrita unos párrafos más arriba. La noción de "heredabilidad" en este contexto, como la transmisión de información epigenética de una entidad parental a sus descendientes, puede resultar poco apropiada. Durante la retrotranscripción, toda la información de metilación de la copia original se borra, lo que implica que cualquier modificación epigenética en las nuevas copias debe surgir posteriormente a este evento. La posibilidad de que las copias hijas mantengan el patrón de metilación de la copia fuente implicaría la transmisión de esa información a través de un mecanismo diferente al ARN que se retrotranscribe, lo cual parece poco probable. Si hubiera considerado este razonamiento antes de plantear la pregunta, podría haber evitado invertir esfuerzo en una cuestión cuya respuesta era poco factible. No obstante, esta reflexión nos ha enseñado una valiosa lección: "***Para obtener buenas respuestas hace falta partir de buenas preguntas***". Con todo, los resultados inesperados pueden abrir puertas hacia direcciones de investigación aún más interesantes que las inicialmente previstas y, desde mi humilde opinión, esta observación merecería ser estudiada minuciosamente en el futuro.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

9.4 FUTUROS DESARROLLOS, NUEVAS PREGUNTAS, POSTERIORES VÍAS DE INVESTIGACIÓN.

La investigación científica es un viaje lleno de descubrimientos y desafíos, y este trabajo no ha sido una excepción. A medida que se resolvían las preguntas iniciales, se abrían nuevos interrogantes y oportunidades que seguir explorando. Mas, desafortunadamente, el tiempo es limitado. Algunas de estas cuestiones podrían haber sido abordadas con los datos y métodos disponibles, como por ejemplo la metilación de las regiones adyacentes del tejido no tumoral en los puntos de inserción de los LINE1, lo que aportaría información acerca de la preferencia en regiones metiladas o no de las inserciones de LINE1. O por ejemplo el análisis de la metilación los distintos alelos, puesto que Methyl-Meiga incorpora esa información en caso de existir.

Otras cuestiones requieren de un enfoque distinto y tecnología innovadora: Como dijo Albert Einstein, “No podemos resolver problemas pensando de la misma manera que cuando los creamos”. A continuación, procedo a enumerar alguna de estas cuestiones que dejo pendientes, así como los posibles enfoques bajo los que se podrían abordar.

Todo método y procedimiento, ya sea experimental o informático, es susceptible a errores. Por ello, es crucial respaldar los resultados con comprobaciones ortogonales, pruebas biológicas y moleculares, y contrastar los análisis con otros programas y metodologías, como Nanopolish y DeepSignal. Además, el avance constante en tecnología también abre nuevas perspectivas y, tal como se indica en la hipótesis de la Reina Roja, la supervivencia de un código informático se asemeja a una carrera evolutiva en la que ha de actualizarse y adaptarse constantemente al desarrollo de las nuevas tecnologías y programas so pena de quedarse obsoleto en el caso de que no lo hiciera. Por ejemplo, la aparición de REMORA, un nuevo *basecaller* de metilación desarrollado por ONT, podría significar mejoras significativas en futuros análisis.

Explorar el estado de metilación de todos los LINE1 (entre quinientas y ochocientas mil copias) y otros tipos de retrotransposones, como Alus, SVAs, retrovirus y pseudogenes, representa un camino prometedor para ampliar nuestra comprensión de la regulación epigenética y la dinámica del genoma. Estos elementos también tienen una amplia distribución en el genoma humano y se sabe que desempeñan funciones cruciales en diversos procesos biológicos, con lo que conociendo qué mecanismos epigenéticos modulan su actividad podríamos tener un mejor entendimiento de su posible implicación en el desarrollo de enfermedades y en la regulación de procesos fisiológicos clave. En particular, los Alus, muy abundantes en el genoma humano, han sido objeto de interés en la investigación genómica debido a su aparente vinculación a procesos como la diferenciación celular y la respuesta inmunitaria. Analizar la metilación de estos elementos en distintos tipos celulares y en diferentes condiciones patológicas podría proporcionar información valiosa sobre su impacto en la función celular y su posible papel en enfermedades.

Es esencial que los estudios epigenéticos se enfoquen holísticamente, abarcando múltiples dimensiones de la biología celular y no limitándose a una fracción mínima (aunque relevante) como la metilación. En este sentido, en un futuro sería sumamente enriquecedor complementar nuestros análisis de metilación y actividad de LINE1 en estas u otras muestras

USC DE COMPOSTELA

con diversos análisis adicionales. Por ejemplo, podríamos investigar la accesibilidad de la cromatina, lo que nos permitiría comprender cómo la estructura de la cromatina afecta la metilación y actividad de los retrotransposones. También podríamos realizar precipitaciones y análisis de proteínas específicamente ligadas a las secuencias de LINE1, para identificar interacciones clave que puedan estar involucradas en su regulación epigenética. Además, la realización de estudios transcriptómicos nos proporcionaría una visión integral de cómo la metilación y actividad de LINE1 afectan la expresión génica global en estas muestras. Podríamos explorar las rutas metabólicas y las vías de señalización que son afectadas por la presencia de LINE1 activos o desmetilados, y comprender mejor su impacto en la función celular y en la respuesta a estímulos externos. Otro aspecto relevante sería el estudio del ARN de interferencia (ARNi) relacionado con los retrotransposones. Sabemos que ciertos ARN pequeños pueden silenciar la expresión de transposones y proteger así la integridad del genoma. Analizar la presencia y actividad de estos ARNi específicos en nuestras muestras podría proporcionar pistas valiosas sobre los mecanismos de regulación de LINE1 y su relevancia en la estabilidad genómica. En última instancia, la integración de todos estos análisis multiómicos nos permitiría obtener una imagen más completa y profunda de la regulación epigenética de los retrotransposones y su impacto en la biología celular.

Un área de investigación que me gustaría explorar en el futuro es el mecanismo molecular que subyace a la hipometilación específica de los promotores de LINE1 en comparación con las regiones adyacentes y el cuerpo del elemento, incluso en las nuevas copias somáticas. Aunque carezco de resultados concretos, tengo la impresión de que este fenómeno podría estar regulado por un bloqueo físico de la acción de las metilasas *de novo*. Es posible que algún complejo proteico específico se una a los promotores de LINE1, impidiendo que las metilasas *de novo* accedan y actúen en estas regiones. Esta hipótesis abre interesantes vías de investigación para comprender cómo se regula la metilación de LINE1 y qué factores proteicos podrían estar involucrados en esta dinámica epigenética.

Otro área muy interesante para futuras investigaciones sería la posibilidad de aplicar estos estudios a la práctica clínica, puesto que las tecnologías y metodologías aquí esbozadas podrían en un futuro tener aplicaciones prácticas ya que permiten estudiar de forma rápida, sencilla y relativamente económica el estado de metilación del ADN y los transposones de muestras de pacientes. Esto permitiría un mejor diagnóstico de las enfermedades y el seguimiento más preciso de los diversos tratamientos.

En resumen, este trabajo, más que la consecución y el legado final de mi estadía en este grupo de investigación, representa un punto de partida para nuevas investigaciones y desarrollos en el campo de la metilación y retrotransposones.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10. CONCLUSIONES

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Imagen portada del capítulo: “Victoria de Samotracia”, Louvre. Lápiz y rotulador sobre libreta (feb 2023)

1. Se ha conseguido la puesta en marcha de una unidad de secuenciación de lecturas largas de Oxford Nanopore Technologies que nos ha permitido contribuir al ensamblaje de un genoma de referencia para un proyecto ERC, a la validación de transposones para un artículo de PCAWG, a varios proyectos de nuestro grupo de investigación y a la generación de los datos para esta Tesis.
2. Se ha desarrollado un programa informático: Methyl-Meiga y multitud de *scripts* (*notebooks* de Python) para el estudio y análisis de los patrones de metilación general de tumores y de los retrotransposones LINE1.
3. Se ha aplicado el método anterior al estudio de la metilación los LINE1 en tumores y tejidos adyacentes en más de veintitrés tumores y dos líneas celulares, todos ellos con tasas de retrotransposición muy altas.
4. Se ha comprobado que los tumores de nuestra cohorte siguen el patrón ya descrito en la literatura de una tendencia general hacia la desmetilación.
5. Se ha constatado que las líneas celulares NCI-H2009 y NCI-H2087 (carcinomas de pulmón) no son adecuadas para el estudio de la metilación de los LINE1 a pesar de la alta tasa de retrotransposición que presentan.
6. No se ha encontrado correlación entre la actividad y los patrones de metilación en los LINE1 en nuestra cohorte de tumores.
7. Se ha descubierto que el promotor y el cuerpo de los LINE1 presentan patrones de metilación diferentes, sobre todo en los tumores, siendo la metilación del cuerpo muy similar a la de las regiones adyacentes, mientras que el promotor muestra una clara tendencia a estar desmetilado, independientemente del estado de metilación de las regiones adyacentes.
8. Se ha comprobado la validez de esta tecnología para el estudio de la metilación del ADN en general y de los LINE1 en particular.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

11. GLOSARIO Y ABREVIATURAS

11.1 GLOSARIO

Amplicones*:* Fragmentos de ADN amplificados mediante PCR.

Contig: Secuencia resultado de concatenar *in silico* datos de secuenciación de fragmentos de ADN o ARN.

Core facility, facility: Recursos de investigación centralizados, que permiten acceso a tecnologías, servicios, instrumentos, experiencia y conocimiento de expertos. En nuestro caso una unidad de secuenciación de lecturas largas.

Epialelo: Alelo distinguible, o segregable por la información epigenética.

Exonización: Proceso por el cual fragmentos de la secuencia de los transposones son incorporados en ARN mensajeros de genes.

Flowcell: Celda de flujo, en este trabajo se refiere en concreto a las celdas desechables que contiene los nanoporos y sensores necesarios para realizar la secuenciación en los dispositivos de ONT.

.fasta: Ficheros de texto con información de secuenciación genómica (nucleótidos).

