INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Martín
Santamarina García

PhD Thesis

Sources of Genomic Instability in the Cancer Genome

Santiago de Compostela, 2024

Doctoral Programme in Molecular Medicine

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DOCTORAL THESIS

SOURCES OF GENOMIC INSTABILITY IN THE CANCER GENOME

Author

Martín Santamarina García

Supervisor: Jose M. C. Tubio

Tutor: Jose M. C. Tubio

PHD PROGRAMME IN MOLECULAR MEDICINE

SANTIAGO DE COMPOSTELA

TABLE OF CONTENTS

AGRADECIMIENTOS 6

ACKNOWLEDGMENTS 8

ABBREVIATIONS AND ACRONYMS 10

SUMMARY 14

RESUMO 20

INTRODUCTION 26

OBJECTIVES 30

CHAPTER I: SOURCES OF L1 RETROTRANSPOSITION AND THEIR IMPACT IN THE HUMAN CANCER GENOME

C1.1 - BACKGROUND 32

C1.1.1 – On the biology of transposable elements 32

C1.1.2 - Impact of L1 retrotransposition in cancer 36

C1.2 - RESULTS AND DISCUSSION 43

C1.2.1 - Analysis of FL-L1 sources in the human germline 43

C1.2.2 - New computational methods to study L1 retrotransposition with long reads 47

C1.2.3 - Dynamics of somatic retrotransposition in the light of long reads 57

C1.3 – CONCLUSIONS OF CHAPTER I 67

C1.4 - PUBLICATIONS 67

CHAPTER II: THE IMPACT OF GENOMIC VARIATION IN THE EVOLUTION OF BIVALVE'S TRANSMISSIBLE NEOPLASIA

C2.1 - BACKGROUND 69

C2.1.1 - Transmissible cancer in bivalves 69

C2.1.2 - *Cerastoderma edule* bivalve transmissible neoplasia (CedBTN) 72

C2.1.3 - Genetic characterisation of bivalve transmissible neoplasia 75

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.2 - RESULTS AND DISCUSSION 80
C2.2.1 - Impact of structural variation in the genome of CedBTN 80
C2.2.2 – Clonal structure and phylogenetic composition of CedBTN 84
C2.2.3 - Candidate drivers of CedBTN genome evolution 92
C2.3 – CONCLUSIONS OF CHAPTER II 102
C2.4 - PUBLICATIONS 103

CHAPTER III: PREDICTION OF MICROSATELLITE INSTABILITY IN COLORECTAL CANCER

C3.1 - BACKGROUND 105
C3.1.1 - Microsatellite DNA sequences 105
C3.1.2 - Sources of microsatellite instability (MSI) 105
C3.1.3 – Role of MSI in human cancer and immunotherapy 107
C3.1.4 - Predictors of the MSI phenotype 109
C3.2 - RESULTS AND DISCUSSION 110
C3.2.1 - Training and optimization of a methylation-based MSI predictor 110
C3.2.2 - Biological features of the MSIMEP classifier 112
C3.2.3 - External validation of MSIMEP classifier 114
C3.2.4 - Comparison of MSIMEP with a MLH1 promoter methylation-based model 117
C3.3 – CONCLUSIONS OF CHAPTER III 118
C3.4 - PUBLICATIONS 118

CONCLUSIONS 119

CONCLUDING REMARKS 119
Conclusion from Chapter I 119
Conclusion from Chapter II 119
Conclusion from Chapter III 120

METHODOLOGY 121

METHODOLOGY OF CHAPTER I 121
HGSVC2 cohort 121
Mobile element insertion calling and MEIGA-PAV annotation 121

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Phylogenetic analysis and age estimation for active FL-L1s 122
Subfamily assignation and ORF status determination for active FL-L1s 122
MEIGA cohort 123
DNA isolation and library preparation 123
Long-read whole-genome sequencing 123
Reconstruction of source L1 elements 124
Somatic retrotransposition calling on long-reads 124
Source inference of solo-L1 retrotransposition 124
Allelic tracing of somatic L1 retrotransposition 125
Identification of MEIGA hot L1 elements and L1 phylogenetic analysis 126
Profiling of polyadenylation signals 126
Simulation of retrotransposition events 127
METHODOLOGY OF CHAPTER II 127
Sample collection, processing and diagnosis 127
DNA isolation and sequencing 128
Computational sequence read alignment 128
Selection of a high-purity tumour set 128
Calling, filtering and annotation of nuclear SNVs and indels 129
Calling, filtering and annotation of structural variants 130
Copy number inference 130
Phylogenetic inference from nuclear variants 131
Identification of candidate driver mutations 131
METHODOLOGY OF CHAPTER III 132
TCGA cohorts 132
External cohorts 133
TCGA datasets annotation 134
Analysis of DNA methylation arrays 134
Machine-learning predictive models 134
Model description 135
MSIMEP exploratory analysis in alternative TCGA cohorts 135
MSIMEP validation in external colorectal cancer cohorts 135
Deconvolution of immune cell populations 136

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

BIBLIOGRAPHY 137

APPENDIX 154

APPENDIX A: LIST OF PUBLICATIONS 154

Manuscript 1 154

Manuscript 2 156

Manuscript 3 158

APPENDIX B: LIST OF FIGURES AND TABLES 161

Index of main figures 161

Index of supplementary figures 162

Index of main tables 163

Index of supplementary tables 164

APPENDIX C: SUPPLEMENTARY MATERIAL 165

Supplementary figures 165

Supplementary tables 185

APPENDIX D: DECLARATIONS 194

Declaration of conflicts of interest 194

Declaration of image use rights 194

APPENDIX E: FUNDING 199

APPENDIX F: LIST OF COLLABORATORS 200

Chapter 1 contributors: 200

Chapter 2 contributors: 202

Chapter 3 contributors: 203

Other relevant contributors 203

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

AGRADECIMIENTOS

A todas aquellas personas que habéis contribuido a que consiguiese realizar este trabajo de investigación, quiero que seáis conscientes de mi gratitud por hacer que esta tesis doctoral sea finalmente posible.

En primer lugar, quiero dar mi más profundo agradecimiento a mi familia, por apoyarme durante todos estos años sin importar las circunstancias. Gracias, *Papá*, *Mamá* y *Sara*, por vuestro cariño, paciencia, comprensión y apoyo incondicional, desde siempre, y muy especialmente durante esta etapa de mi vida. Gracias elmito por tu compañía. Gracias a todos y cada uno de vosotros, *Santamarina* y *García*, por ser como sois, y muy en especial a mis dos abuelas, *Suca* y *Mari*, que habéis sido y seréis siempre un ejemplo y la mayor fuente de inspiración para intentar dar lo mejor de mí.

En segundo lugar, quiero agradecer a mis compañeros de *Mobile Genomes*. Gracias por ser realmente eso, compañeros, algo que a veces se asume, pero no siempre es fácil de conseguir. Gracias por tratar de construir un ambiente sano para trabajar en investigación, por vuestra perseverancia, y por apoyarnos durante los altibajos, sobre todo en los momentos difíciles que ha atravesado el laboratorio. En especial, muchas gracias a quienes habéis compartido conmigo estos últimos años de tesis, trabajando en circunstancias a veces desalentadoras. Gracias por vuestros ánimos, cariño y amistad. Gracias por ayudarme a no olvidarme de lo importante, y por -literalmente- prohibirme tirar la toalla. Posiblemente no estaría escribiendo estas páginas de no haber contado con vuestro apoyo.

Por supuesto, gracias también a mis amigos *Pirilampos*, por vuestra amistad incondicional, por vuestros sabios consejos, por vuestra lealtad y también, por vuestra paciencia conmigo. Sois tan cabezotas que a veces me sacáis de quicio, pero pese a ello me siento muy afortunado de teneros. Muchas gracias también a mis amigos de *Biología*, por vuestro apoyo, vuestro cariño y por los recuerdos compartidos de una bonita etapa. Gracias a mis compañeros de *equipo de tenis* USC, por estar conmigo en mil batallas, por confiar en mí y por ayudarme a no perderme a mí mismo cuando lo más fácil era sencillamente perder el norte.

Volviendo al ámbito científico, quiero agradecer a todas las personas que me habéis enseñado y ayudado durante estos años, permitiendo que me formase como investigador. Soy consciente de que tengo todavía muchísimo por aprender y mejorar, pero sin vuestro apoyo, no habría podido realizar este trabajo. Esto incluye en un lugar destacado a mis *mentores* y *mentoras* de biología computacional, supervisores y colaboradores que durante estos años me habéis dado ejemplo y una gran oportunidad para formarme como científico, aprendiendo a afrontar los retos y problemas de la investigación en el ámbito de la genómica del cáncer. Gracias por confiar en mí y darme la oportunidad de trabajar en proyectos para los que, sinceramente, inicialmente no me sentía capacitado. Gracias por vuestro entusiasmo, paciencia y compromiso para llevarlos a cabo. Con algunos de vosotros he compartido no solo quebraderos de cabeza en ciencia, sino también verdadera amistad y bonitas experiencias (incluyendo alguna que otra aventura trepidante por la costa gallega). Gracias también a los compañeros investigadores del *Transmissible Cancer Group*, por acogerme en vuestro laboratorio, y por contribuir a mi formación durante una de las etapas más bonitas, a la par que desafiantes, que tuvo este doctorado. He tenido la oportunidad de formarme y aprender de personas realmente increíbles, asistir a conferencias fascinantes y trabajar en proyectos realmente especiales, y en ese sentido

he sido muy privilegiado. Gracias también aquellos colaboradores que me han facilitado recursos para investigar, y en especial a los que lo han hecho de manera desinteresada. Y disculpas a aquellos que me han dedicado su tiempo en cuestiones que finalmente no he sido capaz de resolver. Sé que durante este tiempo he dejado alguna cuenta pendiente, y espero que con el tiempo dichos proyectos puedan ser finalizados.

Por último, pero no por ello menos importante, quiero dar mi mayor agradecimiento a mi director de tesis, *Jose*, el principal responsable de las oportunidades que acabo de mencionar, para mí y para muchos otros de mis compañeros durante el doctorado. Gracias por confiar en mí. Gracias por tu determinación, por creer en tu equipo, y por poner tanta pasión y entrega en nuestro trabajo científico a lo largo de este camino. Gracias por apostar por un modelo de investigación colaborativo que no se olvida de respetar a las personas. Es posible que hayamos tenido algunas diferencias a lo largo del camino, pero no olvidaré tu compromiso, generosidad y entrega a la hora de apoyar a tus estudiantes y sus proyectos.

Si bien estas páginas pueden suponer en fin de una etapa (y probablemente por ello priman los buenos recuerdos), quiero dejar claro que el camino ha sido extremadamente duro (para mí y para otros compañeros doctorandos). Bajo mi opinión, el programa de formación en investigación que representa el doctorado debería ser actualizado, de la mano de una reforma de los criterios de evaluación del mérito y la competencia en el ámbito universitario y académico. Los estudiantes de doctorado actualmente presentan como uno de los colectivos con mayor prevalencia de problemas de salud mental en la población [1, 2]. Entre los trastornos más comunes en el ámbito biomédico se encuentran el *burnout*, la ansiedad y la depresión [3]. Espero que esta declaración contribuya a visibilizar esta situación, y al desarrollo de estrategias efectivas para prevenir el deterioro de la salud mental de los estudiantes de doctorado. Creo firmemente que esto constituye una realidad que debe ser afrontada en las diferentes etapas del sistema de educación pública. Creo que debe ser afrontado desde un enfoque integral que tenga en especial consideración a aquellas personas con menos recursos y/o redes de apoyo familiar, económico y social.

Referencias:

1. Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J. & Gisle, L. Work organization and mental health problems in PhD students. *Res. Policy* 46, 868–879 (2017).
2. Hazell, C. M., Chapman, L., Valeix, S. F., Roberts, P., Niven, J. E. & Berry, C. Understanding the mental health of doctoral researchers: a mixed methods systematic review with meta-analysis and meta-synthesis. *Syst. Rev.* 9, 197 (2020).
3. Nagy, G. A., Fang, C. M., Hish, A. J., Kelly, L., Nicchitta, C. V., Dzirasa, K. & Rosenthal, M. Z. Burnout and mental health problems in biomedical doctoral students. *CBE Life Sci. Educ.* 18, ar27 (2019).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ACKNOWLEDGMENTS

To all those who somehow contributed to making this research work possible, I want you to be aware of my gratitude for making this doctoral thesis finally real.

First, I want to express my deepest gratitude to my family for supporting me throughout these years, no matter the circumstances. Thank you, *Dad*, *Mom*, and *Sara*, for your unconditional affection, patience, understanding and support, always and especially during this stage of my life. Thank you, elmito, for your companion. Thanks to every one of you, *Santamarina* and *García*, for being the way you are, and especially to my two grandmothers, *Suca* and *Mari*, who were and will always be an example and the greatest source of inspiration for me to strive to give my best.

Secondly, I want to thank my colleagues at *Mobile Genomes*. Thank you for truly being companions, something that is often assumed but not always easy to achieve. Thank you for striving to create a healthy environment for research, for your perseverance, and for supporting one another through the ups and downs, especially during the difficult times the lab has faced. A special thanks to those of you who have shared these last years of my doctoral work, persevering through sometimes discouraging circumstances. Thank you for your encouragement, kindness, and friendship. Thank you for helping me focus on what's important and for -literally- forbidding me to give up. I likely wouldn't be writing these pages without your support.

Of course, I also want to thank my friends *Pirilampos* for your unconditional friendship, your advice, loyalty, and patience with me. Your stubbornness sometimes drives me a little crazy, but despite that, I feel so fortunate to have you in my life. Many thanks as well to my *Biology* friends for your support, affection, and the shared memories of a beautiful time. Thanks to my USC *tennis team* colleagues for standing by me in countless battles, trusting me, and helping me stay grounded when it would have been easiest to lose my way.

Returning to the scientific sphere, I wish to thank everyone who taught and supported me during all these years, enabling me to grow as a researcher. I know I still have much to learn and improve, and I also know that without your guidance, I couldn't have carried out this work. This prominently includes my *mentors* in computational biology, supervisors, and collaborators who set a true example and offered me incredible opportunities to learning how to tackle the challenges and problems in cancer genomics research. Thank you for trusting me and giving me the chance to work on great projects that, honestly, I initially felt unprepared for. Thank you for your enthusiasm, patience and commitment. With some of you, I've shared not only science but also true friendship (including memorable adventures along the Galician coast).

Thanks also to the researchers at the *Transmissible Cancer Group* for welcoming me into your lab and contributing to my development as a researcher during one of the most rewarding and challenging stages of this doctorate. I've been fortunate to train and learn from truly amazing people, attend fascinating conferences, and work on special projects. In that sense, I've been very privileged. Thanks also to collaborators who provided resources for my research, especially those who did so selflessly. My apologies to those who devoted their time to projects I couldn't complete. I know I've left some unfinished work behind, and I hope those projects can be completed in time.

USC universidade de santiago

Last but not least, my heartfelt thanks to my thesis director, *Jose*, the main enabler of the opportunities I've mentioned, not only for me but for many others during this doctorate. Thank you for believing in me. Thank you for your determination, for trusting your team, and for putting so much passion and dedication into our scientific work along this journey. Thank you for pursuing a collaborative research model that respects people. We may have had some differences along the way, but I won't forget your commitment, generosity and dedication when supporting your students and their projects.

While these pages might mark the end of a chapter (and perhaps for that reason, the good memories prevail), I want to emphasize that this journey has been extremely tough (for me and for other doctoral candidates). In my view, the training and research program that constitutes a doctorate should be updated, alongside a reform of the evaluation criteria for merit and competence in academia. Doctoral students are currently among the groups with the highest prevalence of mental health issues in the population [1, 2]. Common disorders in the biomedical field include burnout, anxiety, and depression [3]. I hope this statement helps raise awareness of this situation and fosters the development of effective strategies to prevent mental health deterioration among doctoral students. I firmly believe this is a reality that must be addressed throughout the public education system. It should be tackled with a comprehensive approach that pays special attention to those with fewer resources and/or family, economic, and social support networks.