.fast5: Ficheros HDF5 con información sin procesar de la señal (*squiggle*) producida por los secuenciadores de ONT

Indel: Contracción de inserción-deleción, referida a una inserción o deleción de bases en el genoma de un organismo

IR(S): Repetición Invertida, (en castellano RI), es una secuencia de nucleótidos monocatenaria seguida corriente abajo por su complemento inverso. Por ejemplo:

5'---CATGCnnnnnnnnnGCATG---3'.

En este caso, la secuencia inicial 'CATGC' y su complemento inverso 'GCATG' están separados por una secuencia intermedia de nucleótidos, que puede tener cualquier longitud, como 'nnnnnnnnn' en este caso. Cuando esta secuencia intermedia tiene una longitud de cero, tendríamos una secuencia palindrómica."

Kilo base (Kb): Mil bases (o pares de bases) referido a las cadenas de ácidos nucleicos.

LINE1 íntegros: Aquellos LINE1 sin truncamiento (full length > 6000pb), que no tienen inversiones internas, y con los marcos de lectura intactos. A priori son los que mantienen su capacidad de movilizarse de forma autónoma.

Mapabilidad: (o mapeabilidad en algunos textos): La palabra *"mapabilidad"* (Sexton y Han 2019) no existe en la lengua castellana, y constituye un anglicismo. No obstante, su uso es tan frecuente en el campo que se ha decidido incluirlo en el presente trabajo. Tiene el significado

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de calidad de un alineamiento o de una lectura con respecto al genoma de referencia, algo equivalente a estar en el "mapa". Recientemente la rae ha aceptado el término 'mapear':
tr. Biol. Localizar y representar gráficamente la distribución relativa de las partes de un todo, como los genes en los cromosomas.

Mega base (Mb): Un millón de bases (o pares de bases) referido a las cadenas de ácidos nucleicos.

MEIGA-LR, MEIGA-SR: Programas desarrollados por nuestro grupo de investigación para el análisis y descubrimiento de transposones usando lecturas largas y cortas respectivamente. Enfocado principalmente a las inserciones somáticas en tumores y las reorganizaciones producidas por estas.

Metil-alelo: Es un tipo de epialelo. Donde los distintos alelos tienen distintos patrones de metilación.

Methyl-Meiga: Programa desarrollado por mi para el estudio de la metilación en LINE1 usando datos de lecturas largas de ONT.

***Notebooks* de Python:** Programas de ordenador interactivos (libros) desarrollados en Python, que permiten la visualización, seguimiento y reproducibilidad, muy útiles para la investigación científica.

Pipelines: Neologismo. Serie de programas de ordenador encadenados unos detrás de otros para realizar una tarea determinada.

Southern blot: Técnica de hibridación para detectar la presencia de cadenas concretas de ADN.

Script: Pequeño programa informático, generalmente lanzado de desde la línea de comandos, para ejecutar pequeñas tareas, como por ejemplo el lanzamiento encadenado de programas.

Splicing: Proceso consistente en la escisión de los intrones, las porciones no codificantes de los genes, del transcrito del ARN mensajero primario, seguido de la unión de los exones, (fragmentos codificantes), para producir el ARN mensajero maduro.

Valores β: Representa una aproximación al porcentaje de metilación de un CpG o locus. (Con valores entre 0 y 1)

$$\beta = \frac{C}{U + C}, (0 \leq \beta \leq 1)$$

siendo C la "intensidad" de metilación (o el número de lecturas) del alelo metilado y U la intensidad (o el número de lecturas) del alelo sin metilar.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Valores M: Es el logaritmo en base 2 de la relación entre las intensidades (o número de lecturas) entre el alelo metilado (C) y el no metilado (U). w es un pseudo número añadido para evitar grandes variaciones en el logaritmo cuando se trabaja con ceros.

$$M = \log_2\left(\frac{C+w}{U+w}\right)$$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

11.2 ABREVIATURAS:

Debido al uso frecuente en el campo de abreviaturas en inglés se ha seguido el criterio de usarlas, excepto en el caso de abreviaturas de uso común en castellano

11.2.1 Castellano

ARNm: Ácido ribonucleico mensajero.

ARNt: Ácido ribonucleico de transferencia.

ATP: Adenosín trifosfato.

CESGA: Centro de supercomputación de Galicia.

CpG: dinucleótido de citosina - enlace fosfodiéster "p" – guanina.

Kb: kilo bases, equivalente a 1,000 bases.

Kpb: kilo pares de bases, equivalente a 1,000 pares de bases.

Mb: Mega bases, equivalente a 1,000,000 bases.

Mpb: Mega pares de bases, equivalente a 1,000,000 pares de bases.

RNP: Ribonucleoproteína.

SAM: S-adenosilmetionina.

RT: Retrotransposón.

11.2.2 Inglés

BP (bp): Base pairs, pares de bases.

DMR: Differential methylated regions. Regiones diferencialmente metiladas

DNMTs: DNA methyltransferases. Metil transferasas de ADN.

dNTP: Deoxyribonucleotide triphosphate. Desoxirribonucleótido trifosfato.

DSB: Double strand break. Rotura de la doble cadena.

HPC: High performance computer. Ordenador de altas prestaciones.

HPLC: High-performance liquid chromatography. Cromatografía líquida de alta eficacia.

IRES: Internal ribosome entry site. Sitio interno de entrada al ribosoma.

IRS: Inverted repeats, inversiones repetidas (ver glosario).

LINE: Long interspersed nuclear elements, elementos nucleares intercalados largos.

L1, LINE-1, LINE1: Un tipo específico de LINE que corresponde al 17% del genoma humano.

USC UNIVERSIDADE DE SANTIAGO DE

NA: Not available, equivalente a NaN, nan (not a number).

NGS: Next generation sequencing. Secuenciación de siguiente generación.

ONT: Oxford Nanopore Technologies. Empresa fabricante de los secuenciadores MinION y otros.

SBS: Sequencing by synthesis. Secuenciación usando reacciones con polimerasas.

SINE: Short interspersed nuclear elements, elementos nucleares intercalados cortos.

SIRPH: methylation-dependent primer extension assays. Ensayos de extensión de primers dependientes de metilación.

SMS: Single molecule sequencing. Secuenciación de molécula única.

SMRT: Single-molecule real-time sequencing. Secuenciación en tiempo real de una sola molécula.

SNPs: Single nucleotide polymorphism. Polimorfismo puntual de un nucleótido (en una secuencia de ADN o ARN).

SNVs: Single nucleotide variants. Variaciones puntuales de un nucleótido (en una secuencia de ADN o ARN).

TDS: Tandem site duplications. Duplicaciones de sitio en tandem.

TE: Transposable element. Elemento transponible.

tRNA. Transfer RNA. Ácido ribonucleico de transferencia (ARNt).

UTR: Untranslated region. Región no traducida, en el extremo de los genes y LINEs.

WGS: Secuenciación de todo el genoma, normalmente referido a lecturas cortas.

ZMW: Zero-mode waveguide. Guía de ondas de modo cero.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

12. BIBLIOGRAFÍA

Adusumalli, Swarnaseetha, Mohd Feroz Mohd Omar, Richie Soong, y Touati Benoukraf. 2015. «Methodological aspects of whole-genome bisulfite sequencing analysis». *Briefings in Bioinformatics* 16 (3): 369-79. https://doi.org/10.1093/bib/bbu016.

Ahmadian, Afshin, Maria Ehn, y Sophia Hober. 2006. «Pyrosequencing: History, Biochemistry and Future». *Clinica Chimica Acta* 363 (1-2): 83-94. https://doi.org/10.1016/j.cccn.2005.04.038.

Antelo, Marina, Francesc Balaguer, Jinru Shia, Yan Shen, Keun Hur, Leticia Moreira, Miriam Cuatrecasas, et al. 2012. «A High Degree of LINE-1 Hypomethylation Is a Unique Feature of Early-Onset Colorectal Cancer». *PLoS ONE* 7 (9): e45357. https://doi.org/10.1371/journal.pone.0045357.

Anwar, Sumadi Lukman, Britta Hasemeier, Elisa Schipper, Arndt Vogel, Hans Kreipe, y Ulrich Lehmann. 2019. «LINE-1 hypomethylation in human hepatocellular carcinomas correlates with shorter overall survival and CIMP phenotype». *PLoS ONE* 14 (5): e0216374. https://doi.org/10.1371/journal.pone.0216374.

Arkhipova, Irina R. 2017. «Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories». *Mobile DNA* 8 (1): 19. https://doi.org/10.1186/s13100-017-0103-2.

Athanikar, Jyoti N., Richard M. Badge, y John V. Moran. 2004. «A YY1-binding site is required for accurate human LINE-1 transcription initiation». *Nucleic Acids Research* 32 (13): 3846-55. https://doi.org/10.1093/nar/gkh698.

Bao, Weidong, Kenji K. Kojima, y Oleksiy Kohany. 2015. «Repbase Update, a database of repetitive elements in eukaryotic genomes». *Mobile DNA* 6 (1): 11. https://doi.org/10.1186/s13100-015-0041-9.

Barchitta, Martina, Annalisa Quattrocchi, Andrea Maugeri, Manlio Vinciguerra, y Antonella Agodi. 2014. «LINE-1 Hypomethylation in Blood and Tissue Samples as an Epigenetic Marker for Cancer Risk: A Systematic Review and Meta-Analysis». *PloS One* 9 (10): e109478. https://doi.org/10.1371/journal.pone.0109478.

Beck, Christine R., Pamela Collier, Catriona Macfarlane, Maika Malig, Jeffrey M. Kidd, Evan E. Eichler, Richard M. Badge, y John V. Moran. 2010. «LINE-1 Retrotransposition Activity in Human Genomes». *Cell* 141 (7): 1159-70. https://doi.org/10.1016/j.cell.2010.05.021.

Beck, Christine R., José Luis Garcia-Perez, Richard M. Badge, y John V. Moran. 2011. «LINE-1 Elements in Structural Variation and Disease». *Annual review of genomics and human genetics* 12: 187-215. https://doi.org/10.1146/annurev-genom-082509-141802.

Belancio, Victoria P., Dale J. Hedges, y Prescott Deininger. 2008. «Mammalian Non-LTR Retrotransposons: For Better or Worse, in Sickness and in Health». *Genome Research* 18 (3): 343-58. https://doi.org/10.1101/gr.5558208.