References:

1. Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden , J. & Gisle , L. Work organization and mental health problems in PhD students. *Policy Res* 46, 868–879 (2017).
2. Hazell, CM, Chapman, L., Valeix, SF, Roberts, P., Niven, JE & Berry, C. Understanding the mental health of doctoral researchers: a mixed methods systematic review with meta-analysis and meta-synthesis. *Syst. Rev.* 9, 197 (2020).
3. Nagy, GA, Fang, CM, Hish, AJ, Kelly, L., Nicchitta, CV, Dzirasa , K. & Rosenthal, MZ Burnout and mental health problems in biomedical doctoral students. *CBE Life Sci. Educ.* 18, ar27 (2019).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABBREVIATIONS AND ACRONYMS

AUC	Area under the curve
BAM	Binary alignment map
BFB	Breakage-fusion-bridge
BND	Breakend
Bp	Base pair
BWA	Burrows-wheeler aligner
CedBTN	*Cerastoderma edule* bivalve transmissible neoplasia
CIMP	CpG island methylator phenotype
CN	Copy number
COAD	Colon adenocarcinoma
COSMIC	Catalogue of somatic mutations in cancer
CRC	Colorectal cancer
CTVT	Canine transmissible venereal tumour
DEL	Deletion
DFTD	Devil facial tumour disease
DNA	Deoxyribonucleic acid
dN/dS	Ratio of substitution rates at non-synonymous and synonymous sites
DMP	Differentially methylated CpG site/probe
DUP	Duplication
ERV	Endogenous retrovirus
FL-L1	Full-length LINE-1
GLM	Generalized linear model
GSEA	Gene set enrichment analysis
HGSCV	Human genome structural variation consortium
ICGC	International cancer genome consortium
IGV	Integrative genomics viewer
INDEL	Insertion / deletion
INV	Inversion
KB	Kilobase
KEGG	Kyoto encyclopedia of genes and genomes
KNN	K-nearest neighbours
LINE	Long interspersed nuclear element
L1Hs	L1 *Homo sapiens*
L1Pt	L1 *Pan troglodytes*
LOH	Loss of heterozygosity
LTR	Long terminal repeat
LUSC	Lung squamous cell carcinoma
MarBTN	*Mya arenaria* bivalve transmissible neoplasia
MB	Megabase
Myr	Million years
MEI	Mobile element insertion
MEIGA	Mobile element insertion genome analyser
ML	Maximum likelihood
MMR	DNA mismatch repair
MNV	Multi-nucleotide variant

USC

MP	Maximum parsimony
MSI	Microsatellite instability
MSIMEP	Microsatellite instability methylation-based predictor
MSS	Microsatellite stability
mRNA	Messenger RNA
ONT	Oxford nanopore technologies
ORF	Open reading frame
PCAWG	Pan-cancer analysis of whole genomes
PCA	Principal component analysis
PCR	Polymerase chain reaction
PON	Panel of normals
RAG	Recombination activating gene
RFE	Recursive feature elimination
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
SINE	Short interspersed nuclear element
SNV	Single nucleotide variant
STAD	Stomach adenocarcinoma
SVM	Support vector machine
SV	Structural variant
SVA	SINE-VNTR-Alus
TCGA	The cancer genome atlas
TFBS	Transcription factor binding site
TPRT	Target primed reverse transcription
TraFiC	Transposon finder in cancer
TSD	Target site duplication
TSG	Tumour suppressor Gene
UCEC	Uterine corpus endometrial carcinoma
UTR	Untranslated region
VAF	Variant allele frequency
VCF	Variant call Format
VEP	Variant Effect Predictor
WGA	Whole Genome Amplification
WGS	Whole Genome Sequencing
WHO	World Health Organization

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Gene symbols

Symbol	Description
ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1
AKT	AKT serine/threonine kinase
ALOX5	Arachidonate 5-lipoxygenase
APC	APC regulator of WNT signalling pathway
AXIN2	Axin 2
BIN1	Bridging integrator 1
BMP2	Bone morphogenetic protein 2
BRCA2	BRCA2 DNA repair associated
CASP3	Caspase 3
CCDN1	Cyclin D1
CCND3	Cyclin D3
CDKN2A	Cyclin dependent kinase inhibitor 2A
CDKN2B	Cyclin dependent kinase inhibitor 2B
CIC	Capicua transcriptional repressor
CTLA4	Cytotoxic T-lymphocyte associated protein 4
CYP11A1	Cytochrome P450 family 11 subfamily A member 1
DNM2	Dynamin 2
EGFR	Epidermal growth factor receptor
EPCAM	Epithelial cell adhesion molecule
ERBB2	Erb-B2 receptor tyrosine kinase 2
EXT1	Exostosin glycotransferase 1
F8	Coagulation factor VIII
GPR160	G protein-coupled receptor 160
HSP	Heat Shock Protein
IDO	Indoleamine 2,3-dioxygenase
KEAP1	Kelch like ECH associated protein 1
KMT2D	Lysine methyltransferase 2D
KRAS	KRAS proto-oncogene, GTPase
LAG3	Lymphocyte activating 3
MAP3K1	Mitogen-activated protein kinase kinase kinase 1
MCL1	MCL1 apoptosis regulator, BCL2 family member
MDM2	MDM2 proto-oncogene
MEN1	Menin 1
MGMT	O-6-Methylguanine-DNA methyltransferase
MLH1	MutL homolog 1
MSH2	MutS homolog 2
MSH6	MutS homolog 6
mtCOI	Mitochondrially-encoded cytochrome C oxidase
MYC	MYC proto-oncogene, BHLH transcription factor
NBN	Nibrin
NF1	Neurofibromin 1
PAS	Polyadenylation signal
PMS2	PMS2 homolog 2, mismatch repair system component
POLE	DNA polymerase epsilon, catalytic subunit
POLD	DNA polymerase delta, catalytic subunit

USC
DE COMPOSTELA

PPM1D	Protein phosphatase, Mg2+/Mn2+ dependent 1D
PTEN	Phosphatase and tensin homolog
PTPRT	Protein tyrosine phosphatase receptor type T
RAD17	RAD17 checkpoint clamp loader component
RAG	Recombination activating gene
RB1	RB transcriptional corepressor 1
RECQL4	RecQ like helicase 4
SELP	Selectin P
SETD2	SET domain containing 2, histone lysine methyltransferase
STAG2	STAG2 cohesin complex component
STARD9	StAR related lipid transfer domain containing 9
SUFU	SUFU negative regulator of hedgehog signalling
TBX3	T-box transcription factor 3
THBS1	Thrombospondin 1
TP53	Tumour protein P53
TP63	Tumour protein P63
TRIP10	Thyroid hormone receptor interactor 10
TSC1	TSC complex subunit 1TSC2 - TSC complex subunit 2
TSC22D2	TSC22 domain family member
WNK2	WNK lysine deficient protein kinase 2
WRN	WRN RecQ like helicase

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

SUMMARY

Cancer is commonly referred to as a complex group of genetic diseases, encompassing over 100 disorders that affect multiple cell types and tissues in humans and many other multicellular organisms. At its core, cancer is characterized by uncontrolled cell proliferation, driven by populations of transformed tumour cells undergoing evolutionary changes. In humans, cancer is a leading global cause of death, responsible for over 8 million deaths each year. Notably, more than 90% of cancer-related deaths are linked to metastasis, a process through which tumour cells disseminate through the circulatory and lymphatic systems and invade distant organs.

Cancer development arises from the interplay of genetic and environmental factors that promote the process of aberrant cell proliferation, invasion and metastasis. Tumours generally initiate from genetic alterations that impair key molecular pathways such as cell-cycle control, DNA repair and apoptosis. Most of the human tumours originate because of the accumulation of de novo mutations in somatic cells. This genetic damage is acquired during organism development due to intrinsic and extrinsic agents that, ultimately, disrupt the genetic programs of somatic cells. This association between DNA damage accumulation and progressive deterioration of stem cells and somatic tissues defines an intrinsic link between the disease and the biological process of aging.

Genomic instability is a molecular condition marked by an accelerated accumulation of genetic alterations across the genome. This instability can arise from diverse source mechanisms, including defects in DNA repair machinery, defects in chromosome segregation, exposure to environmental factors such as radiation or chemicals, and the dysregulation of endogenous of exogenous mutagenic agents, such as transposable elements and viruses. The resulting genetic alterations range from small-scale changes like single nucleotide variants (SNVs) and small insertion-deletions (INDELs) to larger structural variations. Structural variants (SVs) are larger-scale genetic alterations that involve changes in the structure of DNA, typically affecting sequences larger than 50 base pairs. These alterations include deletions, duplications, insertions, inversions and translocations, which are associated to different patterns of DNA gains, losses and rearrangements.

Genomic instability is a hallmark of cancer, driving genetic diversity and promoting somatic clonal evolution processes that contribute to cancer progression and therapy resistance. Chromosomally unstable tumours are generally refractory to immunotherapy and display molecular features associated with immune evasion. Although most structural variants accumulated in cancer usually have a neutral evolutionary effect, collectively, they represent a major source of genetic variation that fuels tumour development. Structural variants can disrupt, amplify or fuse cancer-related genes, and can modify non-coding DNA regulatory sequences altering gene expression. Multiple cancer genes have been identified to be frequently amplified (e.g. *MCL1*, *EGFR*, *MYC*, *ERBB2*) or deleted (e.g. *CDKN2A/B*, *RB1* and *PTEN*) in cancer cells. According to the results of the Pan-cancer analysis of whole genomes (PCAWG) project, structural variants represent the majority (55%) of the cancer driver events generally found in human cancers, keeping a role that is likely underestimated. Notably, structural variants are often associated with complex repetitive DNA sequences subjected to low mappability, which are difficult to profile under the lens of common short-read sequencing technologies and classic bioinformatic methods. Despite recent advances, genomic instability remains as a largely

unknown aspect of cancer biology, hiding relevant implications for understanding tumorigenesis, metastasis and devising new therapeutic approaches against the disease.

In this thesis, we investigated three distinct mechanisms driving genomic instability in various cancer models. Utilizing a combination of high-throughput sequencing technologies and innovative bioinformatic methods, we aimed to study the genetic composition of tumour genomes. This investigation is intended to uncover new molecular features and mutagenic elements that promote genomic instability and the evolution of the cancer genome.

Chapter I: Sources of L1 retrotransposition and their impact in the human cancer genome

The first mechanism of genomic instability we investigated is LINE-1 (L1) somatic retrotransposition. L1 elements are repetitive mobile DNA sequences that serve as key architects of the mammalian genome and are the principal autonomous retrotransposons that retain activity in the human population. L1 retrotransposition plays a significant mutagenic role in certain human cancers of epithelial origin, such as oesophageal adenocarcinoma (ESAD), head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD). In these tumours, retrotransposition might be commanded by a limited set of active L1 source elements known as "hot L1s" (e.g. 22q12.1, 14q23.1, 6p22.1, 6p24.1 and Xp22.2-1), which trigger extensive bursts of somatic retrotransposition. However, the complete characterisation of L1 activity has been constrained by limitations in the read size of sequencing technologies, preventing the analysis of the internal structure of most L1 sequences, and by detection approaches limited to the detection of 3' transductions – a subtype of L1 insertions that carry downstream sequences mobilized alongside the retrotransposon-.

In collaboration with the HGSCV2 consortium, we investigated the potential of long-read assemblies to characterise the complete sequence of full-length L1 elements (FL-L1s). We analysed a diverse catalogue of 142 FL-L1s which showed evidence of activity in the human germline, but often also in tumour tissues. The analysis of diagnostic nucleotides allowed us to ascribe these elements to specific L1Hs subfamilies. Most of these elements were members from the evolutionary-recent *Ta-1* subfamily (55.6%). However, we also found active FL-L1s from the older *Ta-0* (22.5%) and *pre-Ta* (19.7%) subfamilies, in agreement with studies supporting that retrotransposition is sustained by diverse L1Hs lineages. Taking advantage of the polished internal sequences, we assessed the ORF status for this set of active 142 FL-L1s. We found that most (81%) of these competent elements presented two intact open reading frames (ORF1 and ORF2). However, we identified a notable fraction (19%) of active FL-L1s with disrupted ORFs. Remarkably, this included a hot element at 9q32 reported to be highly active in diverse human tumours. These findings suggests either trans-complementation or polymorphisms leading to the recent demise of certain active copies. We then inferred a ML phylogeny based on the alignment of 142 FL-L1 complete sequences and estimated the approximate age of FL-L1 elements considering divergence from their lineage-specific consensus. *Ta-1* elements generally represented younger copies compared with *Ta-0* and *pre-Ta* elements, which were usually older and less polymorphic. Remarkably, cancer-related FL-L1 elements were distributed across multiple clades on the L1 phylogeny. This included a fixed ancient *pre-Ta* insertion in 1p12 that despite integrating into the human genome ~1.8 Myr ago, retains high activity in the germline and somatically within tumour genomes. Our results suggest that a small set of ancient *pre-Ta* L1 copies might still be relevant for cancer development.

USC DE COMPOSTELA

Motivated by the resolution provided by long-reads in the germline, we contributed to the development of MEIGA, a project conceived to investigate the dynamics of somatic L1 retrotransposition in cancer. We designed two specific bioinformatic tools. The first one is MEIsimulator, is a method to generate mock genomes with diverse retrotransposition events integrated at predefined genomic locations. The second one is Source Inference, a method which traces somatic L1 insertions to its element-of-origin based on shared diagnostic nucleotide variants. We assessed the performance of the Source Inference algorithm using in silico and tumour somatic retrotransposition datasets. While insertion length limited the sensitivity of our approach, this evaluation demonstrated the capacity of Source Inference to accurately determine the origin somatic L1 insertions in cancer.

We applied our Source Inference method to the long-read sequencing data from a selection of 10 MEIGA tumours with high retrotransposition rates. As a result, we were able to trace 1,325 events (899 solo-L1 insertions) to its L1 element-of-origin. These events represented 44.5% (1,325/2,980) of the total number of retrotransposition events traced in MEIGA tumours combining various approaches. The mobilizations traced by Source Inference had a substantial impact in the patterns of activity of many source elements. Notably, 688/899 (76.5%) of the newly traced L1-solos were attributed to source elements of known activity in cancer according to PCAWG studies, including 294 events linked to 14 source elements with known hot activity in cancer genomes, highlighting its mutagenic relevance. In addition to the 151 source elements identified in the MEIGA cohort by their derived transductions, another 46 cryptic elements emerged as donors of somatic solo-L1s under our diagnostic nucleotide's strategy. Overall, the addition of solo-L1s increases in average 1.37-fold the activity of MEIGA hot L1s relative to a transduction strategy. We found that some hot L1 source elements (e.g. 1p22.3, 1q24.1, 12q24.22) were typically mediators of solo-L1s insertions, while others (e.g. Xp22.2-1, 6p24.1, 6p22.1) were usually mediators of transduction events. We found evidence supporting that variation in the relative strength of canonical and alternative polyadenylation signals involved in L1 transcription might determine this pattern. Additionally, for a limited set of homozygous FL-L1s, Source Inference allowed to discriminate the allele of origin for specific L1 somatic insertions, showing the potential of phased long-read sequencing data to unravel the biology of L1 source elements in cancer at high resolution.

Chapter II: The impact of genomic variation in the evolution of bivalves' transmissible neoplasia

Next, we explored the impact of structural variation across the genome of *Cerastoderma edule* bivalve transmissible cancer. Transmissible cancers are somatic tumour lineages that spread between individuals through the direct transfer of cancer cells, evading the host immune system. For this reason, they represent a crucial model to understand the molecular basis of metastatic cancer and the dynamics of long-term somatic evolution.

Over the past decade, bivalve transmissible cancers have been described to affect at least eight bivalve species, including common cockles (*Cerastoderma edule*), soft-shell clams (*Mya arenaria*) and bay mussels (*Mytilus trossulus*), among others. These tumours manifest as a leukaemia-like condition, derived from the anomalous proliferation of transformed haemocytes that spread in the ocean saltwater to infect new hosts. Recent studies have identified some molecular alterations associated with BTN genomes, which seem to be characterized by

USC DE COMPOSTELA

unstable karyotypes with marked hyperploid and disruption of essential pathways controlling cell cycle (e.g. p53 dysregulation). Previous research has proposed a polyphyletic origin of the disease in *Cerastodema edule* bivalve transmissible neoplasia (CedBTN), which might be represented by two independent cancer clones. However, the genetic characterization of CedBTN has been limited to the analysis of a small number of DNA sequences and marker genes. The recent publication of bivalve reference whole-genomes, such as the one assembled within the framework of the Scuba Cancers project for the common cockle, provide a key resource to systematically characterize genome-wide alterations (e.g. SNVs, SVs and CNVs) in the CedBTN nuclear genome and to investigate the somatic evolution of these transmissible cancer lineages.

Within the framework of the Scuba Cancers project, we analysed Illumina whole-genome sequencing data from a selection of 10 high-purity CedBTN tumours and 455 common cockles sampled across the species' geographical range in the Atlantic coast. We called structural variants in tumours using a combination of bioinformatic algorithms. Extensive computational SV genotyping enabled to discriminate between somatic and germline alterations, leading to the identification of a total of 18,272 CedBTN-specific structural variants. We analysed the incidence and chromosomal distribution of different types of structural variants across the CedBTN nuclear genome. Deletions, typically with moderate size, represented the most common structural variant type in the genome of CedBTN (79.8%, 14,589/18,272), followed by duplications and potential interchromosomal rearrangements. We detected multiple instances of large-scale rearrangements including deletions, duplications and inversions impacting multiple chromosomes. Our results highlight a significant degree of structural reorganization in the CedBTN genome.

We leveraged the genome-wide genetic variation identified in the Scuba Cancers project to explore the clonal structure and evolutionary origins of CedBTN. We performed MP and ML phylogenetic inferences derived from complementary sources of genomic variation (e.g. 18,272 SVs and 4.3 million SNVs), which in combination with additional evidence obtained in the project (e.g. dN/dS profiling, divergence against extant cockle populations and satellite expansions), confirmed the co-existence of two ancient, clonally independent tumour lineages (CedBTN1 and CedBTN2), distinguishable by molecular and histological features. To facilitate the diagnosis of CedBTN1 and CedBTN2, we designed a PCR-based assay based on somatic deletion profiling to identify and discriminate both lineages in a rapid and cost-effective manner.

Finally, we performed a systematic search of driver alterations in the genomes of CedBTN1 and CedBTN2. We focused on pre-divergence variants with significant impact on the coding sequence of cancer genes. Our genome-wide screening of alterations identified several instances of cancer gene copies potentially disrupted by somatic variants, some of which might contribute to chromosome instability (e.g. *RB1* and *STAG2* splice and frameshift variants). Unfortunately, the integration of these mutations with gene copy number estimations did not allow to identify high-confident driver events. However, the analysis of copy number data and associated gene expression allowed us to detect candidate early driver alterations in CedBTN. We identified an ancestral homozygous somatic deletion of the tumour suppressor *MGMT*, associated with a mutational signature linked to disrupted repair of alkylated DNA bases in CedBTN1. Additionally, we identified several instances of oncogene amplification *(MDM2, CCND3* and *MYC*), potentially involved in cell-cycle dysregulation and disruption of responses

USC DE COMPOSTELA

against aneuploidy in CedBTN. These and other alterations characterized in the framework of the Scuba Cancers project, such as whole genome doubling and recurrent mtDNA capture, might potentially represent molecular adaptations to sustain the fitness of these transmissible cancer genomes during long-term somatic evolution.

Chapter III: Prediction of microsatellite instability in colorectal cancer

The third mechanism we studied is microsatellite instability (MSI) in colorectal cancer (CRC). Microsatellites are short tandem repeat sequences of DNA, which usually expand or contract due to polymerase slippage and defects the mismatch repair machinery (MMR). Deficiency in MMR genes (e.g. *MLH1*, *MSH2*, *MSH6*, *PMS2*) often leads to the development of a MSI hypermutator phenotype, characterized by a widespread incidence of SNVs or INDELs within these microsatellites.

MSI is common observed in hereditary cancer-associated syndromes (e.g. Lynch syndrome), but also in sporadic solid tumours. MSI tumours are usually associated with a better prognosis and are generally less prone to develop metastasis. This tendency has been associated with an increased neoantigen production in MSI cancer, leading to increased tumour immunogenicity. Remarkably, tumours with hypermutator phenotypes have shown improved responses to anti-CTLA-4 and anti-PD-1 treatments in melanoma and lung cancer, respectively. Additionally, CRC patients with MMR generally present a better prognosis and greater benefit from immunotherapy by PD-1 immune checkpoint blockade. This therapeutic relevance has motivated the development of computational approaches to determine MSI phenotype from different sources of molecular information, ranging from direct microsatellite profiling from whole-genome sequencing data to the inference of MSI from epigenetic and transcriptional profiles.

We leveraged the methylation microarray datasets from 388 CRC tumours from The Cancer Genome Atlas (TCGA) project, aiming to develop a methylation-based MSI predictor. We conducted a differential methylation analysis between MSI and MSS primary tumours, leading to the identification of 780 differentially methylated CpG probes. We used this information to train supervised machine learning models on the same dataset and assessed their performance through iterative rounds of 10-fold cross validation. After comparing 7 alternative algorithms for MSI prediction (GLM, Neural Network, Random Forest, SVM, Decision Trees, KNN and Naïve Bayes classifiers), we trained and optimized an effective (Accuracy=0.98, Kappa=0.92) GLM-based MSI/MSS classifier composed of 400 CpG probes we termed MSIMEP. Additionally, considering a potential applicability to liquid biopsy and assuming a limited sacrifice of 1% of classification accuracy, we developed a reduced version of MSIMEP composed of 25 CpGs probes. We analysed the model importance and biological features of MSIMEP predictors, highlighting relevant CpG probes distributed across both genic and intergenic chromosome regions. Remarkably, a significant portion of the MSIMEP classification capacity is driven by *MLH1* associated probes, underscoring the relevance of the epigenetic inactivation of this MMR gene in the development of the MSI condition. We next evaluated the performance of MSIMEP across diverse cancer cohorts. Although its performance was limited on other cancer types (STAD and UCEC), MSIMEP demonstrated a robust performance (Accuracy=0.93, Kappa=0.83) when evaluated on the pooled external cohort composed of 517 colorectal tumours. Although MSIMEP should be considered as a computational resource which requires further evaluation before considering clinic translation,

USC DE COMPOSTELA

our results highlight the potential of epigenetic-based profiling to determine the MSI condition in CRC tumours.

Concluding remarks

This work has presented key findings from my research within the Molecular Medicine PhD program, focusing on genomic instability in cancer. Collectively, these results contribute modestly to a better understanding of the molecular mechanisms driving genomic instability in cancer. Our findings demonstrate the potential of emerging high-throughput technologies, alongside advanced bioinformatic approaches, to identify and characterize genetic and epigenetic elements involved in genomic structural variation in cancer. The analysis of genetic and epigenetic variants, as well as alterations distributed across large-scale chromosome territories, has been valuable in providing insights into the patterns of somatic evolution in tumours. Although these results fall within the framework of basic research in cancer genomics, further development of these research lines in primary, metastatic, and transmissible cancers may contribute to novel therapeutic advances in the fields of personalized medicine and molecular oncology.

Keywords: cancer, genome, somatic evolution, retrotransposition, structural variants, microsatellite, bioinformatics, transmissible cancer, genomic instability.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

RESUMO

O cancro descríbese habitualmente como un grupo complexo de enfermidades xenéticas que inclúe máis de 100 trastornos que afectan múltiples tipos de células e tecidos en humanos e outros organismos multicelulares. En esencia, o cancro caracterízase pola proliferación celular descontrolada, impulsada por poboacións de células tumorais transformadas suxeitas a cambios evolutivos. Nos humanos, o cancro é unha das principais causas de morte a nivel global, responsable de máis de 8 millóns de mortes anuais. É destacable que máis do 90% das mortes relacionadas co cancro están asociadas á metástase, un proceso polo cal as células tumorais disemínase a través dos sistemas circulatorio e linfático e invaden órganos distantes.

O desenvolvemento do cancro xorde da interacción de factores xenéticos e ambientais que promoven a proliferación celular aberrante, a invasión e a metástase. Os tumores adoitan iniciarse a partir de alteracións xenéticas que afectan rutas moleculares clave como o control do ciclo celular, a reparación do ADN e a apoptose. A maioría dos tumores humanos orixínanse pola acumulación de mutacións *de novo* en células somáticas. Este dano xenético adquírese durante o desenvolvemento do organismo debido a axentes intrínsecos e extrínsecos que, en última instancia, alteran os programas xenéticos das células somáticas. Esta asociación entre a acumulación de dano no ADN e o deterioro progresivo das células nai e tecidos somáticos define un vínculo intrínseco entre a enfermidade e o proceso biolóxico do envellecemento.

A inestabilidade xenómica é unha condición molecular caracterizada pola acumulación acelerada de alteracións xenéticas ao longo do xenoma. Esta inestabilidade pode xurdir de diversos mecanismos, incluídos defectos na maquinaria de reparación do ADN, erros na segregación cromosómica, exposición a factores ambientais como radiación ou substancias químicas, e a desregulación de axentes mutaxénicos endóxenos ou exóxenos, como os elementos transpoñibles e os virus. As alteracións xenéticas resultantes abranguen dende cambios a pequena escala, como variantes dun só nucleótido (SNVs) e pequenas inserción ou delecións (INDELs), ata variacións estruturais de maior tamaño. As variantes estruturais (SVs) son alteracións xenéticas de maior escala que implican cambios na estrutura do ADN, tipicamente en secuencias superiores a 50 pares de bases. Estas alteracións inclúen delecións, duplicacións, inserción, inversións e translocacións, asociadas a distintos patróns de ganancias, perdas e reordenamentos de ADN.

A inestabilidade xenómica é unha característica distintiva do cancro, impulsando a diversidade xenética e promovendo procesos de evolución clonal somática que contribúen á progresión tumoral e á resistencia ás terapias. Os tumores con inestabilidade cromosómica adoitan ser refractarios á inmunoterapia e presentan características moleculares asociadas coa evasión inmune. Aínda que a maioría das variantes estruturais acumuladas no cancro teñen un efecto evolutivo neutro, no seu conxunto representan unha fonte importante de variación xenética que alimenta o desenvolvemento tumoral. As variantes estruturais poden interromper, amplificar ou fusionar xenes relacionados co cancro, así como modificar secuencias reguladoras non codificantes, alterando a expresión xénica. Múltiples xenes relacionados co cancro foron identificados como frecuentemente amplificados (p. ex., *MCL1, EGFR, MYC, ERBB2*) ou delecionados (p. ex., *CDKN2A/B, RB1* e *PTEN*) en células tumorais. Segundo os resultados do proxecto *Pan-Cancer Analysis of Whole Genomes* (PCAWG), as variantes estruturais representan a maioría (55%) dos eventos condutores do cancro identificados en tumores humanos, desempeñando un papel probablemente subestimado. É destacable que as variantes

estruturais adoitan estar asociadas a secuencias de ADN repetitivas complexas e de baixa mapabilidade, o que dificulta a súa análise utilizando tecnoloxías comúns de secuenciación de lectura curta e métodos bioinformáticos clásicos. A pesar dos avances recentes, a inestabilidade xenómica segue a ser un aspecto pouco coñecido da bioloxía do cancro, con implicacións relevantes para comprender a tumorixénese, a metástase e o desenvolvemento de novas estratexias terapéuticas contra a enfermidade.

Nesta tese, investigamos tres mecanismos distintos que impulsan a inestabilidade xenómica en varios modelos de cancro. Utilizando unha combinación de tecnoloxías de secuenciación de alto rendemento e métodos bioinformáticos innovadores, o noso obxectivo é estudar a composición xenética dos xenomas tumorais. Esta investigación pretende descubrir novas características moleculares e elementos mutaxénicos que promovan a inestabilidade xenómica e a evolución do xenoma do cancro.

Capítulo I: Fontes de retrotransposición L1 e o seu impacto no xenoma do cancro humano.

O primeiro mecanismo de inestabilidade xenómica que investigamos é a retrotransposición somática de LINE-1 (L1). Os elementos L1 son secuencias móbiles repetitivas de ADN que actúan como arquitectos clave do xenoma dos mamíferos e constitúen os principais retrotransposóns autónomos que conservan actividade na poboación humana. A retrotransposición de L1 desempeña un papel mutaxénico significativo nalgúns cancros humanos de orixe epitelial, como o adenocarcinoma esofáxico (ESAD), o carcinoma escamoso de cabeza e pescozo (HNSC), o carcinoma escamoso de pulmón (LUSC) e o adenocarcinoma de colon (COAD). Nestes tumores, a retrotransposición pode ser dirixida por un conxunto limitado de elementos L1 activos coñecidos como *"hot L1s"* (por exemplo, 22q12.1, 14q23.1, 6p22.1, 6p24.1 e Xp22.2-1), que desencadean extensas ondas de retrotransposición somática. Porén, a caracterización completa da actividade de L1 viuse limitada polas restricións no tamaño de lectura das tecnoloxías de secuenciación, que dificultan a análise da estrutura interna da maioría das secuencias L1, e por métodos de detección que se limitan a identificar transducións 3', un subtipo de inserción L1 que mobilizan secuencias adxacentes ao retrotransposón.

En colaboración co consorcio HGSCV2, investigamos o potencial da secuenciación de lecturas longas para caracterizar a secuencia completa de elementos L1 de lonxitude completa (FL-L1s). Analizamos un catálogo diverso de 142 FL-L1s que amosaron evidencia de actividade na liña xerminal humana e, con frecuencia, tamén no cancro. A análise de nucleótidos diagnóstico permitiunos asignar estes elementos a subfamilias específicas de L1Hs. A maioría destes elementos pertencían á subfamilia evolutivamente máis recente, *Ta-1* (55,6%). Porén, tamén atopamos FL-L1s activos das subfamilias máis antigas *Ta-0* (22,5%) e *pre-Ta* (19,7%), en consonancia con estudos previos que apoian que a retrotransposición humana é sostida por varios linaxes de L1Hs. Aproveitando as secuencias internas de alta calidade, avaliamos o estado dos marcos de lectura abertos (ORF) para este conxunto de 142 FL-L1s activos. Atopamos que a maioría (81%) destes elementos competentes presentaban dous ORFs intactos (ORF1 e ORF2). Porén, identificamos unha fracción notable (19%) de FL-L1s activos con ORFs alterados, incluíndo un elemento L1 *"hot"* en 9q32 reportado como altamente activo en diversos tumores humanos. Estes achados son indicativos ou ben da existencia de trans-complementación entre elementos fonte L1, ou ben de heteroxeneidade alélica asociada á recente perda de actividade de certas copias. Posteriormente, inferimos unha filoxenia ML

USC DE COMPOSTELA

baseada no aliñamento de 142 secuencias completas de FL-L1 e estimamos a idade aproximada dos elementos FL-L1 considerando a súa diverxencia respecto á consenso específica do seu respectivo linaxe. Os elementos *Ta-1* xeralmente representaron copias máis novas en comparación con *Ta-0* e *pre-Ta*, que resultaron ser elementos mais antigos e menos polimórficos. De maneira destacable, os FL-L1 relacionados co cancro distribúense en diversos clados da filoxenia L1. Isto incluíu unha inserción *pre-Ta* fixada en 1p12 que, a pesar de terse integrado no xenoma humano hai ~1,8 millóns de anos, conserva importante actividade tanto na liña xerminal como a nivel somático nos xenomas tumorais. Os nosos resultados suxiren que un pequeno conxunto de copias antigas de L1 *pre-Ta* podería seguir sendo relevante para o desenvolvemento do cancro.

Motivados pola resolución proporcionada por lecturas longas na liña xerminal, contribuímos ao desenvolvemento de MEIGA, un proxecto concibido para investigar a dinámica da retrotransposición somática de L1 no cancro. Deseñamos dúas ferramentas bioinformáticas específicas. A primeira é MEIsimulator, un método para obter xenomas simulados con diversos eventos de retrotransposición en localizacións xenómicas predefinidas. A segunda é Source Inference, un método que rastrexa insercións somáticas de L1 ata o seu elemento de orixe baseándose en combinacións de nucleótidos diagnóstico. Avaliamos o rendemento do algoritmo Source Inference utilizando conxuntos de datos tanto de retrotransposición somática *in silico* coma de tumores de orixe epitelial. Aínda que o frecuente truncamento das insercións L1 limitou a sensibilidade da estratexia, esta avaliación demostrou a capacidade de Source Inference para determinar con precisión a orixe de insercións somáticas de L1 en cancro.

Aplicamos o noso método Source Inference aos datos de secuenciación de lecturas longas dunha selección de 10 tumores de MEIGA con altas taxas de retrotransposición. Como resultado, puidemos rastrexar 1,325 eventos (899 insercións *solo*-L1) ata o seu elemento fonte orixinal. Estes eventos representaron o 44,5% (1,325/2,980) do número total de eventos de retrotransposición rastrexados nos tumores MEIGA utilizando diversos enfoques. As mobilizacións rastrexadas mediante Source Inference tiveron un impacto substancial nos patróns de actividade de numerosos elementos fonte. En particular, 688/899 (76,5%) dos *solo*-L1s rastrexados foron atribuídos a elementos fonte con actividade coñecida no cancro segundo PCAWG, incluíndo 294 eventos vinculados a 14 elementos fonte con actividade *"hot"* coñecida en xenomas tumorais, o que destaca a súa relevancia mutaxénica. Ademais dos 151 elementos fonte identificados na cohorte de MEIGA mediante as súas transducións derivadas, outros 46 elementos crípticos emerxeron como doantes de *solo*-L1s somáticos baixo a nosa estratexia de nucleótidos diagnósticos. A adición de *solo*-L1s incrementou en promedio 1,37 veces a actividade dos L1 *"hot"* en MEIGA en comparación cunha estratexia baseada en transducións. Atopamos que algúns elementos fonte (por exemplo, 1p22.3, 1q24.1, 12q24.22) son tipicamente mediadores de insercións *solo*-L1, mentres que outros (por exemplo, Xp22.2-1, 6p24.1, 6p22.1) median usualmente eventos de transdución. Atopamos evidencia que apoia que a variación na forza relativa dos sinais de poliadenilación canónica e alternativa involucradas na transcrición de L1 podería determinar este patrón. Ademais, para un conxunto limitado de FL-L1s en homocigose, Source Inference permitiu discriminar o alelo de orixe de insercións somáticas específicas de L1, demostrando o potencial dos datos de secuenciación de lecturas longas para desentrañar a bioloxía dos elementos fonte de L1 no cancro cunha alta resolución.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Capítulo II: O impacto da variación xenómica na evolución da neoplasia transmisible en bivalvos

A continuación, exploramos o impacto da variación estrutural no xenoma da neoplasia transmisible de bivalvos en *Cerastoderma edule* (CedBTN). Os cancros transmisibles son liñaxes celular somáticos que se propagan entre individuos mediante a transferencia directa de células tumorais, eludindo o sistema inmune do hóspede. Por este motivo, representan un modelo crucial para entender as bases moleculares do cancro metastático e a dinámica da evolución somática a longo prazo.

Na última década, describíronse cancros transmisibles en polo menos oito especies de bivalvos, incluíndo berberechos comúns (*Cerastoderma edule*), ameixas de cuncha branda (*Mya arenaria*) e mexillóns de bahía (*Mytilus trossulus*), entre outros. Estes tumores maniféstanse como unha condición similar á leucemia, derivada da proliferación anómala de hemocitos transformados que se propagan na auga salgada do océano infectando novos hóspedes. Estudos recentes identificaron algunhas alteracións moleculares asociadas aos xenomas de BTN, caracterizados por cariotipos inestables, marcada hiperploidía e disrupción de mecanismos esenciais que controlan o ciclo celular (por exemplo, alteración de p53). Investigacións previas identificaron unha posible orixe polifilética para CedBTN, que podería estar representada por dous clons tumorais independentes. Non obstante, a caracterización xenética de CedBTN limitouse á análise dun pequeno número de secuencias de ADN e xenes marcadores. A recente publicación de xenomas de referencia completos de bivalvos, como o ensamblado no marco do proxecto *Scuba Cancers* para o berberecho común, proporciona un recurso clave para caracterizar de maneira sistemática as alteracións xenómicas (por exemplo, SNVs, SVs e CNVs) no xenoma nuclear de CedBTN e investigar a evolución somática destes tumores transmisibles.

No marco do proxecto *Scuba Cancers*, analizamos datos de secuenciación do xenoma completo de 10 tumores CedBTN con alta pureza e de 455 berberechos comúns mostreados no rango xeográfico da especie na costa atlántica. Identificamos variantes estruturais nos tumores mediante unha combinación de algoritmos bioinformáticos. O xenotipado de variantes estruturais permitiu discriminar entre alteracións somáticas e xerminais, resultando na identificación dun total de 18,272 variantes estruturais específicas de CedBTN. Analizamos a incidencia e distribución cromosómica dos diferentes tipos de variantes estruturais no xenoma nuclear de CedBTN. As delecións, típicamente de tamaño moderado, representaron o tipo máis frecuente de variante estrutural no xenoma de CedBTN (79.8%, 14,589/18,272), seguidas por duplicacións e potenciais reordenamentos intercromosómicos. Detectamos numerosos reordenamentos a grande escala, incluíndo delecións, duplicacións e inversións distribuídas en diversos cromosomas. Os nosos resultados evidencian un elevado grao de reorganización estrutural no xenoma de CedBTN.

A continuación empregamos a variación xenómica identificada no proxecto *Scuba Cancers* para explorar a estrutura clonal e as orixes evolutivas de CedBTN. Realizamos inferencias filoxenéticas por MP e ML derivadas de fontes complementarias de variación xenómica (18,272 SVs e 4,3 millóns de SNVs), que, en combinación con evidencia adicional obtida no marco do proxecto (patróns de dN/dS, diverxencia respecto a poboacións actuais de berberechos e expansión de secuencias satélite), confirmaron a coexistencia de dous clons tumorais antigos e independentes (CedBTN1 e CedBTN2), distinguibles por características moleculares e histolóxicas. Para facilitar o diagnóstico de CedBTN1 e CedBTN2, deseñamos

USC DE COMPOSTELA

un ensaio baseado en PCR que determina o xenotipo de delecións somáticas para identificar a neoplasia e discriminar ambos linaxes de forma rápida e efectiva.