Berger, Shelley L., Tony Kouzarides, Ramin Shiekhattar, y Ali Shilatifard. 2009. «An operational definition of epigenetics». *Genes & Development* 23 (7): 781-83. https://doi.org/10.1101/gad.1787609.

Bodak, Maxime, Jian Yu, y Constance Ciaudo. 2014. «Regulation of LINE-1 in Mammals». *Biomolecular Concepts* 5 (5): 409-28. https://doi.org/10.1515/bmc-2014-0018.

Branco, Miguel R., Gabriella Ficz, y Wolf Reik. 2012. «Uncovering the Role of 5-Hydroxymethylcytosine in the Epigenome». *Nature Reviews Genetics* 13 (1): 7-13. https://doi.org/10.1038/nrg3080.

Brouha, Brook, Joshua Schustak, Richard M. Badge, Sheila Lutz-Prigge, Alexander H. Farley, John V. Moran, y Haig H. Kazazian. 2003. «Hot L1s Account for the Bulk of Retrotransposition in the

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Human Population». *Proceedings of the National Academy of Sciences of the United States of America* 100 (9): 5280-85. https://doi.org/10.1073/pnas.0831042100.

Burns, Kathleen H. 2017. «Transposable Elements in Cancer». *Nature Reviews Cancer* 17 (7): 415-24. https://doi.org/10.1038/nrc.2017.35.

———. 2022. «Repetitive DNA in disease». *Science* 376 (6591): 353-54. https://doi.org/10.1126/science.abl7399.

Castro-Diaz, Nathaly, Gabriela Ecco, Andrea Coluccio, Adamandia Kapopoulou, Benyamin Yazdanpanah, Marc Friedli, Julien Duc, Suk Min Jang, Priscilla Turelli, y Didier Trono. 2014. «Evolutionally Dynamic L1 Regulation in Embryonic Stem Cells». *Genes & Development* 28 (13): 1397-1409. https://doi.org/10.1101/gad.241661.114.

Chalitchagorn, Krisanee, Shanop Shuangshoti, Nusara Hourpai, Narisorn Kongruttanachok, Pisit Tangkijvanich, Duangporn Thong-ngam, Narin Voravud, Virote Sriuranpong, y Apiwat Mutirangura. 2004. «Distinctive Pattern of LINE-1 Methylation Level in Normal Tissues and the Association with Carcinogenesis». *Oncogene* 23 (54): 8841-46. https://doi.org/10.1038/sj.onc.1208137.

Chen, Long, Jane E. Dahlstrom, Sung-Hun Lee, y Danny Rangasamy. 2012. «Naturally Occurring Endo-SiRNA Silences LINE-1 Retrotransposons in Human Cells through DNA Methylation». *Epigenetics* 7 (7): 758-71. https://doi.org/10.4161/epi.20706.

Choi, In-Seon, Marcos R. H. Estecio, Yasuhiko Nagano, Do Ha Kim, Jill A. White, James C. Yao, Jean-Pierre J. Issa, y Asif Rashid. 2007. «Hypomethylation of LINE-1 and Alu in Well-Differentiated Neuroendocrine Tumors (Pancreatic Endocrine Tumors and Carcinoid Tumors)». *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc* 20 (7): 802-10. https://doi.org/10.1038/modpathol.3800825.

Chow, Jennifer, y Edith Heard. 2009. «X Inactivation and the Complexities of Silencing a Sex Chromosome». *Current Opinion in Cell Biology*, Nucleus and gene expression, 21 (3): 359-66. https://doi.org/10.1016/j.ceb.2009.04.012.

Chu, Chong, Rebeca Borges-Monroy, Vinayak V. Viswanadham, Soohyun Lee, Heng Li, Eunjung Alice Lee, y Peter J. Park. 2021. «Comprehensive Identification of Transposable Element Insertions Using Multiple Sequencing Technologies». *Nature Communications* 12 (1): 3836. https://doi.org/10.1038/s41467-021-24041-8.

Cordaux, Richard, y Mark A. Batzer. 2009. «The Impact of Retrotransposons on Human Genome Evolution». *Nature Reviews. Genetics* 10 (10): 691-703. https://doi.org/10.1038/nrg2640.

Cost, Gregory J., Qinghua Feng, Alain Jacquier, y Jef D. Boeke. 2002. «Human L1 element target-primed reverse transcription in vitro». *The EMBO Journal* 21 (21): 5899-5910. https://doi.org/10.1093/emboj/cdf592.

Cruickshanks, Hazel A., y Cristina Tufarelli. 2009. «Isolation of Cancer-Specific Chimeric Transcripts Induced by Hypomethylation of the LINE-1 Antisense Promoter». *Genomics* 94 (6): 397-406. https://doi.org/10.1016/j.ygeno.2009.08.013.

De Cecco, Marco, Takahiro Ito, Anna P. Petrashen, Amy E. Elias, Nicholas J. Skvir, Steven W. Criscione, Alberto Caligiana, et al. 2019. «L1 Drives IFN in Senescent Cells and Promotes Age-Associated Inflammation». *Nature* 566 (7742): 73-78. https://doi.org/10.1038/s41586-018-0784-9.

Deamer, David, Mark Akeson, y Daniel Branton. 2016. «Three decades of nanopore sequencing». *Nature biotechnology* 34 (5): 518-24. https://doi.org/10.1038/nbt.3423.

Deaton, Aimée M., y Adrian Bird. 2011. «CpG islands and the regulation of transcription». *Genes & Development* 25 (10): 1010-22. https://doi.org/10.1101/gad.2037511.

Deininger, Prescott L., y Mark A. Batzer. 2002. «Mammalian Retroelements». *Genome Research* 12 (10): 1455-65. https://doi.org/10.1101/gr.282402.

Deniz, Özgen, Jennifer M. Frost, y Miguel R. Branco. 2019. «Regulation of Transposable Elements by DNA Modifications». *Nature Reviews Genetics* 20 (7): 417-31. https://doi.org/10.1038/s41576-019-0106-6.

Denli, Ahmet M., Iñigo Narvaiza, Bilal E. Kerman, Monique Pena, Christopher Benner, Maria C. N. Marchetto, Jolene K. Diedrich, et al. 2015. «Primate-Specific ORF0 Contributes to

Retrotransposon-Mediated Diversity». *Cell* 163 (3): 583-93. https://doi.org/10.1016/j.cell.2015.09.025.

Dewannieux, Marie, Cécile Esnault, y Thierry Heidmann. 2003. «LINE-Mediated Retrotransposition of Marked Alu Sequences». *Nature Genetics* 35 (1): 41-48. https://doi.org/10.1038/ng1223.

Ebert, Peter, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, et al. 2021. «Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation». *Science*, febrero. https://doi.org/10.1126/science.abf7117.

Eberwine, James, Jai-Yoon Sul, Tamas Bartfai, y Junhyong Kim. 2014. «The Promise of Single-Cell Sequencing». *Nature Methods* 11 (1): 25-27. https://doi.org/10.1038/nmeth.2769.

England, Robert, y Monica Pettersson. 2005. «Pyro Q-CpG™: Quantitative Analysis of Methylation in Multiple CpG Sites by Pyrosequencing®». *Nature Methods* 2 (10): i-ii. https://doi.org/10.1038/nmeth800.

Esteller, Manel. 2008. «Epigenetics in Cancer». *New England Journal of Medicine* 358 (11): 1148-59. https://doi.org/10.1056/NEJMra072067.

Ewing, Adam D., Nathan Smits, Francisco J. Sanchez-Luque, Jamila Faivre, Paul M. Brennan, Sandra R. Richardson, Seth W. Cheetham, y Geoffrey J. Faulkner. 2020. «Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling». *Molecular Cell* 80 (5): 915-928.e5. https://doi.org/10.1016/j.molcel.2020.10.024.

Feinberg, Andrew. 2014. «DNA methylation in cancer: three decades of discovery». *Genome Medicine* 6 (5): 36. https://doi.org/10.1186/gm553.

Feinberg, Andrew P., y Bert Vogelstein. 1983. «Hypomethylation Distinguishes Genes of Some Human Cancers from Their Normal Counterparts». *Nature* 301 (5895): 89-92. https://doi.org/10.1038/301089a0.

Feng, Yanxiao, Yuechuan Zhang, Cuifeng Ying, Deqiang Wang, y Chunlei Du. 2015. «Nanopore-based Fourth-generation DNA Sequencing Technology». *Genomics, Proteomics & Bioinformatics* 13 (1): 4-16. https://doi.org/10.1016/j.gpb.2015.01.009.

Feschotte, Cédric, y Ellen J. Pritham. 2007. «DNA Transposons and the Evolution of Eukaryotic Genomes». *Annual review of genetics* 41: 331-68. https://doi.org/10.1146/annurev.genet.40.110405.090448.

Finnegan, David J. 1989. «Eukaryotic Transposable Elements and Genome Evolution». *Trends in Genetics* 5 (enero): 103-7. https://doi.org/10.1016/0168-9525(89)90039-5.

Florl, A R, C Steinhoff, M Müller, H-H Seifert, C Hader, R Engers, R Ackermann, y W A Schulz. 2004. «Coordinate hypermethylation at specific genes in prostate carcinoma precedes LINE-1 hypomethylation». *British Journal of Cancer* 91 (5): 985-94. https://doi.org/10.1038/sj.bjc.6602030.

Gao, Fei, Yudong Xia, Junwen Wang, Huijuan Luo, Zhaowei Gao, Xu Han, Juyong Zhang, et al. 2013. «Integrated detection of both 5-mC and 5-hmC by high-throughput tag sequencing technology highlights methylation reprogramming of bivalent genes during cellular differentiation». *Epigenetics* 8 (4): 421-30. https://doi.org/10.4161/epi.24280.

Garcia-Perez, Jose L., Maria Morell, Joshua O. Scheys, Deanna A. Kulpa, Santiago Morell, Christoph C. Carter, Gary D. Hammer, et al. 2010. «Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells». *Nature* 466 (7307): 769-73. https://doi.org/10.1038/nature09209.