Finalmente, realizamos unha procura sistemática de alteracións condutoras nos xenomas de CedBTN1 e CedBTN2. Centrámonos en variantes pre-diverxencia con impacto significativo na secuencia codificante de xenes de cancro. A nosa avaliación a nivel xenómico identificou múltiples instancias de xenes relacionados co cancro potencialmente afectados por variantes somáticas, algunhas das cales poderían contribuír á inestabilidade cromosómica observada en BTN (por exemplo, variantes de *splicing* e cambio de marco de lectura nos xenes *RB1* e *STAG2*). Porén, non logramos identificar mutacións condutoras fiables debido á integración destas mutacións coas estimacións do número de copias dos xenes afectados. Non obstante, o análise de patróns de número de copias e expresión xénica asociada permitiunos detectar posibles alteracións condutoras que poderían terse adquirido nas fases iniciais de CedBTN. Identificamos unha deleción somática ancestral en homocigose do supresor tumoral *MGMT*, asociada cunha sinatura mutacional vinculada á reparación defectuosa de bases de ADN alquiladas en CedBTN1. Ademais, detectamos amplificacións de importantes oncoxenes (*MDM2*, *CCND3* e *MYC*) potencialmente implicados na disrupción do ciclo celular e das respostas fronte a aneuploidía en CedBTN. Estas e outras alteracións caracterizadas no marco do proxecto *Scuba Cancers*, como a duplicación do xenoma completo e capturas recorrentes de ADN mitocondrial, poderían representar adaptacións moleculares para soster a viabilidade destes xenomas de cancro transmisible durante a evolución somática a longo prazo.

Capítulo III: Predición de inestabilidade de microsatélites no cancro colorectal

O terceiro mecanismo que estudamos é a inestabilidade de microsatélites (MSI) no cancro colorectal (CRC). Os microsatélites son secuencias repetitivas curtas de ADN, que adoitan expandirse ou contraerse debido a erros no deslizamento da ADN polimerasa e a defectos na maquinaria de reparación de erros de apareamento (MMR). As deficiencias nos xenes MMR (por exemplo, *MLH1, MSH2, MSH6, PMS2*) con frecuencia conducen ao desenvolvemento dun fenotipo hipermutador MSI, caracterizado por unha alta incidencia de SNVs e INDELs dentro destes microsatélites.

O fenotipo MSI obsérvase comunmente en síndromes hereditarios como o síndrome de Lynch, pero tamén en tumores sólidos esporádicos. Os tumores MSI adoitan estar asociados cun mellor prognóstico e teñen menor probabilidade de desenvolver metástases. Esta tendencia está relacionada cun aumento na produción de neoantíxenos nos tumores MSI, o que incrementa a súa inmunoxenicidade. De maneira notable, os tumores con fenotipos hipermutadores demostraron mellores respostas aos tratamentos anti-CTLA-4 e anti-PD-1 no melanoma e cancro de pulmón, respectivamente. Ademais, os pacientes con CRC e deficiencias en MMR xeralmente presentan un mellor prognóstico e un maior beneficio coa inmunoterapia anti-PD-1. Esta relevancia terapéutica motivou o desenvolvemento de enfoques computacionais para determinar o fenotipo MSI a partir de diferentes fontes de información molecular, desde a cuantificación directa a partir de datos de secuenciación do xenoma completo ata a súa inferencia a partir de perfís epixenéticos e transcriptómicos.

Neste proxecto, empregamos os datos de microarrays de metilación correspondentes a 388 tumores de CRC do proxecto The Cancer Genome Atlas (TCGA) co obxectivo de desenvolver un predictor de MSI baseado en metilación. Realizamos un análisis de metilación diferencial entre tumores primarios MSI e MSS, o que nos permitiu identificar 780 sondas CpG

USC DE COMPOSTELA

diferencialmente metiladas. Con esta información, adestramos modelos supervisados de aprendizaxe automática sobre o mesmo conxunto de datos e avaliamos o seu rendemento a través de roldas iterativas de validación cruzada con 10 particións. Tras comparar sete algorítmos alternativos para a predición de MSI (GLM, Neural Network, Random Forest, SVM, Decision Trees, KNN e Naïve Bayes), optimizamos un clasificador efectivo (*Accuracy*=0.98, *Kappa*=0.92) baseado en GLM composto por 400 sondas CpG, que denominamos MSIMEP. Ademais, considerando unha potencial aplicabilidade a biopsia líquida, e asumindo un sacrificio máximo dun 1% de exactitude, desenvolvemos unha versión reducida de MSIMEP composta por 25 sondas CpG. Analizamos a importancia no modelo e as características biolóxicas dos predictores de MSIMEP, identificando sondas relevantes distribuídas en rexións cromosómicas tanto xénicas como interxénicas. Notablemente, unha porcentaxe importante da capacidade de clasificación de MSIMEP asociouse á rexión *MLH1*, destacando a relevancia da inactivación epixenética deste xene de MMR no desenvolvemento da condición MSI. A continuación, avaliamos o rendemento de MSIMEP en diversas cohortes de cancro. Malia presentar un rendemento limitado noutros tipos de cancro (STAD e UCEC), MSIMEP demostrou un rendemento robusto (*Accuracy* = 0.93, *Kappa* = 0.83) na súa avaliación no conxunto externo de 517 tumores colorrectais. Aínda que MSIMEP debe considerarse como un recurso computacional que require unha maior avaliación antes de contemplar a súa aplicabilidade clínica, estes resultados destacan o potencial da caracterización epixenética para determinar o estado MSI en tumores de CRC.

Conclusións

Este traballo presentou achados relevantes da miña investigación no programa de doutoramento en Medicina Molecular, centrado na inestabilidade xenómica no cancro. En conxunto, estes resultados contribúen modestamente a unha mellor comprensión dos mecanismos moleculares que impulsan a inestabilidade xenómica no cancro. Os nosos achados demostran o potencial das tecnoloxías emerxentes de alto rendemento, xunto con enfoques bioinformáticos avanzados, para identificar e caracterizar elementos xenéticos e epixenéticos implicados na variación estrutural xenómica no cancro. O análise de variantes xenéticas e epixenéticas e o estudo da súa distribución ao longo de extensas rexións cromosómicas, foi valioso para proporcionar información sobre os patróns de evolución somática en tumores. Aínda que estes resultados enmárcanse na investigación básica en xenómica do cancro, o desenvolvemento futuro destas liñas de investigación en cancros primarios, metastásicos e transmisibles podería contribuír a avances terapéuticos novidosos nos campos da medicina personalizada e a oncoloxía molecular.

Palabras chave: cancro, xenoma, evolución somática, retrotransposición, variantes estruturais, microsatélites, bioinformática, cancro transmisible, inestabilidade xenómica.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

INTRODUCTION

Cancer is commonly referred to as a complex group of genetic diseases, encompassing over 100 disorders that affect various cell types and tissues in humans and many other multicellular organisms (Stratton, Campbell, and Futreal 2009; Aktipis et al. 2015). At its core, cancer involves uncontrolled cell proliferation, driven by populations of transformed tumour cells undergoing evolutionary changes (Brown et al. 2023). Despite their molecular heterogeneity, these disorders share a hallmark feature: the loss of control over specific cell lineages, leading to unrestricted clonal cell division, tissue invasion and metastasis, often resulting in severe damage to the host (Martínez-Jiménez et al. 2020).

According to WHO statistics, nearly 20 million cases of cancer occur annually (Sung et al. 2021; Ferlay et al. 2021), with an increasing incidence linked to population aging and exposure to risk factors. Currently, more than half of cancer cases can be cured or managed as a chronic condition. Nevertheless, cancer remains a leading cause of death globally, responsible for over 8 million deaths each year Pan-Cancer Analysis of Whole Genomes Consortium 2020). Notably, over 90% of cancer-related deaths are associated with metastasis, a process in which tumour cells spread through the circulatory and lymphatic systems to distant organs (Ganesh and Massagué 2021).

Cancer development arises from the interplay of genetic and environmental factors that drive the process of aberrant cell proliferation, invasion and metastasis (Stratton, Campbell, and Futreal 2009). Tumours generally initiate from mutations (genetic alterations) that impair key molecular pathways such as cell-cycle control, DNA repair and apoptosis (Martínez-Jiménez et al. 2020). A minor fraction -approximately 5% to 10%- of the human cancers show a strong hereditary component, arising from deleterious mutations inherited via the parental germline, often leading to different cancer-associated syndromes (Giardiello et al. 2014). However, most of the human tumours originate because of the accumulation of de novo genetic alterations in somatic cells (Figure 1). This genetic damage is acquired during organism development due to intrinsic (e.g. replication errors, metabolic oxidative stress, mutational activation) and extrinsic (e.g. UV-light, toxin exposure, viral infections) agents that, ultimately, impact on the genetic programs of somatic cells (Stratton, Campbell, and Futreal 2009). This connection between DNA damage accumulation and progressive deterioration of stem cells and somatic tissues defines an intrinsic link between the disease and the biological process of aging.

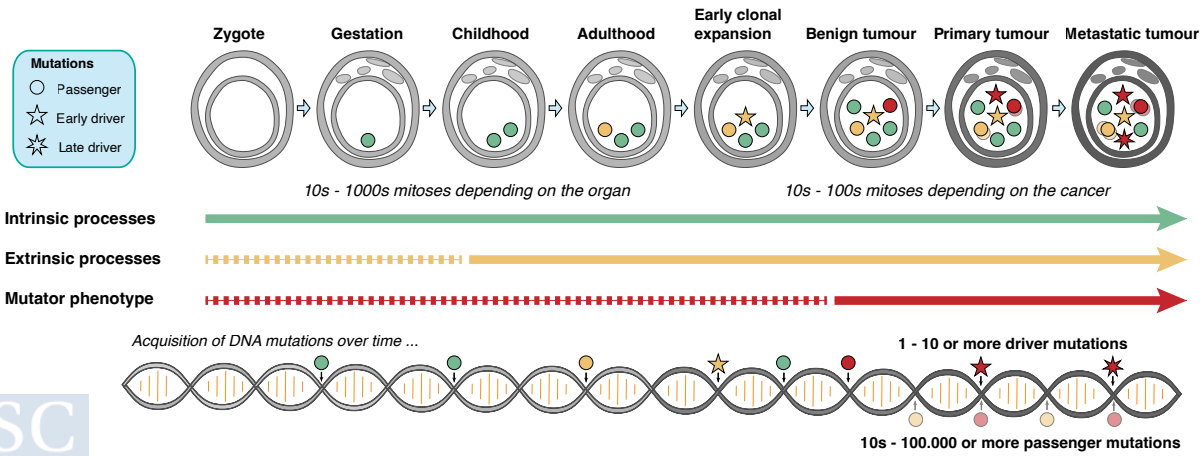


Figure 1: The role of somatic mutation in cancer development.

Cancer generally arises from the accumulation of genetic alterations into the genome of somatic cells. These mutations are incorporated and fixed into the genome of somatic cell lineages, and they result from both intrinsic and extrinsic mutational processes operating over time. Eventually, these mutations can lead to unrestricted cell clonal expansions which potentially can promote the development of cancer. However, only a small fraction of these somatic mutations represent driver events involved in cancer progression and metastasis. Adapted from: Stratton, Campbell, and Futreal 2009 with permission of Springer Nature.

Cancer development is characterized by the process of somatic clonal evolution (Gerstung et al. 2020; Baez-Ortega et al. 2019; Stammnitz et al. 2023). In this process, the mitotic division of tumour cells, and their constant acquisition of variants through diverse mutational processes, leads to the diversification of the tumour populations into clonal lineages with different genetic backgrounds (Gerstung et al. 2020). During cancer development, the success of these clonal lineages and their associated genetic variation is driven by evolutionary forces such as selection, genetic drift and gene flow (Martincorena et al. 2017; Baez-Ortega et al. 2019). As a result, heterogeneous tumour clones evolve and acquire novel features, such as increased growth rate, metastatic potential or resistance to an eventual therapy (Stratton, Campbell, and Futreal 2009). The evolution of these clones is conditioned by multiple factors, such as changes in the mutation rates, metabolic limitation and host immune response, among other selective pressures from the tumour microenvironment (Ganesh and Massagué 2021).

Importantly, the vast majority of the short genetic variants that are acquired by the cancer genome during its development are passenger mutation events, with minimal to no influence in the clonal evolution of the cancer genome and in the progression of the disease (Martincorena et al. 2017; Stratton, Campbell, and Futreal 2009). Even when affecting to the small proportion of DNA represented by coding regions, most (~99%) point mutations are expected to be neutrality selected in cancer (Martincorena et al. 2017). This limited impact can be attributed to various protective mechanisms acquired along species evolution. These factors include the redundancy in gene contents, the degeneration of the genetic code, and the resilience of most cellular pathways, especially those involved in cell-cycle control. Only certain variants altering key genomic sequences, such as the regulatory and coding regions of tumour suppressor genes (TSG) and oncogenes, as well as those disrupting 3D genome architecture, are expected to be critical for cancer initiation and progression. These alterations that confer selective advantages to the cancer clones and support its development are known as driver mutations. Recent studies have shown that the number of driver mutations required for cancer development can sometimes be relatively low -1 to 10 drivers per tumour- (Martincorena et al. 2017). Remarkably, some of these driver events show high tumour specificity (e.g. *EGFR* mutations in glioblastoma and lung adenocarcinoma), while others (e.g. common *TP53*, *KRAS* and *BRAF* alterations) are generally cancer wide drivers (Martínez-Jiménez et al. 2020).

Over recent decades, important efforts have been made to understand the genetic basis of multiple cancer types, leading to the discovery of a large catalogue of more than 700 cancer-related genes and to the identification of novel therapeutic targets (Sondka et al. 2024). These discoveries have been paved by the development of high-throughput sequencing technologies, enabling the identification and characterization of numerous cancer-associated genetic variants, particularly those affecting coding sequences (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Notably, large-scale international sequencing projects such as The Cancer Genome Atlas (TCGA) and its extension, the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) project, have advanced our understanding of common patterns of somatic mutation trough the systematic analysis of tumours and matched normal tissues (Sondka et al. 2024). However, we are only beginning to unravel the mechanisms underlying

large-scale chromosomal alterations in cancer genomes, and to understand how these mechanisms interplay with tumorigenesis and metastasis (Xuelan Chen et al. 2024).

Genomic instability is a molecular condition characterized by an accelerated accumulation of genetic alterations through the genome (Shen 2011; Tubio 2015; Cosenza, Rodriguez-Martin, and Korbel 2022). This instability can arise from different sources, including defects in DNA repair machinery, chromosome missegregation, exposure to environmental factors such as radiation or chemicals, and the dysregulation of endogenous and exogenous mutagenic elements (Xuelan Chen et al. 2024). The resulting genetic alterations range from small-scale changes like single nucleotide variants (SNVs) and small insertion-deletions (INDELs) to larger structural variations. Structural variants (SVs) are larger-scale genetic alterations that involve changes in the structure of DNA, typically affecting sequences larger than 50 base pairs (bp) (Mahmoud et al. 2019). These alterations include deletions, duplications, insertions, inversions and translocations, which are associated to different patterns of DNA gains, losses and rearrangements (Yi and Ju 2018). Some of these complex alterations, such as chromotripsis and chromoplexia, can be triggered by critical events such as telomere erosion, chromosome breakage-fusion-bridge (BFB) cycles and subsequent micronuclei formation and degradation (Xuelan Chen et al. 2024). Additionally, SVs can also be induced by the integration of both endogenous and exogenous DNA sequences, including transposable elements, viral sequences and mitochondrial DNA (Tubio 2015; Yi and Ju 2018).

Genomic instability is a hallmark of cancer, driving genetic diversity and promoting clonal expansions that contribute to cancer progression and therapy resistance (Stratton, Campbell, and Futreal 2009; Xuelan Chen et al. 2024). Chromosomally unstable tumours are generally refractory to immunotherapy and display several hallmarks of immune evasion (Xuelan Chen et al. 2024). Although most structural variants accumulated in cancer are still expected to have a neutral evolutionary effect, collectively, they represent a major source of genetic variation that fuels tumour development (Tubio 2015). Structural variants can disrupt, amplify or fuse cancer-related genes, and also can modify non-coding DNA regulatory sequences altering gene expression (Cosenza, Rodriguez-Martin, and Korbel 2022). Multiple cancer genes have been identified to be frequently amplified (e.g. *MCL1*, *EGFR*, *MYC*, *ERBB2*) or deleted (e.g. *CDKN2A/B*, *RB1* and *PTEN*) in cancer cells (Yi and Ju 2018). According to the results of PCAWG studies, structural variants represent the majority (55%) of the cancer driver events generally found in human cancers (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Notably, structural variants are often associated with complex repetitive DNA sequences subjected to low mappability, which are difficult to profile under the lens of common short-read sequencing technologies and classic bioinformatic methods (Yi and Ju 2018). For this reason, the relevance of these alterations might be underestimated in many cancer types (Cosenza, Rodriguez-Martin, and Korbel 2022).

The impact of genomic instability varies notably across cancer types and subtypes (Tubio 2015; Rodriguez-Martin et al. 2020; Cosenza, Rodriguez-Martin, and Korbel 2022). Malignant epithelial tumours (carcinomas) for instance, often exhibit higher rates of genomic rearrangements, while many non-epithelial tumours (e.g. tumours of mesenchymal, hematopoietic or neuroectodermal origin) tend to be more stable (Cosenza, Rodriguez-Martin, and Korbel 2022). However, specific cancer types in both cases are associated to characteristic driver genomic rearrangements and variant patterns (Table 1). Examples of this diverse genomic impact can be found in the incidence of chromoplexy in prostate cancer, RAG-

mediated deletions in acute lymphoblastic leukaemia, microsatellite instability in colorectal cancer, and somatic retrotransposition on oesophageal cancer (Tubio et al. 2014; Cortes-Ciriano et al. 2017; Rodriguez-Martin et al. 2020). This underscores the complex and diverse roles that genomic instability plays in cancer development, with distinct mechanisms of structural variation contributing uniquely to tumorigenesis across different cancer types and subtypes (Cosenza, Rodriguez-Martin, and Korbel 2022).

Table 1: Hallmarks of Genomic instability and Structural Variation in Human Cancer.

Hallmark(s)	Cause(s)	Mutation pattern(s)	Cancer type(s)	Reference(s)
Microsatellite instability	DNA mismatch repair disruption	INDELs and SNVs	COAD, READ, STAD, UCEC	Cortes-Ciriano et al. 2017
Widespread structural variant accumulation	Double-Strand Breaks, Replication stress, DNA repair disruption	Deletions Duplications Inversions Translocations	PAAD, LUAD, GBM, UCEC	Cosenza, Rodriguez-Martin, and Korbel 2022
Somatic retrotransposition	Epigenetic alterations	Insertions, transductions, pseudogene mobilizations	ESAD, HNSC, LUSC, COAD	Rodriguez-Martin et al. 2020
Viral DNA integration	HBV, HPV, EBV, KSHV infection	Somatic rearrangements	LIHC, CESC, NPC	Álvarez et al. 2021
Bacterial DNA integration	*Helicobacter pylori*	Somatic rearrangements	STAD	Salvatori et al. 2023
Chromothripsis Chromoplexy	Telomere erosion BFB cycles Micronucleus rupture	Chromosome shattering Translocations, deletions	Sarcomas PRAD	Stephens et al. 2011; Baca et al. 2013
RAG-mediated deletions	Off-target activity of RAG enzymes	Deletions with RSS motifs	ALL	Papaemmanuil et al. 2014
Whole genome doubling	Chromosome missegregation	Tetraploidy (4n)	Widespread	Storchova and Pellman 2004

An overview of representative hallmarks of genomic instability and associated causes and mutation patterns is presented, together with typically associated cancer types. Abbreviations: COAD=colon adenocarcinoma; READ= rectum adenocarcinoma; STAD=stomach adenocarcinoma; UCEC=uterine corpus endometrial carcinoma; PAAD=pancreatic adenocarcinoma; LUAD=lung adenocarcinoma; GBM=glioblastoma multiforme; ESAD= oesophageal adenocarcinoma; HNSC=head and neck squamous cell carcinoma; LUSC=lung squamous cell carcinoma; LIHC=liver hepatocellular carcinoma; CESC=cervical squamous cell carcinoma and endocervical adenocarcinoma; NPC=nasopharyngeal carcinoma; PRAD=prostate adenocarcinoma; ALL=acute lymphoblastic leukaemia; INDELs=insertion-deletions; SNVs=single nucleotide variants; BFB=breakage-fusion-bridge; RAG=recombination activating gene; RSS=recombination signal sequence; HBV=hepatitis B virus; HPV=human papillomavirus; EBV=Epstein-Barr virus; KSHV=Kaposi sarcoma-associated herpesvirus.

Genomic instability remains as a largely unknown aspect of cancer biology, and gaining insights about its determinants might have relevant implications for understanding tumorigenesis, metastasis and devising new therapeutic approaches against the disease (Yi and Ju 2018; Cosenza, Rodriguez-Martin, and Korbel 2022; Xuelan Chen et al. 2024). In this thesis, we will integrate insights from large-scale genomic analysis to study three relevant sources of structural variation and genomic instability in cancer: somatic retrotransposition, chromosomal alterations and microsatellite instability. We will analyse patterns of genetic structural variation in cancer genomes and develop novel bioinformatic approaches to explore the sources and mechanisms of genetic alteration operating in these tumours. Collectively, this research might contribute with some insights to improve our knowledge about the genetic basis of the disease.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

OBJECTIVES

In this thesis, we aim to investigate three distinct mechanisms driving genomic instability in various cancer models. Utilizing a combination of high-throughput sequencing technologies and innovative bioinformatic methods, we aim to study the genetic composition of tumour genomes. This investigation is intended to uncover new molecular features and mutagenic elements that promote genomic instability and the evolution of the cancer genome.

The specific objectives of this thesis project are as follows:

Objective I: To characterize the sequence and the activity patterns of L1 source elements driving retrotransposition in human cancer. In this project, we will utilize a combination of long-read sequencing technologies to investigate the internal DNA sequences, evolutionary origins, and mobilization patterns of L1 elements that drive retrotransposition across the human population and in various cancer types. We will develop and refine bioinformatic methods designed to trace the somatic activity of these L1 source elements within tumour genomes. By applying these methods, we aim to uncover and understand the activity patterns of L1 source elements, particularly in human tumours with high retrotransposition rates. This research will enhance our understanding of the dynamics of L1 retrotransposition in cancer, potentially revealing new avenues for therapeutic intervention.

Objective II: To explore the impact of structural variation across the genome of *Cerastoderma edule* bivalve transmissible neoplasia. This project will analyse extensive whole-genome sequencing data from CedBTN tumours and diverse common cockles to investigate the impact of somatic structural variants on this transmissible cancer genome. By leveraging nuclear genomic variation identified in CedBTN, we aim to explore the clonal structure and evolutionary origins of this haemic neoplasia, and to search for potential driver alterations that may influence the evolution of CedBTN lineages. The insights gained from this study will deepen our understanding of the role of structural variation and genomic instability in the long-term somatic evolution of transmissible cancers.

Objective III: To develop a predictive model to detect microsatellite instability in colorectal cancer. In this project, we aim to develop an advanced computational model to predict the MSI phenotype in colorectal cancer, leveraging tumour methylation profiles. This model will consider the significant role that epigenetic silencing of mismatch repair genes plays in this condition. We plan to evaluate and compare several algorithmic approaches to construct an effective MSI/MSS classifier. This classifier will be rigorously validated across multiple colorectal cancer cohorts. Upon completion, we will analyse the epigenetic factors most critical to our model, assess their distribution across chromosomes, and explore their functional implications. The findings from this research will enhance our comprehension of the molecular underpinnings of microsatellite instability in cancer, offering the scientific community a new computational tool for MSI identification and potentially aiding in the stratification of patients for immunotherapy.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER I: SOURCES OF L1 RETROTRANSPOSITION AND THEIR IMPACT IN THE HUMAN CANCER GENOME

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER I

Sources of L1 retrotransposition and their impact in the human cancer genome

C1.1 - BACKGROUND

C1.1.1 – On the biology of transposable elements

C1.1.1.1 – What are transposable elements?

Transposable elements (TEs) are repetitive DNA sequences with the ability to move or "jump" from one location to another within the host genome (Haig H. Kazazian and Moran 2017; Haig H. Kazazian 2004). Primarily considered as intracellular molecular parasites, these sequences have independently emerged multiple times during the course of eukaryotic evolution (Ivancevic et al. 2016). They are also common in several forms of *Bacteria* and have been identified in *Archaea* (Brügger et al. 2002). The evolutionary origins of transposable elements are diverse, with some transposable elements likely derived from retroviral sequences and some others likely derived from host cells components, such as introns and tRNAs. The intrinsic ability of transposable elements to mobilize and proliferate in bursts, can lead to a dramatic impact on their host genomes, which are often colonized by these sequences within short evolutionary time (P. Deininger 2011; Hancks and Kazazian 2010; Haig H. Kazazian 2004).

C1.1.1.2 – TEs are classified into DNA transposons and retrotransposons

Transposable elements can be broadly classified into two major classes based on their mobilization intermediates and general mode of transposition: DNA transposons and retrotransposons (Platt, Vandewege, and Ray 2018) (Figure 2).

DNA transposons (Class II elements) mobilize as a DNA molecule intermediate (Platt, Vandewege, and Ray 2018). They generally move via a "cut-and-paste" mechanism, which consist in the excision of the element from a genomic site to become later integrated into another location. This process is generally mediated by an enzyme named transposase. While DNA transposons remain active in several species (Brügger et al. 2002), they have achieved limited success in mammals and are no longer mobile in the human genome (Haig H. Kazazian and Moran 2017).

Retrotransposons (Class I elements), on the other hand, mobilize as an RNA molecule intermediate (Platt, Vandewege, and Ray 2018). They generally move under a "copy-and-paste" mechanism, which involves the transcription of the element into an RNA copy which is later reverse transcribed and inserted into another genomic location. Retrotransposons are divided into two major groups: long-terminal repeat (LTR) elements and non-LTR elements, distinguished by the presence of direct terminal repeats. LTR elements, such as endogenous retroviruses (ERVs), are structurally similar to retroviruses, and some of them are considered to be remnants of ancient viral infections adapted to the host cell (Guallar et al. 2018; Frank et al. 2022). The non-LTR elements include long-interspersed elements (LINEs) and short-

interspersed elements (SINEs), which are the dominant transposable elements in mammalian genomes (Platt, Vandewege, and Ray 2018; Haig H. Kazazian and Moran 2017).

C1.1.1.3 - L1 structure and the L1 retrotransposition cycle

The retrotransposition cycle can be exemplified with the mobilization of a LINE-1 (L1) element (Figure 3), which is the dominant LINE retrotransposon in mammals (Platt, Vandewege, and Ray 2018). L1 elements possess a full-length structure of 6.4 KB that contains the genetic components necessary for their autonomous mobilization (Ivancevic et al. 2016; Haig H. Kazazian and Moran 2017). Their sequence starts with a GC-rich internal promoter on its 5'UTR region. This region is followed by an ORF1, which encodes a nuclear chaperone protein, and by an ORF2, which encodes an endonuclease and a reverse transcriptase. The L1 element also presents a 3'UTR region ending in a characteristic poly-A tail which is also critical for its retrotransposition cycle (Haig H. Kazazian and Moran 2017; Beck et al. 2011). Recent research has also identified a short, primate-specific antisense open reading frame that might play a role in the regulation of L1 activity (Denli et al. 2015).

The retrotransposition cycle of L1 begins in the nucleus when a full-length element becomes transcribed by RNA polymerase II from its 5'UTR promoter sequence (Beck et al. 2011). This leads to the formation of a L1 mRNA intermediate, which is exported to the cytoplasm and directed to the ribosomal complexes for its translation (Haig H. Kazazian and Moran 2017). The ORF1 encodes a ~40-KDa protein (ORF1p) with RNA binding capacity and chaperone activity which is required to stabilize the L1 molecule. ORF2 encodes a ~150-KDa protein (ORF2p) with endonuclease and reverse transcriptase activities that are essential for the L1 insertion stage (Beck et al. 2011). In the cytoplasm, the ORF1p and ORF2p preferentially associate with their encoding L1 mRNA molecule creating a L1 ribonucleoprotein particle (Wei et al. 2001). This L1 ribonucleoprotein is imported back to the nucleus, where it mediates the insertion of the L1 copy into a novel genomic location. This final stage of the retrotransposition cycle usually occurs through the process of target-primed reverse transcription (TPRT). (Devine 2023; Haig H. Kazazian and Moran 2017).

The TPRT process starts with the L1-encoded endonuclease nicking the target DNA sequence, which is characterized by a specific motif (consensus 5'-TTTT/A-3') (Sultana et al. 2017; Haig H. Kazazian and Moran 2017). The hybridization between the target DNA single strand overhang and the L1 poly-A tail provides then a primer for reverse transcription. This process generates a new L1 DNA copy which gets inserted into the chromosome. Reverse transcription leads to the integration of the retrotransposon sequence into both strands, usually flanked by small (~15 bp) duplications of the target site. Variations in the TPRT process often generates 5' truncated and / or partially inverted L1 sequences (Porubsky et al. 2022; Ostertag and Kazazian 2001). These L1 insertions are generally inactive and probably less deleterious; for this reason are presumed to undergo more neutral evolutionary patterns compared to their source elements (Sultana et al. 2017; Song and Boissinot 2007).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

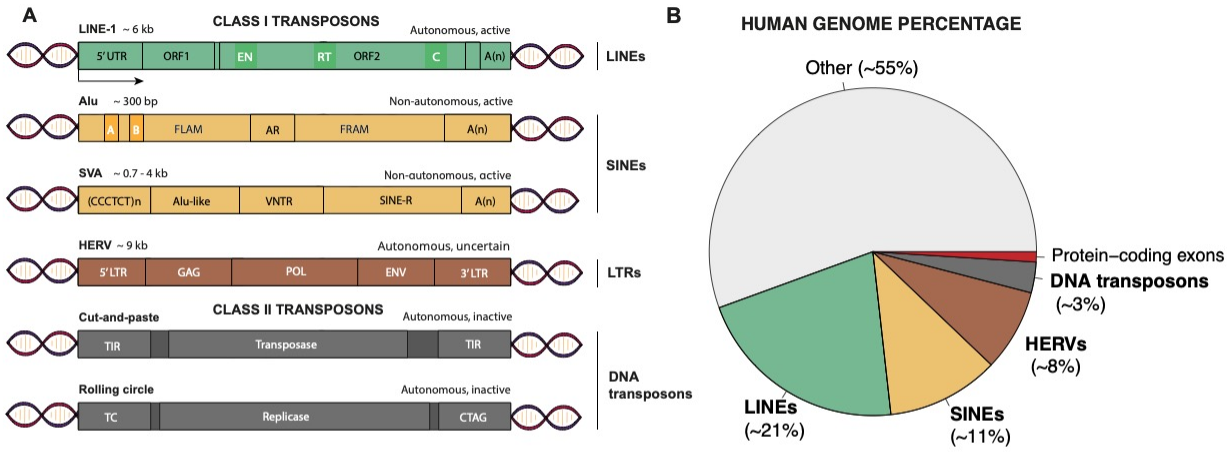


Figure 2: Transposable element biology.
A) Genetic structure of representative transposable elements: 1) LINE-1 elements (LINEs), 2) Alu elements (SINEs), 3) SVA elements (SINEs), 4) HERV elements (LTRs), 5) Cut-and-paste elements (DNA transposons), 6) Rolling circle elements (DNA transposons). Approximate size, mobilization strategy and activity status in the human genome is indicated. B) Percentage of the human genome sequence conformed by different transposable elements, including LINEs, SINEs, HERVs and DNA transposons (total ~43%). Adapted with permission from: Haig H. Kazazian and Moran 2017. Copyright Massachusetts Medical Society.

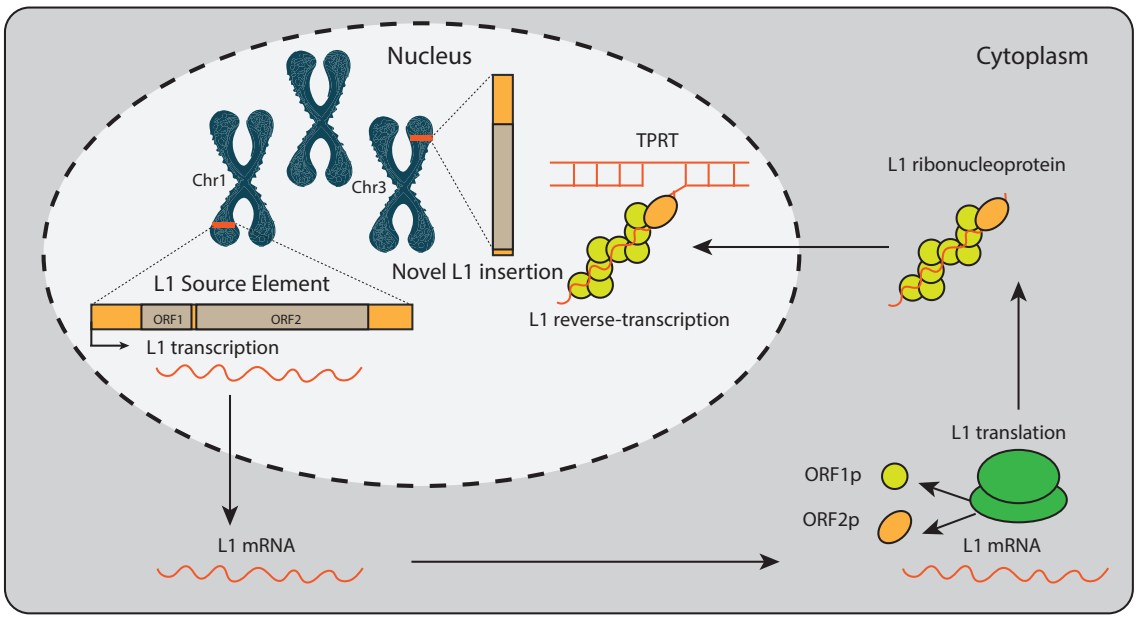


Figure 3: The L1 retrotransposition cycle.
A competent L1 source element with two preserved ORFs (ORF1 and ORF2) is transcribed from its 5'UTR promoter, leading to the generation of a L1 mRNA molecule, which is exported to the cytoplasm of the cell. Once in the cytoplasm, the L1 mRNA molecule is translated to generate ORF1p and ORF2p, which associate with the L1 mRNA molecule generating a L1 ribonucleoprotein. This L1 ribonucleoprotein is imported back to the nucleus, where it mediates the acquisition of a new L1 insertion, generally through the process of target-primed reverse transcription (TPRT). Adapted from: Devine 2023.

C1.1.1.4 - Genomic relevance of TEs

In eukaryotic lineages such as plants and animals, TEs have displayed remarkable success, and they have been great contributors both to the evolution of genome size and regulatory complexity (Haig H. Kazazian 2004; Ivancevic et al. 2016; Haig H. Kazazian and Moran 2017). As illustrative examples, they constitute around 20% of the fruit fly genome and up to 90% of the maize genome, being involved in essential aspects such as the control of chromosomal

structure and genome stability in these species (Lim and Simmons 1994). Another clear evidence of TE success can be found in vertebrate genomes, where transposon-derived sequences are estimated to represent around 50% of total genomic content, and have triggered the evolution of adaptive immunity components such as the major histocompatibility complex (Tsakou-Ngouafo et al. 2020). In mammals, transposon-derived sequences represent more than 40% of total genomic content (Platt, Vandewege, and Ray 2018). A large diversity of TE lineages emerged during mammalian radiation, near 65 Myr ago, suggesting a relevant role in the diversification of this phylum (Ivancevic et al. 2016). Since then, several TEs have co-evolved with their mammalian hosts along the evolution of this animal group, leading to adaptations at multiple biological levels, including gene regulation and 3D genome reorganization (Choudhary et al. 2023). In rodents, TEs might have been involved in the evolution of circadian rhythm control (Judd, Sanderson, and Feschotte 2021). In canids, TE lineages emerged coinciding with major speciation events during wolf evolution (Halo et al. 2019), and extensive genetic variation created by novel insertions likely played a role in the evolution of traits related to domestication and social behaviour. In primates, TE insertions might have promoted key adaptations, such as the evolution of antiviral response (Frank et al. 2022), locomotor modulation (Xia et al. 2021) and nutrient uptake (Ebert et al. 2021).