Gardiner-Garden, M., y M. Frommer. 1987. «CpG Islands in Vertebrate Genomes». *Journal of Molecular Biology* 196 (2): 261-82. https://doi.org/10.1016/0022-2836(87)90689-9.

Gardner, Eugene J., Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, The 1000 Genomes Project Consortium, y Scott E. Devine. 2017. «The Mobile Element Locator Tool (MELT): Population-Scale Mobile Element Discovery and Biology». *Genome Research* 27 (11): 1916-29. https://doi.org/10.1101/gr.218032.116.

Garg, Shilpa, Marcel Martin, y Tobias Marschall. 2016. «Read-based phasing of related individuals». *Bioinformatics* 32 (12): i234-42. https://doi.org/10.1093/bioinformatics/btw276.

USC DE SANTIAGO DE COMPOSTELA

Garrison, Erik, y Gabor Marth. 2012. «Haplotype-based variant detection from short-read sequencing». arXiv. https://doi.org/10.48550/arXiv.1207.3907.

Gazdar, Adi F., Luc Girard, William W. Lockwood, Wan L. Lam, y John D. Minna. 2010. «Lung Cancer Cell Lines as Tools for Biomedical Discovery and Research». *JNCI Journal of the National Cancer Institute* 102 (17): 1310-21. https://doi.org/10.1093/jnci/djq279.

Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, M. Math, James Larkin, David Endesfelder, Eva Gronroos, et al. 2012. «Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing». *The New England journal of medicine* 366 (10): 883-92. https://doi.org/10.1056/NEJMoa1113205.

Goel, Ajay, Rosa M. Xicola, Thuy-Phuong Nguyen, Brian J Doyle, Vanessa R. Sohn, Prathap Bandipalliam, Josep Reyes, et al. 2010. «Aberrant DNA Methylation in Hereditary Non-Polyposis Colorectal Cancer without Mismatch Repair Deficiency». *Gastroenterology* 138 (5): 1854-62. https://doi.org/10.1053/j.gastro.2010.01.035.

Gonçalves, Ana, Jorge Oliveira, Teresa Coelho, Ricardo Taipa, Manuel Melo-Pires, Mário Sousa, y Rosário Santos. 2017. «Exonization of an Intronic LINE-1 Element Causing Becker Muscular Dystrophy as a Novel Mutational Mechanism in Dystrophin Gene». *Genes* 8 (10): 253. https://doi.org/10.3390/genes8100253.

Goodier, John L. 2016. «Restricting retrotransposons: a review». *Mobile DNA* 7 (agosto): 16. https://doi.org/10.1186/s13100-016-0070-z.

Goodier, John L., y Haig H. Kazazian. 2008. «Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites». *Cell* 135 (1): 23-35. https://doi.org/10.1016/j.cell.2008.09.022.

Grandi, Fiorella C., James M. Rosser, Simon J. Newkirk, Jun Yin, Xiaoling Jiang, Zhuo Xing, Leanne Whitmore, et al. 2015. «Retrotransposition Creates Sloping Shores: A Graded Influence of Hypomethylated CpG Islands on Flanking CpG Sites». *Genome Research* 25 (8): 1135-46. https://doi.org/10.1101/gr.185132.114.

Gujar, Hemant, Daniel J. Weisenberger, y Gangning Liang. 2019. «The Roles of Human DNA Methyltransferases and Their Isoforms in Shaping the Epigenome». *Genes* 10 (2): 172. https://doi.org/10.3390/genes10020172.

Guler, Gulfem Dilek, Charles Albert Tindell, Robert Pitti, Catherine Wilson, Katrina Nichols, Tommy KaiWai Cheung, Hyo-Jin Kim, et al. 2017. «Repression of Stress-Induced LINE-1 Expression Protects Cancer Cell Subpopulations from Lethal Drug Exposure». *Cancer Cell* 32 (2): 221-237.e13. https://doi.org/10.1016/j.ccell.2017.07.002.

Guo, Hongshan, Ping Zhu, Fan Guo, Xianlong Li, Xinglong Wu, Xiaoying Fan, Lu Wen, y Fuchou Tang. 2015. «Profiling DNA Methylome Landscapes of Mammalian Cells with Single-Cell Reduced-Representation Bisulfite Sequencing». *Nature Protocols* 10 (5): 645-59. https://doi.org/10.1038/nprot.2015.039.

Hanahan, Douglas, y Robert A. Weinberg. 2000. «The Hallmarks of Cancer». *Cell* 100 (1): 57-70. https://doi.org/10.1016/S0092-8674(00)81683-9.

———. 2011. «Hallmarks of Cancer: The Next Generation». *Cell* 144 (5): 646-74. https://doi.org/10.1016/j.cell.2011.02.013.

Hancks, Dustin C., y Haig H. Kazazian. 2016. «Roles for retrotransposon insertions in human disease». *Mobile DNA* 7 (mayo): 9. https://doi.org/10.1186/s13100-016-0065-9.

Hattori, Naoko, y Toshikazu Ushijima. 2017. «Analysis of Gene-Specific DNA Methylation». En *Handbook of Epigenetics*, 113-23. Elsevier. https://doi.org/10.1016/B978-0-12-805388-1.00008-0.

Heather, James M., y Benjamin Chain. 2016. «The sequence of sequencers: The history of sequencing DNA». *Genomics* 107 (1): 1-8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Helman, Elena, Michael L. Lawrence, Chip Stewart, Carrie Sougnez, Gad Getz, y Matthew Meyerson. 2014. «Somatic Retrotransposition in Human Cancer Revealed by Whole-Genome and Exome Sequencing». *Genome Research*, mayo, gr.163659.113. https://doi.org/10.1101/gr.163659.113.

Hoesel, Anneke Q. van, Cornelis J. H. van de Velde, Peter J. K. Kuppen, Gerrit Jan Liefers, Hein Putter, Yusuke Sato, David A. Elashoff, et al. 2012. «Hypomethylation of LINE-1 in Primary

Tumor Has Poor Prognosis in Young Breast Cancer Patients: A Retrospective Cohort Study». *Breast Cancer Research and Treatment* 134 (3): 1103-14. https://doi.org/10.1007/s10549-012-2038-0.

Holley, Robert W. 1965. «Structure of an Alanine Transfer Ribonucleic Acid». *JAMA* 194 (8): 868-71. https://doi.org/10.1001/jama.1965.03090210032009.

Holliday, Robin, Robin Holliday, Marilyn Monk, y J. E. Pugh. 1997. «DNA methylation and epigenetic inheritance». *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 326 (1235): 329-38. https://doi.org/10.1098/rstb.1990.0015.

Howard, G., R. Eiges, F. Gaudet, R. Jaenisch, y A. Eden. 2008. «Activation and Transposition of Endogenous Retroviral Elements in Hypomethylation Induced Tumors in Mice». *Oncogene* 27 (3): 404-8. https://doi.org/10.1038/sj.onc.1210631.

Huang, Kevin Yu Yuan, Yan-Jiun Huang, y Pao-Yang Chen. 2018. «BS-Seeker3: ultrafast pipeline for bisulfite sequencing». *BMC Bioinformatics* 19 (1): 111. https://doi.org/10.1186/s12859-018-2120-7.

Hubley, Robert, Robert D. Finn, Jody Clements, Sean R. Eddy, Thomas A. Jones, Weidong Bao, Arian F.A. Smit, y Travis J. Wheeler. 2016. «The Dfam database of repetitive DNA families». *Nucleic Acids Research* 44 (D1): D81-89. https://doi.org/10.1093/nar/gkv1272.

Iwasaki, Yuka W., Mikiko C. Siomi, y Haruhiko Siomi. 2015. «PIWI-Interacting RNA: Its Biogenesis and Functions». *Annual Review of Biochemistry* 84: 405-33. https://doi.org/10.1146/annurev-biochem-060614-034258.

Jachowicz, Joanna W., Xinyang Bing, Julien Pontabry, Ana Bošković, Oliver J. Rando, y Maria-Elena Torres-Padilla. 2017. «LINE-1 Activation after Fertilization Regulates Global Chromatin Accessibility in the Early Mouse Embryo». *Nature Genetics* 49 (10): 1502-10. https://doi.org/10.1038/ng.3945.

Johnstone, Sarah E., Vadim N. Gladyshev, Martin J. Aryee, y Bradley E. Bernstein. 2022. «Epigenetic clocks, aging, and cancer». *Science* 378 (6626): 1276-77. https://doi.org/10.1126/science.abn4009.

Jones, Peter A., y Stephen B. Baylin. 2007. «The Epigenomics of Cancer». *Cell* 128 (4): 683-92. https://doi.org/10.1016/j.cell.2007.01.029.

Jou, W. Min, G. Haegeman, M. Ysebaert, y W. Fiers. 1972. «Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein». *Nature* 237 (mayo): 82-88. https://doi.org/10.1038/237082a0.

Jürgens, B., B. J. Schmitz-Dräger, y W. A. Schulz. 1996. «Hypomethylation of L1 LINE Sequences Prevailing in Human Urothelial Carcinoma». *Cancer Research* 56 (24): 5698-5703.

Kannan, Manoj, Jingfeng Li, Sarah E. Fritz, Kathryn E. Husarek, Jonathan C. Sanford, Teresa L. Sullivan, Pawan Kumar Tiwary, Wenfeng An, Jef D. Boeke, y David E. Symer. 2017. «Dynamic silencing of somatic L1 retrotransposon insertions reflects the developmental and cellular contexts of their genomic integration». *Mobile DNA* 8 (1): 8. https://doi.org/10.1186/s13100-017-0091-2.

Kasianowicz, John J., Eric Brandin, Daniel Branton, y David W. Deamer. 1996. «Characterization of Individual Polynucleotide Molecules Using a Membrane Channel». *Proceedings of the National Academy of Sciences* 93 (24): 13770-73.

Kawano, Hiroyuki, Hiroshi Saeki, Hiroyuki Kitao, Yasuo Tsuda, Hajime Otsu, Koji Ando, Shuhei Ito, et al. 2014. «Chromosomal Instability Associated with Global DNA Hypomethylation Is Associated with the Initiation and Progression of Esophageal Squamous Cell Carcinoma». *Annals of Surgical Oncology* 21 (4): 696-702. https://doi.org/10.1245/s10434-014-3818-z.

Kazazian, Haig H. 2004. «Mobile Elements: Drivers of Genome Evolution». *Science* 303 (5664): 1626-32. https://doi.org/10.1126/science.1089670.

Kazazian, Haig H. Jr., y John V. Moran. 2017. «Mobile DNA in Health and Disease». *New England Journal of Medicine* 377 (4): 361-70. https://doi.org/10.1056/NEJMra1510092.

Kazazian, Haig. H., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, y S. E. Antonarakis. 1988. «Haemophilia A Resulting from de Novo Insertion of L1 Sequences Represents a Novel

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Mechanism for Mutation in Man». *Nature* 332 (6160): 164-66. https://doi.org/10.1038/332164a0.

Koning, A. P. Jason de, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, y David D. Pollock. 2011. «Repetitive Elements May Comprise Over Two-Thirds of the Human Genome». *PLOS Genetics* 7 (12): e1002384. https://doi.org/10.1371/journal.pgen.1002384.

Kraft, Florian, y Ingo Kurth. 2019. «Long-Read Sequencing in Human Genetics». *Medizinische Genetik* 31 (2): 198-204. https://doi.org/10.1007/s11825-019-0249-z.

Krerowicz, Samuel J.W., Juan P. Hernandez-Ortiz, y David C. Schwartz. 2019. «A Simple Dialysis Device for Large DNA Molecules». *BioTechniques* 66 (2): 93-95. https://doi.org/10.2144/btn-2018-0133.

Kroutter, Emily N., Victoria P. Belancio, Bradley J. Wagstaff, y Astrid M. Roy-Engel. 2009. «The RNA Polymerase Dictates ORF1 Requirement and Timing of LINE and SINE Retrotransposition». *PLoS Genetics* 5 (4): e1000458. https://doi.org/10.1371/journal.pgen.1000458.

Krueger, Felix, y Simon R. Andrews. 2011. «Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications». *Bioinformatics* 27 (11): 1571-72. https://doi.org/10.1093/bioinformatics/btr167.

Lanciano, Sophie, Claude Philippe, Arpita Sarkar, Cécilia Domrane, Aurélien J. Doucet, Dominic van Essen, Simona Saccani, Laure Ferry, Pierre-Antoine Defossez, y Gael Cristofari. 2023. «Resolving the Heterogeneity of L1 DNA Methylation Reveals the Epigenetic and Transcriptional Interplay between L1s and Their Integration Sites». bioRxiv. https://doi.org/10.1101/2023.01.03.522582.

Lao, Victoria Valinluck, y William M. Grady. 2011. «Epigenetics and Colorectal Cancer». *Nature Reviews Gastroenterology & Hepatology* 8 (12): 686-700. https://doi.org/10.1038/nrgastro.2011.173.

Lazaro-Camp, Vanessa J., Kiarash Salari, Xiangbing Meng, y Shujie Yang. 2021. «SETDB1 in Cancer: Overexpression and Its Therapeutic Implications». *American Journal of Cancer Research* 11 (5): 1803-27.

Lederberg, Esther M., y Joshua Lederberg. 1953. «Genetic Studies of Lysogenicity in Escherichia Coli». *Genetics* 38 (1): 51-64.

Lee, Eunjung, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace J. Luquette, Jens G. Lohr, et al. 2012. «Landscape of Somatic Retrotransposition in Human Cancers». *Science* 337 (6097): 967-71. https://doi.org/10.1126/science.1222077.

Lee, Haeun, Jun Won Min, Seyoung Mun, y Kyudong Han. 2022. «Human Retrotransposons and Effective Computational Detection Methods for Next-Generation Sequencing Data». *Life* 12 (10): 1583. https://doi.org/10.3390/life12101583.

Li, Heng. 2013. «Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM». *arXiv:1303.3997 [q-bio]*, marzo. http://arxiv.org/abs/1303.3997.

———. 2018. «Minimap2: pairwise alignment for nucleotide sequences». *Bioinformatics* 34 (18): 3094-3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, y Richard Durbin. 2010. «Fast and accurate long-read alignment with Burrows–Wheeler transform». *Bioinformatics* 26 (5): 589-95. https://doi.org/10.1093/bioinformatics/btp698.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, y 1000 Genome Project Data Processing Subgroup. 2009. «The Sequence Alignment/Map Format and SAMtools». *Bioinformatics (Oxford, England)* 25 (16): 2078-79. https://doi.org/10.1093/bioinformatics/btp352.

Lin, Jianhuang, Dajiang Guo, Heng Liu, Wei Zhou, Chen Wang, Iris Müller, Andrew V. Kossenkov, et al. 2021. «The SETDB1–TRIM28 complex suppresses antitumor immunity». *Cancer immunology research* 9 (12): 1413-24. https://doi.org/10.1158/2326-6066.CIR-21-0754.

Liu, Qian, Li Fang, Guoliang Yu, Depeng Wang, Chuan-Le Xiao, y Kai Wang. 2019. «Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data». *Nature Communications* 10 (1): 2449. https://doi.org/10.1038/s41467-019-10168-2.

Liu, Yang, Wojciech Rosikiewicz, Ziwei Pan, Nathaniel Jillette, Ping Wang, Aziz Taghbalout, Jonathan Foox, et al. 2021. «DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation». *Genome Biology* 22 (1): 295. https://doi.org/10.1186/s13059-021-02510-z.

Loyfer, Netanel, Judith Magenheim, Ayelet Peretz, Gordon Cann, Joerg Bredno, Agnes Klochendler, Ilana Fox-Fisher, et al. 2023. «A DNA Methylation Atlas of Normal Human Cell Types». *Nature*, enero, 1-10. https://doi.org/10.1038/s41586-022-05580-6.

Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schoenhuth, y Tobias Marschall. 2016. «WhatsHap: Fast and Accurate Read-Based Phasing». *BioRxiv*, noviembre, 085050. https://doi.org/10.1101/085050.

Marzese, Diego M., Dave S. B. Hoon, Kelly K. Chong, Francisco E. Gago, Javier I. Orozco, Olga M. Tello, Laura M. Vargas-Roig, y María Roqué. 2012. «DNA Methylation Index and Methylation Profile of Invasive Ductal Breast Tumors». *The Journal of Molecular Diagnostics* 14 (6): 613-22. https://doi.org/10.1016/j.jmoldx.2012.07.001.

Marzese, Diego M., y Dave S.B. Hoon. 2015. «Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer». *Expert review of molecular diagnostics* 15 (5): 647-64. https://doi.org/10.1586/14737159.2015.1027194.

Matsunoki, Aika, Kazuyuki Kawakami, Masanori Kotake, Mami Kaneko, Hirotaka Kitamura, Akishi Ooi, Go Watanabe, y Toshinari Minamoto. 2012. «LINE-1 methylation shows little intra-patient heterogeneity in primary and synchronous metastatic colorectal cancer». *BMC Cancer* 12 (1): 574. https://doi.org/10.1186/1471-2407-12-574.

Maxam, A. M., y W. Gilbert. 1977. «A New Method for Sequencing DNA». *Proceedings of the National Academy of Sciences* 74 (2): 560-64.

McClintock, Barbara. 1950. «The Origin and Behavior of Mutable Loci in Maize». *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344-55.

Meissner, Alexander, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, y Rudolf Jaenisch. 2005. «Reduced Representation Bisulfite Sequencing for Comparative High-Resolution DNA Methylation Analysis». *Nucleic Acids Research* 33 (18): 5868-77. https://doi.org/10.1093/nar/gki901.

Miller, Ian, Max Totrov, Lioubov Korotchkina, Denis N Kazyulkin, Andrei V Gudkov, y Sergey Korolev. 2021. «Structural dissection of sequence recognition and catalytic mechanism of human LINE-1 endonuclease». *Nucleic Acids Research*, n.º gkab826 (septiembre). https://doi.org/10.1093/nar/gkab826.

Miousse, Isabelle R., y Igor Koturbash. 2015. «The Fine LINE: Methylation Drawing the Cancer Landscape». *BioMed Research International* 2015. https://doi.org/10.1155/2015/131547.

Mir, Ashfaq A., Claude Philippe, y Gaël Cristofari. 2015. «euL1db: the European database of L1HS retrotransposon insertions in humans». *Nucleic Acids Research* 43 (Database issue): D43-47. https://doi.org/10.1093/nar/gku1043.

Mojarad, Ehsan Nazemalhosseini, Peter JK Kuppen, Hamid Asadzadeh Aghdaei, y Mohammad Reza Zali. 2013. «The CpG island methylator phenotype (CIMP) in colorectal cancer». *Gastroenterology and Hepatology From Bed to Bench* 6 (3): 120-28.

Monot, Clément, Monika Kuciak, Sébastien Viollet, Ashfaq Ali Mir, Caroline Gabus, Jean-Luc Darlix, y Gaël Cristofari. 2013. «The Specificity and Flexibility of L1 Reverse Transcription Priming at Imperfect T-Tracts». *PLOS Genetics* 9 (5): e1003499. https://doi.org/10.1371/journal.pgen.1003499.

Moore, Tom. 2002. «Southern Analysis Using Methyl-Sensitive Restriction Enzymes». En *Genomic Imprinting: Methods and Protocols*, editado por Andrew Ward, 193-203. Methods in Molecular Biology™. Totowa, NJ: Humana Press. https://doi.org/10.1385/1-59259-211-2:193.

Moran, J. V., R. J. DeBerardinis, y H. H. Kazazian. 1999. «Exon Shuffling by L1 Retrotransposition». *Science (New York, N.Y.)* 283 (5407): 1530-34. https://doi.org/10.1126/science.283.5407.1530.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Moran, J. V., H. H. Kazazian, y J. D. Boeke. 1996. «Human L1 Retrotransposon Encodes a Conserved Endonuclease Required for Retrotransposition». *Cell* 87 (5): 905-16. https://doi.org/10.1016/s0092-8674(00)81997-2.