Despite this impressive adaptive relevance and genomic success, only few members of transposable element families usually remain active at the same evolutionary stage (Ivancevic et al. 2016; Platt, Vandewege, and Ray 2018). TE activity is a strictly regulated phenomenon in the germline, and a process restricted to certain TE lineages and cellular types in the soma. This is needed in order to limit its intrinsic mutagenic potential (Faulkner and Billon 2018). For this reason, most transposable elements usually remain in the host genome as inactive elements, repressed by genetic or epigenetic mechanisms (Platt, Vandewege, and Ray 2018).

C1.1.1.5 – Retrotransposition-competent elements in the human genome

According to several lines of evidence, three main types of retrotransposons remain mobilizing in the human genome: L1, Alu and SINE-VNTR-Alu (SVA) elements. All these are non-LTR retrotransposons which depend on the L1 molecular machinery to become mobilized.

The most prolific retrotransposon in the human germline is represented by the Alu family (P. Deininger 2011). This family includes primate-specific non-autonomous SINE retrotransposons of around 300 bp, which are mobilized in trans by the L1 machinery. With more than 1,100,000 copies, Alu constitutes ~ 11% of the human genomic content. Even when these retrotransposons might have exhibited greater activity during earlier stages of primate evolution, certain Alu subfamilies, such as Ya5 and Yb8, continued mobilizing at elevated rates (1 new insertion / 20 meiosis) (Haig H. Kazazian and Moran 2017). Nowadays they still generate novel polymorphisms, many of which have relevant implications in gene regulation and human disease (Ebert et al. 2021; P. Deininger 2011).

Another active retrotransposon in the human genome is represented by the SVA family (Hancks and Kazazian 2010). This family includes hominid-specific non-autonomous SINE retrotransposons of 700 – 4,000 bp, which also rely on L1 for mobilization. Interestingly, SVA are composite retrotransposons composed of three different components: a SINE region, a VNTR region, and an Alu-like region. With approximately 2,700 copies, SVA constitutes only ~ 0.1% of the human genome content. Despite showing a low mobilization rate (1 new insertion / 900 meiosis) (Haig H. Kazazian and Moran 2017), certain SVA subfamilies, such as SVA-E

USC UNIVERSIDADE

and SVA-J, are still active and can mediate novel polymorphisms with relevant molecular consequences.

The active autonomous retrotransposons in the human genome are represented by the L1 family (Beck et al. 2010). This family includes LINE retrotransposons of up to 6.4 KB which encode the molecular elements required for their mobilization. With more than 500,000 copies, L1 which constitutes ~ 17% of the human genomic content. During primate evolution, several L1 subfamilies emerged and expanded (e.g. L1PA8 - L1PA1 in anthropoid primates (Lanciano et al. 2024)), usually with few lineages competing at the same evolutionary stage (Khan, Smit, and Boissinot 2006). Over the course of modern human diversification, at least three main human-specific L1 (L1Hs) lineages have sequentially emerged: *pre-Ta*, *Ta-0* and *Ta-1*, being *Ta-1* the most recent and prolific (Boissinot, Chevret, and Furano 2000; Marchani et al. 2009). These lineages drive new L1 polymorphic insertion at moderate rates (1 new insertion / 200 meiosis) (Haig H. Kazazian and Moran 2017), which are relevant drivers of genetic variation in modern human populations.

L1 elements also drive significant retrotransposition in the human somatic cells, even when there is limited knowledge about their actual mobilization rates (Faulkner and Billon 2018). L1 activation is known to occur more often at early embryonic stages, likely promoted by favourable epigenetic conditions such as open chromatin and genome-wide hypomethylation (Lanciano et al. 2024). Certain pluripotent cell types, such as human embryonic stem cells, can express and acquire novel L1 insertions, as demonstrated experimentally (Garcia-Perez et al. 2010; 2007). L1 retrotransposition can also occur in multipotent stem cells, and even on more differentiated cell types. Certain extraembryonic tissues, such as placenta, can be severely impacted by L1-mediated mutagenesis, which is probably linked to its limited lifespan and specialized functions (Coorens et al. 2021; Lanciano et al. 2024). However, somatic mobilization of L1 is not confined to embryonic stages alone (Faulkner and Billon 2018). Certain human tissues, such as those from the adult brain (including postmitotic neurons), have been known to be impacted by the activation of L1 retrotransposons. Recent research has demonstrated that L1 somatic activity might be notable in other tissues, such as colorectal epithelium (Nam et al. 2023; Lawson et al. 2020), stomach or small bowel (Zumalave 2024). Remarkably, L1 mosaicism might be involved in the regulation of tissue-specific gene pathways, but somatic L1 activation might be also associated with the development of genetic diseases, including cancer and neurological disorders (Faulkner and Billon 2018; Haig H. Kazazian and Moran 2017; Rodriguez-Martin et al. 2020). To which extent somatic L1 activation is tolerated and even promoted from the human host genome, or on the other hand arises from a selfish, parasitic behaviour from the retrotransposons is an intriguing biological question.

C1.1.2 - Impact of L1 retrotransposition in cancer

C1.1.2.1 - Early observations linking retrotransposon activation and cancer

The first bona fide evidence of L1 impact in the cancer genome was obtained in the early 1990s. Miki et al. (1992) discovered a somatic L1 insertion with TPRT features that disrupted the 16th coding exon of the *APC* tumour suppressor gene in a patient with colorectal cancer (CRC). In gastrointestinal cancers, disruption of the *APC* gene is one of the most common early driver mutations during tumour development in non-MSI tumours (~85% of the cases). Importantly,

USC DE COMPOSTELA

this study evidenced the potential role of somatic retrotransposition as a relevant source of genetic variation driving tumour evolution (Miki et al. 1992).

Over the next decades, technological development and novel research on cancer retrotransposition led to the discovery of new instances where L1-mediated mutagenesis might play a role in the biology of certain human tumours. For example, Iskow et al. (2010) developed Transposon-Seq for systematic L1 profiling, and found that somatic L1 activity might occur at high frequencies in lung tumours (Iskow et al. 2010). Lee et al. (2012) developed Tea and applied it to five cancer types, finding somatic retrotransposition in tumours of epithelial cell origin (colorectal, prostate and ovarian), but absence of mobilizations in blood and brain tumours (Lee et al. 2012). In the same year, Solyom et al. found extensive mobilization in colorectal tumours, and reported a correlation between the patient age and the number of somatic L1 insertions (Solyom et al. 2012). Tubio et al. (2014) studied 14 cancer types, finding that 3' transductions make up 24% of somatic L1 activity in cancer, which is driven by a small number of active source elements with hypomethylated promoters (Tubio et al. 2014). Rodić et al. (2015) found evidence of persistent but discontinuous somatic L1 retrotransposition activity along pancreatic tumour evolution. Scott et al. (2016) identified an additional somatic L1 insertion in the 16th coding exon of *APC* probably driving CRC tumorigenesis, and identified population-specific L1 copies showing remarkable activity in cancer (Scott et al. 2016).

In order to obtain a better overview of the role of somatic retrotransposition in human cancer, our group participated in the PCAWG project (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Here, we studied somatic retrotransposition on 2,954 cancer whole-genomes across 38 different histological tumour subtypes (Rodriguez-Martin et al. 2020). We identified 19,166 somatic retrotransposition events, affecting 35% of the tumours (Figure 4). Notably, we found marked variation across tumour types, with oesophageal adenocarcinoma (ESAD), head-and-neck squamous carcinoma (HNSC), lung squamous carcinoma (LUSC) and colorectal adenocarcinoma (COAD) accounting for 70% of the total retrotransposition events. Remarkably, retrotransposition was identified as the most frequent source of somatic structural variation in ESAD, and the second more frequent in HSNC and COAD, including L1-mediated rearrangements occasionally involved in driver alterations of the cancer genomes.

Collectively, these studies demonstrated that somatic retrotransposition is a characteristic feature of several human cancers, being more frequent on tumours of epithelial origin, and rarely found in tumours derived from the blood or the nervous system. These studies also confirmed that somatic retrotransposition in cancer is commanded by L1 insertions, with Alu and SVA somatic mobilizations being rarely mobilized, as might happen in other somatic tissues. In addition, some of these studies found evidence that somatic retrotransposition is an active process along different stages of the tumour development, in some cases being detected in precancerous lesions.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

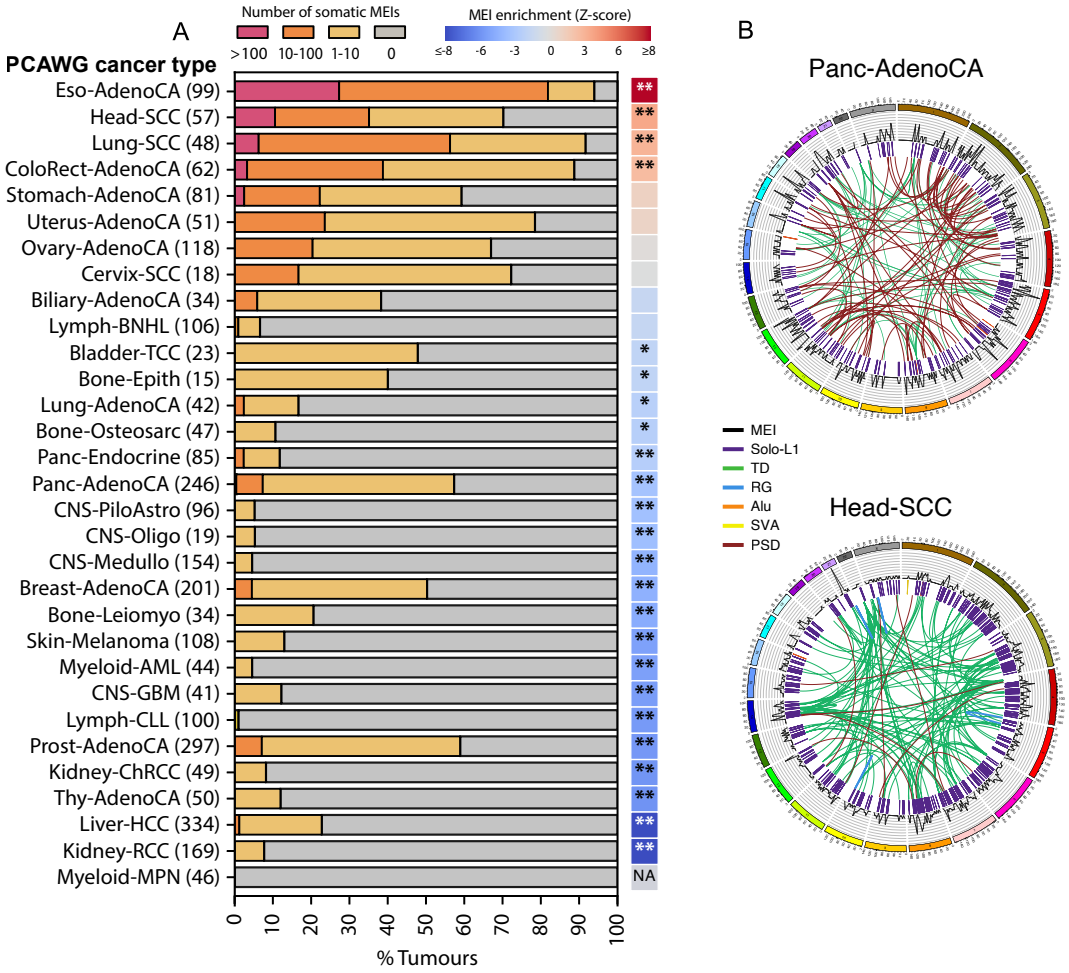


Figure 4: The landscape of somatic retrotransposition across human cancers.
A) For 31 PCAWG cancer types with sample size ≥ 15, bar plots show the proportion of tumour samples with >100 (red), 10-100 (orange), 1-10 (yellow) and 0 (grey) somatic retrotranspositions. The number of samples analysed for each tumour type is shown in parentheses. Retrotransposition enrichment or depletion for each tumour type together with the level of significance (zero-inflated negative binomial regression) is shown. B) Circos plots showing two PCAWG epithelial tumours with high rates of somatic retrotransposition events. Figure adapted from: Rodriguez-Martin et al. 2020.

C1.1.2.2 - Mechanisms linking retrotransposons to cancer development

Somatic retrotransposition is a relevant mutagenic agent in the cancer genome (Haig H. Kazazian and Moran 2017). The mechanisms by which these mutations are generated are diverse, ranging from the local impact of insertional mutagenesis, up to the promotion of large-scale genomic rearrangements (Rodriguez-Martin et al. 2020).

Insertional mutagenesis is probably the best studied mechanism of genetic alteration mediated by L1 (H. H. Kazazian et al. 1988; Miki et al. 1992; Helman et al. 2014). In this process, the somatic integration of the L1 sequence (usually mediated by TPRT) can lead to the local disruption of relevant genetic sequences for the host cell. These sequences can comprise protein-coding exons, but also regulatory elements such as promoters, enhancers and insulators. Since the genomic integration of L1 is widespread, but associated with the A-rich endonuclease motif and with late-replicating DNA regions (Mita et al. 2018; Rodriguez-Martin et al. 2020), most L1 insertions usually do not disrupt protein coding regions, and usually affect intronic or intergenic regions (Sultana et al. 2017). The majority of somatic L1 insertions are expected to

be passenger events for cancer genome evolution (Martincorena et al. 2017). However, some of these events can have functional impacts, leading to relevant alterations for cancer development.

One of the mechanisms by which gene expression is altered, is the promotion of changes in the splicing pattern of the target gene (V. P. Belancio, Roy-Engel, and Deininger 2008; Elbarbary, Lucas, and Maquat 2016). Examples of these splicing alterations are exonization, where the somatic insertion leads to the incorporation of intronic sequences into the spliced mRNA, or exon skipping, where the somatic insertion leads to the non-incorporation of one or more exonic sequences (Jung, Choi, and Lee 2018). Another relevant mechanism of gene alteration is the induction of epigenetic changes (methylation, often associated with gene repression) into the target gene (Beck et al. 2011). This mechanism likely arises as a defensive mechanism of the host cell, in order to prevent further mobilization of the recently-integrated retrotransposons (Yoder, Walsh, and Bestor 1997).

Another relevant mutation process driven by L1 is the generation of 3' somatic transductions (Tubio et al. 2014; Helman et al. 2014; Rodriguez-Martin et al. 2020). This process occurs when the transcription of a source L1 bypasses its canonical termination signal located in the 3' end of the element and continues in the downstream genomic sequence until an alternative polyadenylation site is found (Moran, DeBerardinis, and Kazazian 1999). This process leads to the inclusion of non-repetitive sequences into the L1 mRNA molecule, which are mobilized during the retrotransposition cycle. This process can be remarkably frequent in the context of cancer (~24% of total L1 mobilizations), and often leads to the mobilization of regulatory sequences (TFBS motifs, exons, etc.) to alternative genomic locations (Tubio et al. 2014). This represents an important source of genetic variation, which might be exploited by the cancer cell genome during somatic clonal evolution.

Another mutagenic process driven by L1 is the mobilization of processed pseudogenes (Beck et al. 2011; Rodriguez-Martin et al. 2020). In this process, L1-mediated retrotransposition leads to the propagation of non-functional copies of normal genes, which are incorporated into different genomic locations from its source locus. The L1 machinery applies reverse transcription to copy the processed mRNA of an expressed gene into a cDNA molecule which is integrated into a new genomic location. Processed pseudogenes lack introns and other critical regulatory regions, and for this reason are rarely expressed and usually undergo neutral evolution. However, the sequences carried by these pseudogenes can alter cellular processes and potentially contribute to the somatic clonal evolution observed in cancer cells.

L1 retrotransposons can also promote large-scale genomic rearrangements, such as deletions, duplications, inversions and translocations (Rodriguez-Martin et al. 2020; Beck et al. 2011; Cosenza, Rodriguez-Martin, and Korbel 2022). Some of these events can be acquired along the process of L1 integration, possibly under alternative variants of the canonical TPRT mechanism (e.g. secondary L1 cDNA invasion) (Gilbert, Lutz-Prigge, and Moran 2002). And others can be acquired after integration, driven for example by ectopic recombination between non-homologous repetitive sequences (Song and Boissinot 2007). L1-mediated rearrangements are rare events, less frequent compared with basic L1 retrotransposition. However, they can represent a large source of genetic variation for somatic cancer cells (Rodriguez-Martin et al. 2020), probably driving the selective pressures attributable to L1 retrotransposition (Song and Boissinot 2007). L1-mediated rearrangements can span several megabases of genomic

USC DE COMPOSTELA

sequence. For this reason, these rearrangements simultaneously affect multiple gene functions and can have an impact on the integrity and stability of one or more human chromosomes. In the context of cancer, these rearrangements have been found to disrupt critical genes involved in cell-cycle control, such as tumour suppressor genes and oncogenes. One example is the promotion of rounds of *CCDN1* oncogene amplification driven by breakage-fusion-bridge cycles (Rodriguez-Martin et al. 2020). Another is the promotion of large-scale deletions leading to the loss of tumour suppressor genes, such as *CDKN2A* (Rodriguez-Martin et al. 2020). And another one is the mediation of reciprocal translocations, coupled to additional genetic alterations (Rodriguez-Martin et al. 2020). Interestingly, recent research by our group found that L1-mediated rearrangements could be more frequent than previously observed, probably representing a significant fraction of the genomic rearrangements occurring in certain cancer types (Zumalave et al. 2024). Remarkably, some of these alterations can represent cancer driver events in certain human tumours, highlighting the biological relevance of L1-mediated mutation in the context of cancer.

C1.1.2.3 – Patterns of L1 source elements activity in human cancer

The mutagenic impact of L1 in human germline and somatic tissues is driven by the mobilization of competent L1 copies, which are termed full-length L1 (FL-L1) source elements (Tubio et al. 2014). The study of FL-L1 source elements is crucial for understanding the biology of L1 retrotransposons and their impact on genomic stability.

FL-L1s sources are the only autonomous transposons which remain active in the human genome (Haig H. Kazazian and Moran 2017). They present a complete L1 structure, including an internal promoter in the 5' UTR region, ORF1 and ORF2 sequences, and a poly-A tail on its 3' UTR region. The majority (~75%) of L1 copies in the human genome do not present this functional structure, since they tend to acquire 5' truncation, which prevents them for transcription (Devine 2023). However, many of the remaining full-length L1 copies are inactive as well. This inactivation likely occurs because of the time-dependent accumulation of mutations along the retrotransposon sequence, and due to restrictive epigenetic marks affecting the L1 promoter and to the chromatin surrounding the element. Understanding the genetic and epigenetic factors which determine the FL-L1s source mobilization potential, remains as a relevant question in the field (Devine 2023).