Moran, John V., y Nicolas Gilbert. 2007. «Mammalian LINE-1 Retrotransposons and Related Elements». En *Mobile DNA II*, 836-69. John Wiley & Sons, Ltd. https://doi.org/10.1128/9781555817954.ch35.

Morrish, Tammy A., Nicolas Gilbert, Jeremy S. Myers, Bethaney J. Vincent, Thomas D. Stamato, Guillermo E. Taccioli, Mark A. Batzer, y John V. Moran. 2002. «DNA Repair Mediated by Endonuclease-Independent LINE-1 Retrotransposition». *Nature Genetics* 31 (2): 159-65. https://doi.org/10.1038/ng898.

Muotri, Alysson R., Maria C. N. Marchetto, Nicole G. Coufal, Ruth Oefner, Gene Yeo, Kinichi Nakashima, y Fred H. Gage. 2010. «L1 Retrotransposition in Neurons Is Modulated by MeCP2». *Nature* 468 (7322): 443-46. https://doi.org/10.1038/nature09544.

Murata, A., Y. Baba, M. Watanabe, H. Shigaki, K. Miyake, T. Ishimoto, M. Iwatsuki, et al. 2013. «Methylation Levels of LINE-1 in Primary Lesion and Matched Metastatic Lesions of Colorectal Cancer». *British Journal of Cancer* 109 (2): 408-15. https://doi.org/10.1038/bjc.2013.289.

Nam, Chang Hyun, Jeonghwan Youk, Jeong Yeon Kim, Joonoh Lim, Jung Woo Park, Soo A. Oh, Hyun Jung Lee, et al. 2022. «Extensive Mosaicism by Somatic L1 Retrotransposition in Normal Human Cells». bioRxiv. https://www.biorxiv.org/content/10.1101/2022.05.18.492429v1.

Navin, Nicholas E. 2015. «The first five years of single-cell cancer genomics and beyond». *Genome Research* 25 (10): 1499-1507. https://doi.org/10.1101/gr.191098.115.

Ni, Peng, Jinrui Xu, Zeyu Zhong, Jun Zhang, Neng Huang, Fan Nie, Feng Luo, y Jianxin Wang. 2022. «DNA 5-Methylcytosine Detection and Methylation Phasing Using PacBio Circular Consensus Sequencing». bioRxiv. https://doi.org/10.1101/2022.02.26.482074.

Nik-Zainal, Serena, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, et al. 2016. «Landscape of Somatic Mutations in 560 Breast Cancer Whole-Genome Sequences». *Nature* 534 (7605): 47-54. https://doi.org/10.1038/nature17676.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. «The Life History of 21 Breast Cancers». *Cell* 149 (5): 994-1007. https://doi.org/10.1016/j.cell.2012.04.023.

Nüsgen, Nicole, Wolfgang Goering, Albertas Dauksa, Arijit Biswas, Muhammad Ahmer Jamil, Ioanna Dimitriou, Amit Sharma, et al. 2015. «Inter-locus as well as intra-locus heterogeneity in LINE-1 promoter methylation in common human cancers suggests selective demethylation pressure at specific CpGs». *Clinical Epigenetics* 7 (1): 17. https://doi.org/10.1186/s13148-015-0051-y.

Ogino, Shuji, Takako Kawasaki, Katsuhiko Nosho, Mutsuko Ohnishi, Yuko Suemoto, Gregory J. Kirkner, y Charles S. Fuchs. 2008. «LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype (CIMP) in colorectal cancer». *International journal of cancer. Journal international du cancer* 122 (12): 2767-73. https://doi.org/10.1002/ijc.23470.

Ogino, Shuji, Katsuhiko Nosho, Gregory J. Kirkner, Takako Kawasaki, Andrew T. Chan, Eva S. Schernhammer, Edward L. Giovannucci, y Charles S. Fuchs. 2008. «A Cohort Study of Tumoral LINE-1 Hypomethylation and Prognosis in Colon Cancer». *JNCI Journal of the National Cancer Institute* 100 (23): 1734-38. https://doi.org/10.1093/jnci/djn359.

Ostertag, E. M., y H. H. Kazazian. 2001. «Biology of Mammalian L1 Retrotransposons». *Annual Review of Genetics* 35: 501-38. https://doi.org/10.1146/annurev.genet.35.102401.091032.

Ostertag, Eric M., y Haig H. Kazazian. 2001. «Twin Priming: A Proposed Mechanism for the Creation of Inversions in L1 Retrotransposition». *Genome Research* 11 (12): 2059-65. https://doi.org/10.1101/gr.205701.

Pavlovic, Sonja, Kristel Klaassen, Biljana Stankovic, Maja Stojiljkovic, y Branka Zukic. 2020. «Chapter 9 - Next-Generation Sequencing: The Enabler and the Way Ahead». En

Microbiomics, editado por Manousos E. Kambouris y Aristea Velegraki, 175-200. Translational and Applied Genomics. Academic Press. https://doi.org/10.1016/B978-0-12-816664-2.00009-8.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, y Matthew Loose. 2019. «BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files». *Bioinformatics* 35 (13): 2193-98. https://doi.org/10.1093/bioinformatics/bty841.

Peddigari, Suresh, Patrick Wai-Lun Li, Jennifer L. Rabe, y Sandra L. Martin. 2013. «hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition». *Nucleic Acids Research* 41 (1): 575-85. https://doi.org/10.1093/nar/gks1075.

Penzkofer, Tobias, Marten Jäger, Marek Figlerowicz, Richard Badge, Stefan Mundlos, Peter N. Robinson, y Tomasz Zemojtel. 2017. «L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes». *Nucleic Acids Research* 45 (D1): D68-73. https://doi.org/10.1093/nar/gkw925.

Petri, Rebecca, Per Ludvik Brattås, Yogita Sharma, Marie E. Jönsson, Karolina Pircs, Johan Bengzon, y Johan Jakobsson. 2019. «LINE-2 transposable elements are a source of functional human microRNAs and target sites». *PLoS Genetics* 15 (3): e1008036. https://doi.org/10.1371/journal.pgen.1008036.

Phokaew, Chureerat, Supakit Kowudtitham, Keskanya Subbalekha, Shanop Shuangshoti, y Apiwat Mutirangura. 2008. «LINE-1 methylation patterns of different loci in normal and cancerous cells». *Nucleic Acids Research* 36 (17): 5704-12. https://doi.org/10.1093/nar/gkn571.

Piégu, Benoît, Solenne Bire, Peter Arensburger, y Yves Bigot. 2015. «A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity». *Molecular Phylogenetics and Evolution* 86 (mayo): 90-109. https://doi.org/10.1016/j.ympev.2015.03.009.

Poulin, Matthew, Jeffrey Y. Zhou, Liying Yan, y Toshi Shioda. 2018. «Pyrosequencing Methylation Analysis». *Methods in Molecular Biology (Clifton, N.J.)* 1856: 283-96. https://doi.org/10.1007/978-1-4939-8751-1_17.

Pradhan, Barun, Tatiana Cajuso, Riku Katainen, Päivi Sulo, Tomas Tanskanen, Outi Kilpivaara, Esa Pitkänen, Lauri A. Aaltonen, Liisa Kauppi, y Kimmo Palin. 2017. «Detection of Subclonal L1 Transductions in Colorectal Cancer by Long-Distance Inverse-PCR and Nanopore Sequencing». *Scientific Reports* 7 (1): 14521. https://doi.org/10.1038/s41598-017-15076-3.

Protasova, Maria Sergeevna, Tatiana Vladimirovna Andreeva, y Evgeny Ivanovich Rogaev. 2021. «Factors Regulating the Activity of LINE1 Retrotransposons». *Genes* 12 (10): 1562. https://doi.org/10.3390/genes12101562.

Qian, Yaping, Debora Mancini-DiNardo, Thaddeus Judkins, Hannah C. Cox, Krystal Brown, Maria Elias, Nanda Singh, et al. 2017. «Identification of pathogenic retrotransposon insertions in cancer predisposition genes». *Cancer Genetics* 216 (Supplement C): 159-69. https://doi.org/10.1016/j.cancergen.2017.08.002.

Rahrmann, Eric P., y David A. Largaespada. 2013. «Retrotransposons: A New and Credible Source of Inherited and Somatically Acquired Hepatocellular Carcinoma Mutations». *Journal of Hepatology* 59 (3): 616-17. https://doi.org/10.1016/j.jhep.2013.05.001.

Rauch, Tibor A., y Gerd P. Pfeifer. 2017. «Methods for Assessing DNA Cytosine Modifications Genome-Wide». En *Handbook of Epigenetics*, 125-34. Elsevier. https://doi.org/10.1016/B978-0-12-805388-1.00009-2.

Richardson, Sandra R, Iñigo Narvaiza, Randy A Planegger, Matthew D Weitzman, y John V Moran. 2014. «APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition». Editado por Anne Ferguson-Smith. *eLife* 3 (abril): e02008. https://doi.org/10.7554/eLife.02008.

Rodriguez-Martin, Bernardo, Eva G. Alvarez, Adrian Baez-Ortega, Jorge Zamora, Fran Supek, Jonas Demeulemeester, Martin Santamarina, et al. 2020. «Pan-Cancer Analysis of Whole Genomes Identifies Driver Rearrangements Promoted by LINE-1 Retrotransposition». *Nature Genetics* 52 (3): 306-19. https://doi.org/10.1038/s41588-019-0562-0.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Rollins, Robert A., Fatemeh Haghighi, John R. Edwards, Rajdeep Das, Michael Q. Zhang, Jingyue Ju, y Timothy H. Bestor. 2006. «Large-Scale Structure of Genomic Methylation Patterns». *Genome Research* 16 (2): 157-63. https://doi.org/10.1101/gr.4362006.

Sabatella, Mariangela, Tuomo Mantere, Esmé Waanders, Kornelia Neveling, Arjen R Mensenkamp, Freerk van Dijk, Jayne Y Hehir-Kwa, et al. 2021. «Optical Genome Mapping Identifies a Germline Retrotransposon Insertion in SMARCB1 in Two Siblings with Atypical Teratoid Rhabdoid Tumors». *The Journal of Pathology* 255 (2): 202-11. https://doi.org/10.1002/path.5755.

Sánchez, Carlos, Paolo Dessì, Maeve Duffy, y Piet N. L. Lens. 2020. «OpenTCC: An Open Source Low-Cost Temperature-Control Chamber». *HardwareX* 7 (abril): e00099. https://doi.org/10.1016/j.ohx.2020.e00099.

Sandoval, Juan, y Manel Esteller. 2012. «Cancer Epigenomics: Beyond Genomics». *Current Opinion in Genetics & Development*, Cancer genomics, 22 (1): 50-55. https://doi.org/10.1016/j.gde.2012.02.008.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, y M. Smith. 1977. «Nucleotide Sequence of Bacteriophage ΦX174 DNA». *Nature* 265 (5596): 687-95. https://doi.org/10.1038/265687a0.

Sanger, F., G. G. Brownlee, y B. G. Barrell. 1965. «A Two-Dimensional Fractionation Procedure for Radioactive Nucleotides». *Journal of Molecular Biology* 13 (2): 373-IN4. https://doi.org/10.1016/S0022-2836(65)80104-8.

Sanger, F., S. Nicklen, y A. R. Coulson. 1977. «DNA Sequencing with Chain-Terminating Inhibitors». *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463-67.

Sarkar, Arpita, Sophie Lanciano, y Gael Cristofari. 2022. «Targeted Nanopore Resequencing and Methylation Analysis of LINE-1 Retrotransposons». bioRxiv. https://doi.org/10.1101/2022.06.25.497594.

Schalamun, Miriam, Ramawatar Nagar, David Kainer, Eleanor Beavan, David Eccles, John P. Rathjen, Robert Lanfear, y Benjamin Schwessinger. 2019. «Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora». *Molecular Ecology Resources* 19 (1): 77-89. https://doi.org/10.1111/1755-0998.12938.

Schauer, Stephanie N., Patricia E. Carreira, Ruchi Shukla, Daniel J. Gerhardt, Patricia Gerdes, Francisco J. Sanchez-Luque, Paola Nicoli, et al. 2018. «L1 Retrotransposition Is a Common Feature of Mammalian Hepatocarcinogenesis». *Genome Research* 28 (5): 639-53. https://doi.org/10.1101/gr.226993.117.

Schulz, Wolfgang A. 2006. «L1 Retrotransposons in Human Cancers». *Journal of Biomedicine and Biotechnology* 2006: 83672. https://doi.org/10.1155/JBB/2006/83672.

Sexton, Corinne, y Mira Han. 2019. «Paired-end mappability of transposable elements in the human genome». *Mobile DNA* 10 (julio). https://doi.org/10.1186/s13100-019-0172-5.

Shademan, Milad, Khadijeh Zare, Morteza Zahedi, Hooman Mosannen Mozaffari, Hadi Bagheri Hosseini, Kamran Ghaffarzadegan, Ladan Goshayeshi, y Hesam Dehghani. 2020. «Promoter methylation, transcription, and retrotransposition of LINE-1 in colorectal adenomas and adenocarcinomas». *Cancer Cell International* 20 (1): 426. https://doi.org/10.1186/s12935-020-01511-5.

Sharma, Amit, Muhammad A. Jamil, Nicole Nuesgen, Albertas Dauksa, Antanas Gulbinas, Wolfgang A. Schulz, Johannes Oldenburg, y Osman El-Maarri. 2019. «Detailed Methylation Map of LINE-1 5'-Promoter Region Reveals Hypomethylated CpG Hotspots Associated with Tumor Tissue Specificity». *Molecular Genetics & Genomic Medicine* 7 (5): e601. https://doi.org/10.1002/mgg3.601.

Shen, Shu Yi, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael H. A. Roehrl, Dianne Chadwick, Philip C. Zuzarte, et al. 2018. «Sensitive Tumour Detection and Classification Using Plasma Cell-Free DNA Methylomes». *Nature* 563 (7732): 579-83. https://doi.org/10.1038/s41586-018-0703-0.

Shigaki, Hironobu, Yoshifumi Baba, Masayuki Watanabe, Asuka Murata, Shiro Iwagami, Keisuke Miyake, Takatsugu Ishimoto, Masaaki Iwatsuki, y Hideo Baba. 2013. «LINE-1 hypomethylation in gastric cancer, detected by bisulfite pyrosequencing, is associated with poor prognosis». *Gastric Cancer* 16 (4): 480-87. https://doi.org/10.1007/s10120-012-0209-7.

Simmen, Martin W. 2008. «Genome-Scale Relationships between Cytosine Methylation and Dinucleotide Abundances in Animals». *Genomics* 92 (1): 33-40. https://doi.org/10.1016/j.ygeno.2008.03.009.

Siomi, Mikiko C., Kaoru Sato, Dubravka Pezic, y Alexei A. Aravin. 2011. «PIWI-Interacting Small RNAs: The Vanguard of Genome Defence». *Nature Reviews. Molecular Cell Biology* 12 (4): 246-58. https://doi.org/10.1038/nrm3089.

Smit, A. F., G. Tóth, A. D. Riggs, y J. Jurka. 1995. «Ancestral, Mammalian-Wide Subfamilies of LINE-1 Repetitive Sequences». *Journal of Molecular Biology* 246 (3): 401-17. https://doi.org/10.1006/jmbi.1994.0095.

Stetson, Daniel B., Joan S. Ko, Thierry Heidmann, y Ruslan Medzhitov. 2008. «Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity». *Cell* 134 (4): 587-98. https://doi.org/10.1016/j.cell.2008.06.032.

Stratton, Michael R., Peter J. Campbell, y P. Andrew Futreal. 2009. «The cancer genome». *Nature* 458 (7239): 719-24. https://doi.org/10.1038/nature07943.

Suarez, Nicole A., Angela Macia, y Alysson R. Muotri. 2018. «LINE-1 Retrotransposons in Healthy and Diseased Human Brain». *Developmental Neurobiology* 78 (5): 434-55. https://doi.org/10.1002/dneu.22567.

Sultana, Tania, Dominic van Essen, Oliver Siol, Marc Bailly-Bechet, Claude Philippe, Amal Zine El Aabidine, Léo Pioger, et al. 2019. «The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-Insertion Sequence Biases and Post-Insertion Selection». *Molecular Cell* 74 (3): 555-570.e7. https://doi.org/10.1016/j.molcel.2019.02.036.

Tarailo-Graovac, Maja, y Nansheng Chen. 2009. «Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences». *Current Protocols in Bioinformatics* 25 (1): 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

Thayer, R. E., M. F. Singer, y T. G. Fanning. 1993. «Undermethylation of Specific LINE-1 Sequences in Human Cells Producing a LINE-1 -Encoded Protein». *Gene* 133 (2): 273-77. https://doi.org/10.1016/0378-1119(93)90651-I.

Thomas, Charles A., Leon Tejwani, Cleber A. Trujillo, Priscilla D. Negraes, Roberto H. Herai, Pinar Mesci, Angela Macia, Yanick J. Crow, y Alysson R. Muotri. 2017. «Modeling of TREX1-Dependent Autoimmune Disease Using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation». *Cell Stem Cell* 21 (3): 319-331.e8. https://doi.org/10.1016/j.stem.2017.07.009.

Tischler, German, y Steven Leonard. 2014. «biobambam: tools for read pair collation based algorithms on BAM files». *Source Code for Biology and Medicine* 9 (junio): 13. https://doi.org/10.1186/1751-0473-9-13.

Tollefsbol, Trygve O. 2017. «An Overview of Epigenetics». En *Handbook of Epigenetics*, 1-6. Elsevier. https://doi.org/10.1016/B978-0-12-805388-1.00001-8.

Tubio, Jose M. C., Yilong Li, Young Seok Ju, Inigo Martincorena, Susanna L. Cooke, Marta Tojo, Gunes Gundem, et al. 2014. «Mobile DNA in Cancer. Extensive Transduction of Nonrepetitive DNA Mediated by L1 Retrotransposition in Cancer Genomes». *Science (New York, N.Y.)* 345 (6196): 1251343. https://doi.org/10.1126/science.1251343.

Vidal, E., S. Sayols, S. Moran, A. Guillaumet-Adkins, M. P. Schroeder, R. Royo, M. Orozco, et al. 2017. «A DNA Methylation Map of Human Cancer at Single Base-Pair Resolution». *Oncogene* 36 (40): 5648-57. https://doi.org/10.1038/onc.2017.176.

Vu, Ha, y Jason Ernst. 2022. «Universal annotation of the human genome through integration of over a thousand epigenomic datasets». *Genome Biology* 23 (1): 9. https://doi.org/10.1186/s13059-021-02572-z.

Wang, Yong, y Nicholas E. Navin. 2015. «Advances and Applications of Single-Cell Sequencing Technologies». *Molecular Cell* 58 (4): 598-609. https://doi.org/10.1016/j.molcel.2015.05.005.

USC DE SANTIAGO DE COMPOSTELA

Watson, J. D., y F. H. C. Crick. 1953. «Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid». *Nature* 171 (4356): 737-38. https://doi.org/10.1038/171737a0.

Weber, B., S. Kimhi, G. Howard, A. Eden, y F. Lyko. 2010. «Demethylation of a LINE-1 Antisense Promoter in the CMet Locus Impairs Met Signalling through Induction of Illegitimate Transcription». *Oncogene* 29 (43): 5775-84. https://doi.org/10.1038/onc.2010.227.

Weirather, Jason L., Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, y Kin Fai Au. 2017. «Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis». F1000Research. https://doi.org/10.12688/f1000research.10571.2.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. «A Unified Classification System for Eukaryotic Transposable Elements». *Nature Reviews Genetics* 8 (12): 973-82. https://doi.org/10.1038/nrg2165.

Xi, Yuanxin, y Wei Li. 2009. «BSMAP: whole genome bisulfite sequence MAPping program». *BMC Bioinformatics* 10 (1): 232. https://doi.org/10.1186/1471-2105-10-232.