Research on FL-L1s source element biology was boosted in the 1990s, when Kazazian's team discovered a FL-L1 source potentially involved in the mutagenesis of the *F8* gene (coagulation factor VIII) in patients with haemophilia A (Dombroski et al. 1991). This source element (L1.2B) was identified in chromosome 22. The L1 copy, which was found based on matching sequence identity with derived insertions, presented intact ORFs, which supported its competence for driving retrotransposition. Years later, additional work from Kazazian's lab demonstrated that this L1 source element presented two allelic variants (L1.2A and L1.2B) showing differential retrotransposition capacity in cultured human HeLa cells (Lutz et al. 2003).

Over next decades, novel source elements were identified based on alternative screening strategies, including tracing of 3' transductions and profiling of specific L1 substitutions (Brouha et al. 2002; Holmes et al. 1994; Gardner et al. 2017; Scott et al. 2016; Tubio et al. 2014). The development of *in vitro* retrotransposon assays enabled to quantify the activity levels of cloned FL-L1s (Moran et al. 1996). Interestingly, cloned FL-L1s copies with mutation on conserved ORF residues showed drastic decreases in retrotransposition activity. Along these

years, discussion grew up about the controversy between the master gene model (where a limited set of active L1 retrotransposons might produce generally inactive copies), and the precursor gene model, where the L1 elements might be capable of generating active derived copies in the genome (P. L. Deininger et al. 1992; Dombroski et al. 1991).

The publication of the human reference genome and the development of high-throughput sequencing technologies was crucial to promote the study of FL-L1 dynamics. Kazazian and colleagues identified 90 FL-L1s reference elements that had two intact ORFs in the draft of the human genome. They tested 82 of these FL-L1s in a cell-culture-based assay for retrotransposition and found that eight of the elements were highly active "hot L1" source elements (Brouha et al. 2003). Beck et al. later sequenced 68 non-REF FL-L1s that were identified in cosmid clones from eight diverse humans and tested 67 of these elements in the cell-culture-assay for retrotransposition (Beck et al. 2010). They found that 37/67 (55%) of these elements were highly active in the cell culture assay (compared to 8/82, or 9.8%, of the REF elements tested in the Brouha et al. study). These data indicated that the non-REF collection of FL-L1s studied by Beck et al. was enriched in younger hot L1's.

Tubio et al. 2014 developed TraFiC, one of the first computational methods to systematically study somatic L1 mobilizations in cancer whole genomes (Tubio et al. 2014). This method was applied to study the L1 retrotransposition patterns in a cohort of 244 cancer genomes from 12 different cancer types. This led to the detection of 2,756 somatic L1 retrotransposition events. Interestingly, L1 transductions were traced back to a limited set of 72 germline progenitors, representing a first catalogue of FL-L1s with somatic activity in cancer. Interestingly, two hot-L1 loci, located at chromosomes 22q12 and 6p24.1, accounted for more than a third of the traced somatic retrotransposition.

In 2016, Scott et al. discovered a new somatic insertion impacting the *APC* gene likely involved in the CRC carcinogenesis and were able to trace 18/27 somatic L1 insertions to three active source elements in chr17, chr14 and chr12 by profiling internal diagnostic nucleotides. They also showed that population-specific hot L1s might represent a novel form of cancer risk (Scott et al. 2016).

In 2020, research by our group demonstrated that a limited set of 124 (114) FL-L1s might be active in the context of human cancer (Rodriguez-Martin et al. 2020; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). This conclusion was obtained through the analysis of 2,954 cancer whole-genomes from 38 histological subtypes within the framework of the PCAWG project. These active source elements were identified through the analysis of 3' transductions with the software TraFiC-mem, which allowed to track 3669 somatic mobilizations to different L1 source loci. Remarkably, we found that only a small subset of 16 elements presented a relevant contribution (67%) to this mutation burden, and these copies were termed as "hot L1 elements". Notably, five of these elements (22q12.1, 14q23.1, 6p22.1, 6p24.1 and Xp22.2-1) alone gave rise to half of the transductions found in the PCAWG cohort. Interestingly, we described two alternative patterns of activity for those relevant source elements in cancer. We distinguished between Strombolian elements, which produce a small number of retrotransposition events in individual cancer genomes but were often active, and Plinian elements, which were rarely active, but when activated, leaded to extreme numbers of somatic mobilizations.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Our PCAWG analysis of cancer genomes presented however important limitations considering the study of FL-L1 dynamics. First, the observed patterns of source L1 activity were primarily derived from the analysis of 3' transductions, detected under short-read sequencing technologies. While this is a well-established and reliable method for tracing individual L1 retrotransposition events, it does not provide a comprehensive overview of L1 mobilization in cancer genomes. Notably, the major fraction of the observed somatic L1 integrations in cancer genomes is mediated by solo-L1 insertions (76 – 80%) (Tubio et al. 2014; Rodriguez-Martin et al. 2020), and there is missing information about the molecular origin of these retrotransposition events.

Furthermore, while 3' transduction account for 20 – 24% of the insertions, they are not expected to occur uniformly across FL-L1 sources (Moran, DeBerardinis, and Kazazian 1999). FL-L1s sources with weaker termination signals or with non-repetitive downstream sequences, are more likely to be profiled by this approach, while those FL-L1s sources with stronger termination signals or repetitive downstream genomic contexts might be underrepresented, introducing a notable bias in the reconstruction of L1 activation dynamics.

Moreover, the study of somatic retrotransposition events under the light of short-read sequencing technology implies that internal L1 sequences cannot be resolved for most insertions (those larger than 350 bp). Internal L1 sequences are expected to contain a substantial amount of diagnostic mutations which can be mined to trace individual L1 insertions to their original source locus (Scott et al. 2016). In addition, these internal sequences can be studied to determine the status of ORF1 and ORF2 sequences, being fundamental sequences to determine the competence of FL-L1s (Chuang et al. 2021).

Finally, short-read technology is less sensitive to detect retrotransposon insertions in certain genomic locations, such as nested mobile element insertions and repetitive regions subjected to low read mappability, which are now accessible by long reads (Wang et al. 2021). For this reason, novel approaches based on long-read sequencing technologies might be required to improve our understanding of the FL-L1 retrotransposition dynamics both in the human germline and in cancer genomes.

During this PhD thesis, we developed new computational methods to study the FL-L1 biology and its dynamics in the cancer genome. These methodologies were developed under the framework of the MEIGA project (Mobile Element Insertion Genome Analyser). The application of these methods to a diverse set of human whole-genomes sequenced under novel long-read sequencing technologies might allow to discover and characterize hidden active FL-L1 source elements and to improve our knowledge about their activation dynamics in relation with the evolution of the cancer genome.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C1.2 - RESULTS AND DISCUSSION

C1.2.1 - Analysis of FL-L1 sources in the human germline

C1.2.1.1 - Reconstruction and characterization of active FL-L1 source elements

Based on phased genome assemblies from diverse human genomes across 25 populations, we identified a collection of sequence-resolved FL-L1s (see Methods). We used a transduction-based approach to find a set of 142 active FL-L1 source elements driving L1 retrotransposition in the human germline (Figure 5). This set comprised both reference (66/142, 46.5%) and non-reference (76/142, 53.5%) L1 sources, and included elements known to be active in vitro (Brouha et al. 2003; Beck et al. 2010), in human populations (Gardner et al. 2017), and in human cancers (Tubio et al. 2014; Jung, Choi, and Lee 2018; Rodriguez-Martin et al. 2020). We took advantage of the full reconstruction of their internal DNA sequences, to explore key aspects of their biology.

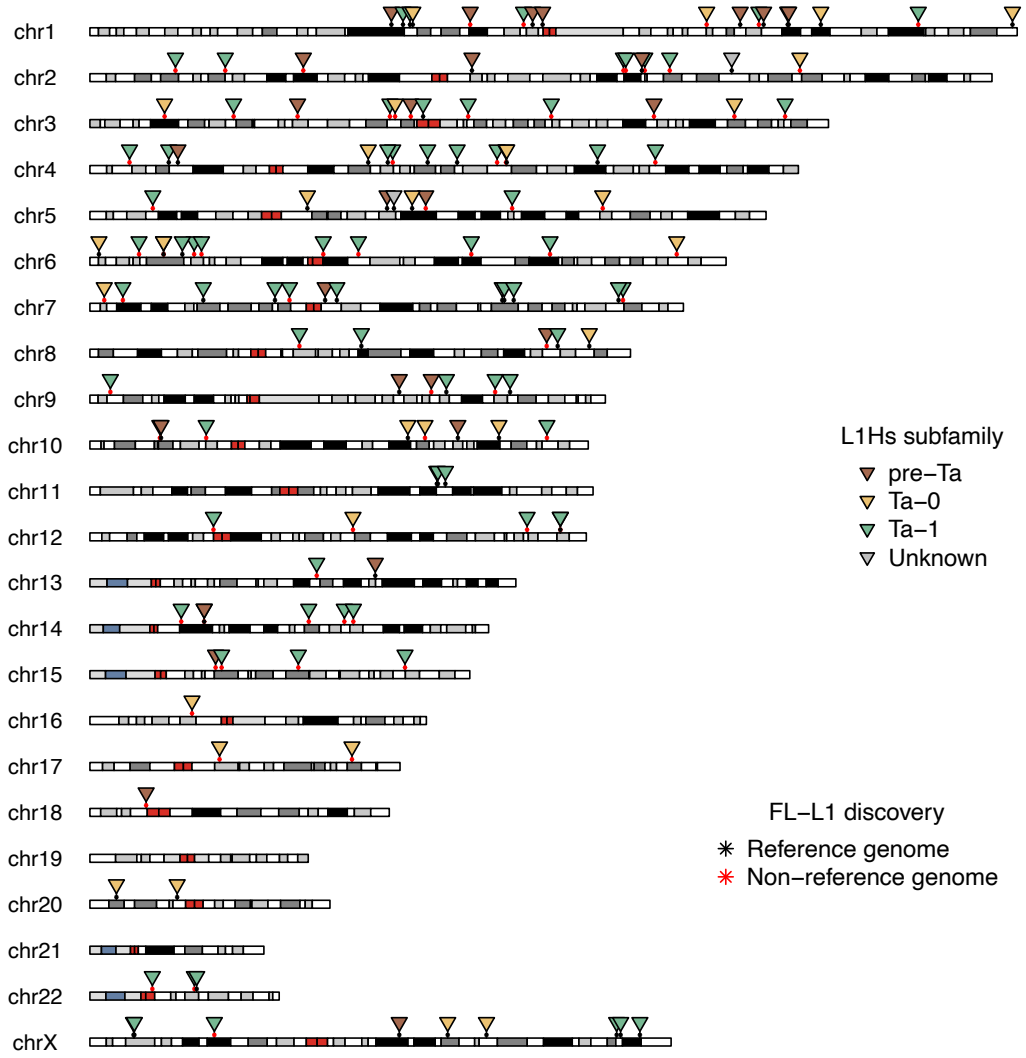


Figure 5: Genomic distribution of active L1 source elements in the human germline.
The genomic location of the 142 active FL-L1 sequences identified across the human genome (HGSVC2). Each triangle indicates the chromosomal coordinate of an active L1 source element, filled according to its L1Hs

USC DE SANTIAGO DE COMPOSTELA

subfamily. Non-reference L1 sources are denoted with a basal red point, and reference L1 sources, with a basal black point. Human reference genome and associated cytobands correspond to GRCh38.

We used L1Hs subfamily diagnostic nucleotide positions (Boissinot, Chevret, and Furano 2000) to determine the lineage for each active FL-L1 (Figure 6). As expected, most active FL-L1s belonged to the *Ta-1* subfamily (79/142; 55.6%), which is known to have mastered retrotransposition across modern human evolutionary history (Boissinot, Chevret, and Furano 2000; Beck et al. 2010), and is estimated to have emerged around 2.01 Myr (95% CI: 2.00, 2.02 Myr). Notably, we also found multiple active copies from the older subfamilies *Ta-0* (32/142; 22.5%) and *pre-Ta* (28/142; 19.7%), supporting that germline L1Hs activity is sustained by different lineages (Figure 6).

Taking advantage of the high-quality resolution for internal L1 sequences provided by PACBIO long reads, we assessed the ORF status for this set of active FL-L1s. We found that the majority (116/142; 81%) of the active FL-L1 sources presented two intact open reading frames (ORF1 and ORF2), encoding all the proteins required for L1, Alu, SVA and processed pseudogene mobilization (Figure 6). Interestingly, we found that 19% of the active FL-L1 sources presented at least one ORF disrupted (14% preserved only ORF1 and 4.2% preserved only ORF2). Remarkably, the set of elements with disrupted ORFs includes a hot element at 9q32 reported to be highly active in diverse human tumours (Rodriguez-Martin et al. 2020) (Figure 6). This high proportion of active source elements with defective ORFs might be indicative of either trans-complementation (Wei et al. 2001) or polymorphisms leading to the recent demise of these active copies.

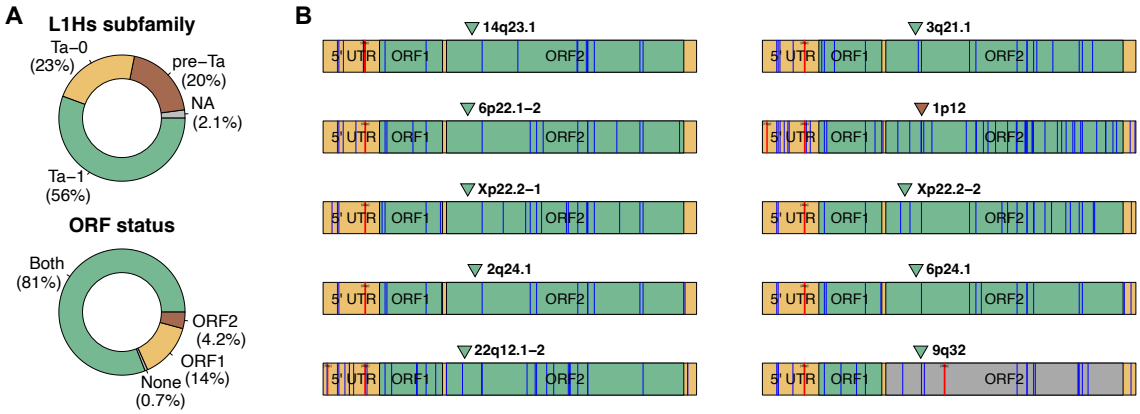


Figure 6: Characterization of germline active FL-L1s with long-read sequencing.
A) Circle charts showing the proportion of active source elements belonging to *pre-Ta*, *Ta-0* and *Ta-1* L1Hs subfamilies (Top chart), and the proportion of active source elements with both ORFs intact, ORF1 intact, ORF2 intact, or none ORF intact (Bottom chart). B) FL-L1 ideograms showing the reconstructed sequences and ORF status of 10 source elements with "hot" activity in cancer genomes (PCAWG). Vertical lines indicate SNVs (blue) and INDELs (red) compared against a L1Hs consensus. The source element 9q32 presents a disrupted ORF2 sequence, denoted in grey colour. Source elements are termed according to their associated GRCh38 cytoband.

C1.2.1.2 - The evolutionary history of active FL-L1 source elements

Using L1 *Pan troglodytes* as an outgroup, we built a Maximum Likelihood (ML) phylogeny of active FL-L1s, which revealed the potential diversification of L1Hs into multiple *pre-Ta*, *Ta-0* and *Ta-1* sublineages (Figure 7, Supp. Figure 3). Additionally, we estimated the age in million years (Myr) of FL-L1 copies based on their divergence from their associated L1Hs subfamily consensus. As expected, copies of the *Ta-1* subfamily were the youngest sources (mean age =

USC DE SANTIAGO

1.00 [95% CI: 0.88 - 1.13]), followed by *Ta-0* (mean age = 1.63 [95% CI: 1.49-1.77]) and *pre-Ta* (mean = 2.15 [95% CI: 1.91-2.40]), which generally represented older sources (Supp. Figure 1). Notably, the evolutionary age of the source elements correlated with L1 features such as subfamily, level of activity and allele frequency, with the youngest FL-L1s typically corresponding to highly polymorphic and active *Ta-1* sequences (Supp. Figure 2). Indeed, three out of the four youngest active FL-L1s, termed 2q24.1, 6p24.1 and 6p22.1-2, are *Ta-1* copies reported to be extremely active in cancer genomes (Rodriguez-Martin et al. 2020). In contrast, 1p12 is a fixed *pre-Ta* insertion that despite integrating into the human genome approximately 1.8 Myr ago, remains highly active both in the germline (Gardner et al. 2017) and somatically within tumours (Tubio et al. 2014; Jung, Choi, and Lee 2018; Rodriguez-Martin et al. 2020). This suggests that a small set of *pre-Ta* representatives possibly remain very active in the human population and might still be relevant for cancer development.