Xing, Jinchuan, Yuhua Zhang, Kyudong Han, Abdel Halim Salem, Shurjo K. Sen, Chad D. Huff, Qiong Zhou, et al. 2009. «Mobile Elements Create Structural Variation: Analysis of a Complete Human Genome». *Genome Research* 19 (9): 1516-26. https://doi.org/10.1101/gr.091827.109.

Xu, Liu, y Masahide Seki. 2020. «Recent Advances in the Detection of Base Modifications Using the Nanopore Sequencer». *Journal of Human Genetics* 65 (1): 25-33. https://doi.org/10.1038/s10038-019-0679-0.

Yang, Allen S., Marcos R. H. Estécio, Ketan Doshi, Yutaka Kondo, Eloiza H. Tajara, y Jean-Pierre J. Issa. 2004. «A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements». *Nucleic Acids Research* 32 (3): e38. https://doi.org/10.1093/nar/gnh032.

Yang, Fang, y P. Jeremy Wang. 2016. «Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline». *Seminars in cell & developmental biology* 59 (noviembre): 118-25. https://doi.org/10.1016/j.semcdb.2016.03.001.

Yuen, Zaka Wing-Sze, Akanksha Srivastava, Runa Daniel, Dennis McNevin, Cameron Jack, y Eduardo Eyras. 2021. «Systematic Benchmarking of Tools for CpG Methylation Detection from Nanopore Sequencing». *Nature Communications* 12 (1): 3438. https://doi.org/10.1038/s41467-021-23778-6.

Zhang, Xiang H.-F., y Lawrence A. Chasin. 2006. «Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons». *Proceedings of the National Academy of Sciences of the United States of America* 103 (36): 13427-32. https://doi.org/10.1073/pnas.0603042103.

Zhang, Xiao, Rui Zhang, y Jinpu Yu. 2020. «New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation». *Frontiers in Cell and Developmental Biology* 8. https://www.frontiersin.org/articles/10.3389/fcell.2020.00657.

Zhao, Ke, Juan Du, Xue Han, John L. Goodier, Peng Li, Xiaohong Zhou, Wei Wei, et al. 2013. «Modulation of LINE-1 and Alu/SVA Retrotransposition by Aicardi-Goutières Syndrome-Related SAMHD1». *Cell Reports* 4 (6): 1108-15. https://doi.org/10.1016/j.celrep.2013.08.019.

Zhou, Weichen, Sarah B. Emery, Diane A. Flasch, Yifan Wang, Kenneth Y. Kwan, Jeffrey M. Kidd, John V. Moran, y Ryan E. Mills. 2020. «Identification and Characterization of Occult Human-Specific LINE-1 Insertions Using Long-Read Sequencing Technology». *Nucleic Acids Research* 48 (3): 1146-63. https://doi.org/10.1093/nar/gkz1173.

Zhu, Shijia, John Beaulaurier, Gintaras Deikus, Tao P. Wu, Maya Strahl, Ziyang Hao, Guanzheng Luo, et al. 2018. «Mapping and Characterizing N6-Methyladenine in Eukaryotic Genomes Using Single-Molecule Real-Time Sequencing». *Genome Research* 28 (7): 1067-78. https://doi.org/10.1101/gr.231068.117.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANEXO I. APROBACIÓN DEL COMITÉ DE ÉTICA

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral Técnica

Secretaría Técnica
Comité Autonómico de Ética da Investigación de Galicia
Secretaría Xeral. Consellería de Sanidade
Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Tel: 881546425. Correo-e: ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ DE ÉTICA DE LA INVESTIGACIÓN DE SANTIAGO-LUGO

Cristina Márquez Riveras, Secretaria del Comité de Ética de la Investigación de Santiago-Lugo,

CERTIFICA:

Que este Comité evaluó en su reunión del día 22/01/19 el estudio:

Título: The impact of L1-mediated somatic rearrangements in the origin and development of human cancer
Versión:
Promotor/a:José Manuel Castro Tubío
Investigador/a: José Manuel Castro Tubío
Código de Registro:2018/586

Y que este Comité, tomando en consideración la pertinencia del estudio, el conocimiento disponible, los requisitos legales aplicables y los Procedimientos Normalizados de Trabajo del Comité, emite un dictamen **FAVORABLE** para la realización del citado estudio.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Documento firmado digitalmente por:
Márquez Riveras, Cristina: 31/01/2019 10:39
SAOC-Z4G4-BOAH-CFHK-6M15-4892-7577-549

Y HACE CONSTAR QUE:

1.- El Comité Territorial de Ética de la Investigación de Santiago-Lugo cumple tanto en su composición como en sus PNTs los requisitos legales vigentes.

2.- La composición actual del Comité Territorial de Ética de la Investigación de Santiago-Lugo es:

- **Juan Manuel Vázquez Lago (Presidente).** Médico especialista en Medicina Preventiva y Salud Pública. Área de Gestión Integrada de Santiago.
- **Pilar Rodríguez Ledo (Vicepresidenta).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Cristina Márquez Riveras (Secretaria).** Enfermera. Dirección Xeral de Saúde Pública.
- **Lorenzo Armenteros del Olmo (Secretario Suplente).** Médico especialista en Medicina Familiar y Comunitaria. Área de Gestión Integrada de Lugo.
- **Francisco Campos Pérez.** Biólogo. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Rosana Castelo Domínguez.** Farmacéutica de Atención Primaria. Área de Gestión Integrada de Santiago.
- **Ricardo García Martínez.** Licenciado en Derecho. Área de Gestión Integrada de Lugo.
- **Jaime Gulín Dávila.** Farmacéutico especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Lugo.
- **María Jesús Lamas Díaz.** Farmacéutica especialista en Farmacia Hospitalaria. Área de Gestión Integrada de Santiago.
- **Guillermo José Prada Ramallal** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago. Fundación Instituto de Investigación Sanitaria de Santiago de Compostela.
- **Carlos Rodríguez Moreno.** Médico especialista en Farmacología Clínica. Área de Gestión Integrada de Santiago.
- **Sandra Vidal Martínez.** Enfermera. Área de Gestión Integrada de Santiago
- **Rafael Carlos Vidal Pérez.** Médico especialista en Cardiología. Área de Gestión Integrada de Lugo.

Para que conste donde proceda, y a petición de quien proceda, en Santiago de Compostela,

La Secretaria del Comité Territorial de Ética de la Investigación de Santiago Lugo,

Cristina Márquez Riveras

Documento firmado digitalmente por:
Márquez Riveras, Cristina: 31/01/2019 10:39
SAOC-Z4G4-BOAH-CFHK-6M15-4892-7577-549

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ANEXO II. PERMISOS DE REPRODUCCIÓN DE IMÁGENES Y GRÁFICAS

Todas las imágenes y gráficas de esta tesis son de autoría propia excepto las indicadas en la siguiente tabla.

Figura	Página	Tipo	Artículo / fuente de la imagen	Permiso reproducción
2: Clases de transposones.	34	Adaptación de esquema de libro	Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics	Permiso CCC marketplace 1366632
5: Proceso de retrotransposición de un LINE1.	37	Adaptación de esquema de artículo científico	Regulation of LINE-1 in mammals	Open Access Creative Commons CCC3
6: Modelos de inserción de un LINE1.	39	Adaptación de esquema de artículo científico	Title: Extensive mosaicism by somatic L1 retrotransposition in normal human cells	Open Access Creative Common CCC4
8: Evolución de las células cancerosas.	43	Adaptación de esquema de artículo científico	The cancer genome	Permiso CCC marketplace 1366667
9: “Hallmarks of cancer”.	43	Adaptación de esquema de artículo científico	Hallmarks of cancer: the next generation	Permiso CCC marketplace 1366554
11: Mecanismos de silenciamiento del LINE1 y su mantenimiento.	48	Adaptación de esquema de artículo científico	Regulation of transposable elements by DNA modifications	Permiso CCC marketplace 1409518
14: Actividad del LINE1.	52	Adaptación de esquema de artículo científico	Regulation of transposable elements by DNA modifications	Permiso CCC marketplace 1409518
24: Secuenciación por nanoporos.	63	Imagen	Oxford Nanopore Technologies Media	Permiso explicito del fabricante para su reproducción
25: Dispositivos de secuenciación de Oxford Nanopore Technologies.	63	Imagen	Oxford Nanopore Technologies media	Permiso explicito del fabricante para su reproducción
26: “The Cost of Sequencing a Human Genome”.	64	Gráfica	www.genome.gov	Dominio público
27: Cánceres con alta tasa de inserciones somáticas de retrotransposones.	65	Adaptación de gráfica de artículo científico	Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition	Open acces CC by
29: Espectrofotómetro One.		Imagen	Thermofisher media	Imagen de uso permitido por el fabricante
34: Secuenciador NGS miseq de Illumina	85	Imagen	Illumina media	Imagen de uso permitido por el fabricante
59: Fragmentación del ADN con el Covaris G-Tube.	123	Imagen	Covaris	Permiso explicito del fabricante para su reproducción
116: Circos plot con información de la línea celular NCI-H2087.	187	Gráfica	https://cancer.sanger.ac.uk/cosmic	Libre para usos no comerciales
117: Circos plot con información de la línea celular NCI-H2009.	188	Gráfica	https://cancer.sanger.ac.uk/cosmic	Libre para usos no comerciales
163: Comparación de epialelos informativos y no informativos de LINE1.	255	Gráfica	Comprehensive locus-specific L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites.	CC-BY 4.0 International license

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

El genoma humano es altamente repetitivo, en gran parte debido a elementos móviles como los retrotransposones LINE1, los únicos autónomos y activos en nuestro genoma con más de 500.000 copias. Sus patrones de metilación pueden afectar a su actividad y suelen estar destregulados en ciertas enfermedades y cáncer. Debido a la naturaleza repetitiva de los LINE1, su gran longitud y similitud, el estudio de esos patrones representa todo un desafío.

En esta Tesis se caracterizan los patrones de metilación de los LINE1 en genomas humanos tumorales y no tumorales, respondiendo cuestiones biológicas sobre su actividad. Para ello hacemos uso de la tecnología de secuenciación de nanoporos 'single molecule sequencing' la cual permite de forma nativa obtener lecturas de gran longitud con información de su estado de metilación.