Collectively, this work provides a valuable collection of annotated and sequence-resolved L1 source elements mobilizing in the human germline (https://zenodo.org/records/4475905). The analysis of their complete genetic sequences showed that cancer-associated FL-L1s have diverse and in some cases ancient evolutionary origins. Some of the older elements with disrupted ORFs, might sustain their activity through trans-complementation, but also could present polymorphic alleles in the population, some of which are still able to mobilize. Of note, some of the youngest L1 copies in our analysis (e.g. 6p22.1-1 and 2q24.1) have been reported to be rare polymorphisms able to drive massive bursts of somatic retrotransposition in cancer genomes (Rodriguez-Martin et al. 2020; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). This observation suggests that recently acquired L1s, which have not yet reached an equilibrium with the human host genome, might contribute disproportionately to disease-causing variation (Cordaux and Batzer 2009).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

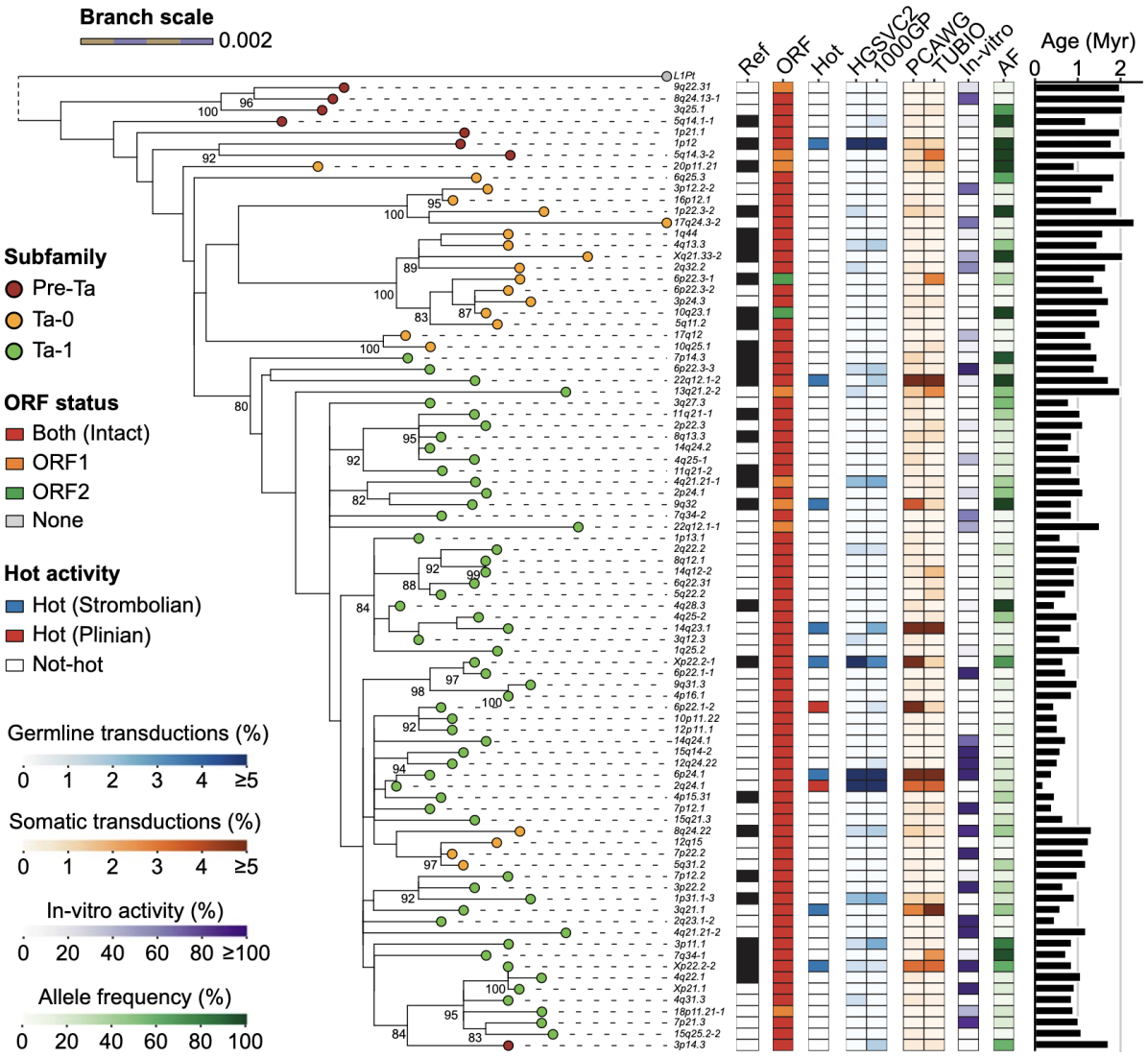


Figure 7: Phylogenetic tracing of L1Hs source elements.
Maximum-likelihood phylogenetic tree for highly active sequence-resolved FL-L1s annotated by subfamily designation, presence / absence on the reference genome, ORF content, and hot activity profile. Bootstrap values ≥ 80% are shown. Tree branch lengths are scaled according to the average number of substitutions per base position across the L1 sequence. Dashed lines map each L1 cytoband identifier to its corresponding branch on the tree. *Pan troglodytes* (L1Pt) is included as an outgroup. Heatmaps represent allele frequency (AF) based on the assembly discovery set, activity estimates based on in vitro assays, and the number of transduction events detected in human populations or cancer studies.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C1.2.2 - New computational methods to study L1 retrotransposition with long reads

To address the study of the sources of L1 somatic retrotransposition in human cancer, we developed new computational strategies mining the resolution provided by long reads. These methods were developed within the framework of the Mobile Element Insertion Genome Analyser (MEIGA), a retrotransposition detection tool developed by our research group at Universidade de Santiago de Compostela.

MEIGA is a bioinformatic tool that detects and characterizes retrotransposition events in whole-genomes sequenced with long-reads. It is specially designed for the somatic analysis of tumours and matched-normal genomes. Remarkably, MEIGA can identify canonical insertions (including solo-L1s and 3' orphan and 3' partnered L1 transductions), but also more complex variation as the one represented by processed pseudogenes and L1-mediated rearrangements.

To contribute to the development of MEIGA and enhance its capabilities, we developed two additional programs. The first one, MEIsimulator, is a program that generates in silico mobile element insertions along the human genome, which can be used to assess the performance of retrotransposition callers. MEIsimulator enables the simulation and distribution of diverse retrotransposition insertions, with different features and configurations. Importantly, it implements the simulation of both long and short reads, with variants acquired at predefined levels of cell clonality. The second one, Source Inference, is a specific method to trace the origin of somatic L1 insertions based on diagnostic substitution profiling as provided by long-read sequencing reconstruction. The rationale is that during transcription and retrotranscription, nucleotide substitutions present in a source element are transferred to their derived copies, allowing to potentially link each somatic insertion to a particular element-of-origin (Scott et al. 2016). This method was used to trace the source locus of somatic L1 solo insertions, being to our knowledge the first bioinformatic tool that systematically addresses this challenge.

C1.2.2.1 - Development of the MEIsimulator

The MEIsimulator implements a bioinformatic workflow that generates mock genomes peppered with specific retrotransposition events at multiple chromosome locations (Figure 8). The program takes as input three files. First, the master configuration file, which provides a table-based platform to specify the required number of retrotransposition events to generate, under each retrotransposition class (L1, Alu or SVA) and configuration. Second, the retrotransposition database in FASTA file format, which contains the collection of mobile element sequences that will be integrated as commanded by the configuration file. Third, the reference sequence, which provides the genome to be modified (GRCh38 by default but can be easily adapted for other assemblies or species). Additionally, coverage and variant allele frequency (VAF) must be indicated. Additional input files and parameters can optionally be included to feed the program with other customizable features (e.g. specific retrotransposition size distributions, sets of target chromosomes, sequencing read profiles). Finally, MEIsimulator requires to set the location of the output files, which essentially include an aligned, sorted and indexed binary aligned map (BAM) file and a report file describing the genetic features of the simulated retrotransposition events.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

To generate its simulations, MEIsimulator implements the following three core stages:

I - Distribution of in silico retrotransposition events
In this stage, retrotransposon sequences (L1, Alu or SVA) are sampled from the retrotransposition database according to the counts specified in the master configuration file. For each retrotransposition event, specific features such as poly-A tails and downstream transduced sequences are added according to the configuration associated with each integration mode. To enhance sequence diversity among integrated sequences, a specified rate of point mutations or distinct truncation patterns can be introduced into the sequences. Finally, each retrotransposition event is assigned a random genomic insertion position and a strand. This assignment is performed by sampling chromosome coordinates within 10 KB genomic windows, with the condition of avoiding GAP-associated positions.

II - Genome modification
In this stage, the reference genome is modified to accommodate the set of retrotransposition events. This is done rebuilding the chromosomes impacted by the simulated retrotransposons on their insertion points. For each retrotransposition event, upstream and downstream chromosome sequences are split, and the retrotransposon sequence is incorporated to the locus under a specific orientation, associated with TPRT features. In this stage, L1-mediated structural variation is also added, modelled as a deletion, duplication, inversion or translocation of the surrounding genomic region. As a result, a modified version of the reference genome is generated, incorporating the complete set of retrotransposition events and their associated features across the entire genome sequence.

III - BAM file generation
At this stage, the genome peppered with retrotransposition events is scanned to generate in silico reads from the desired sequencing technology up to the required depth of coverage. MEIsimulator implements two alternatives for the simulation of the sequencing reads. If short reads are required, the program implements a workflow that relies on ART v2.5.8 ((Huang et al. 2012) for sequencing read simulation, BWA-MEM v0.7.17 for alignment and Samtools v1.9 for processing BAM files. If the user requires long-reads, the program uses PBSIM (Ono, Asai, and Hamada 2013) as the read simulator, followed by minimap2 v2.17 (H. Li 2018) and Samtools v1.9 for aligment and BAM file processing respectively. Somatic retrotransposition events can be subclonal, being present only in a subset of cells. Additionally, they generally impact only one chromosome of a homologous pair. As a result of both factors, somatic retrotranspositions are profiled only by a fraction of sequencing reads at their integration points. To represent this scenario, MEIsimulator samples and combines reads from both modified and unmodified genome according to a predefined VAF threshold (e.g. 50%). After the final merging step of the sequencing data, MEIsimulator generates a sorted and indexed BAM file that can be analysed with different retrotransposition callers.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

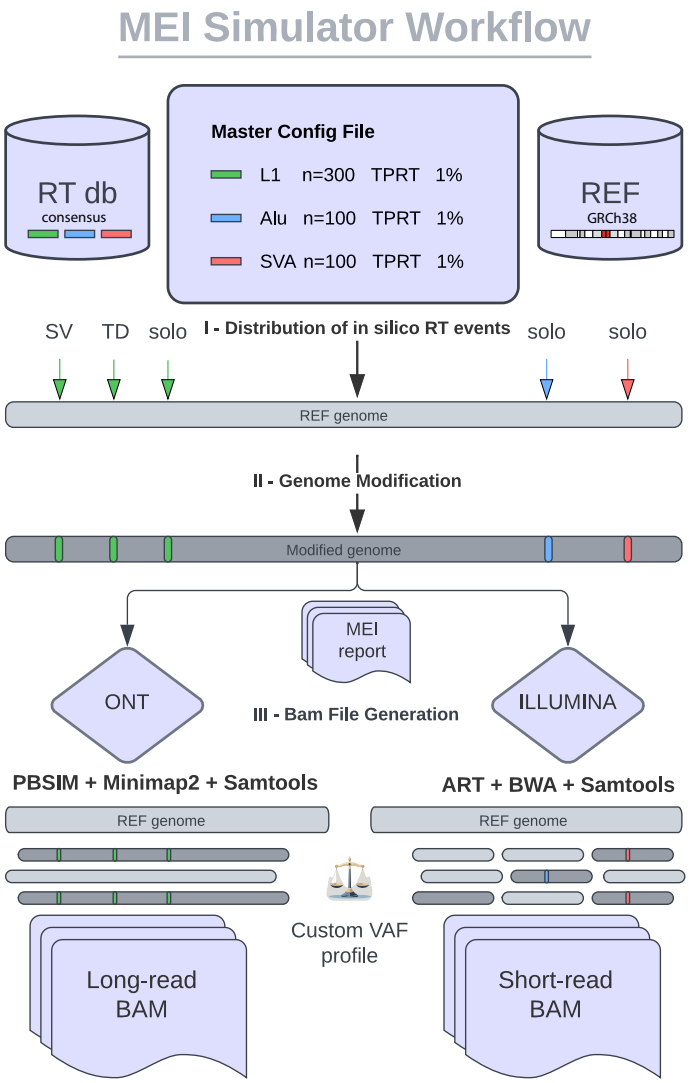


Figure 8: The workflow of MEIsimulator.
Schematic representation of the MEIsimulator pipeline, allowing the simulation of mock genomes with mobile element insertion-associated events under long-read (ONT) and short-read (ILLUMINA) sequencing profiles. Step 1: Distribution of in silico retrotransposition events. Step 2: Genome modification. Step 3: BAM file generation.

The MEIsimulator tool is written in Python, and it was developed within the framework of the MEIGA project (Mobile Genomes & Disease Laboratory, CiMUS, USC). The program is available at: https://gitlab.com/mobilegenomesgroup/MEIsimulator

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Figure 9: IGV view of a simple retrotransposition event generated with MEIsimulator.
Somatic L1 partnered transduction (chr6:45014950) of 4,660 bp integrated in the positive strand showing characteristic TPRT features. Top: Long-read bam file with ONT sequencing read profile. Bottom: Short-read bam file with paired-end ILLUMINA sequencing profile. In both cases the event was simulated according to VAF 50% profile and sequencing coverage 33.7X, with reads aligned against the GRCh38 reference genome.

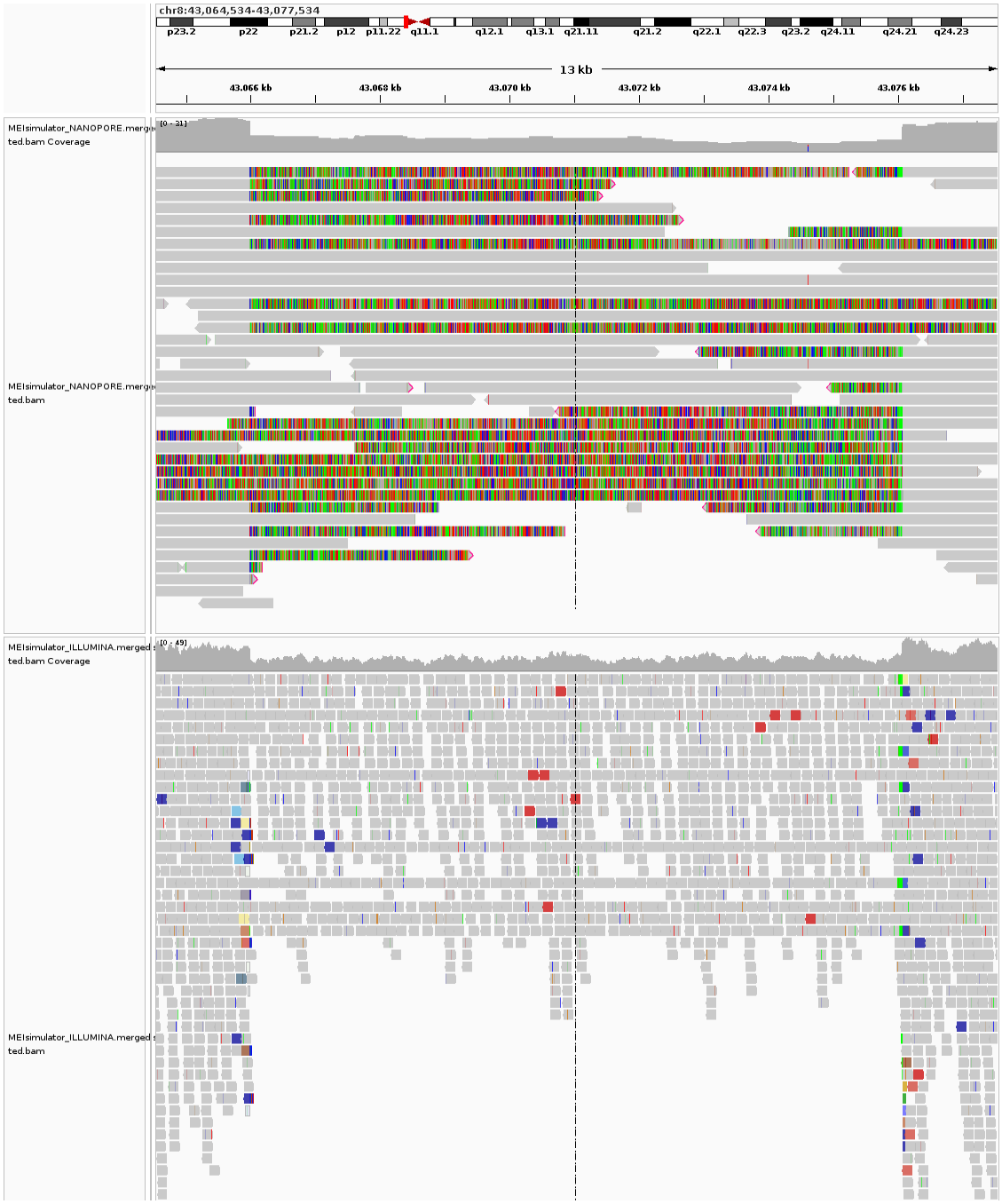


Figure 10: IGV view of a complex retrotransposition generated with MEIsimulator.
Somatic L1-mediated deletion of 10 KB (chr8:43065993-43076077 region showing drop of coverage) including a L1 retrotransposition bridge of 2.5 KB. Top: Long-read bam file with ONT sequencing read profile. Bottom: Short-read bam file with paired-end ILLUMINA sequencing profile. In both cases the event was simulated according to VAF 50% profile and sequencing coverage 33.7X, with reads aligned against the GRCh38 reference genome.

Table 2: Computational monitoring of MEIsimulator executions.

VAF	Time (h:m:s)	MaxRSS	MaxVMSize	MaxDiskWrite	MaxDiskRead
15%	20:57:02	18.47G	0.21G	0.18M	0.05G
30%	21:05:48	16.39G	0.21G	0.18M	0.05G
50%	22:36:55	17.28G	0.21G	0.18M	0.05G

Human mock genomes seeded with 6480 synthetic retrotranspositions, sequenced to 33.7X on short and long-read in silico datasets. Abbreviations: VAF=variant allele frequency; MaxRSS=maximum resident set size of all tasks in job; MaxVMSize=maximum virtual memory size of all tasks in job; MaxDiskWrite=maximum number of bytes written by all tasks in job; MaxDiskRead=maximum number of bytes read by all tasks in job.

C1.2.2.2 - Development of the Source Inference

The Source Inference method implements a bioinformatic workflow that identifies L1 diagnostic variants and uses them to trace somatic retrotransposition events (e.g. solo L1 insertions) to specific L1 source elements according to the shared set of mutations found across their sequences.

The Source Inference method requires polished sequences for both somatic insertions and candidate L1 source elements, ideally derived from phased long-read sequencing data. Non-phased data can also be used, although it may reduce resolution (see Methods: "Somatic retrotransposition calling on long-reads" and "Reconstruction of source L1 elements repertoire"). Candidate source elements are defined as those FL-L1s found in the genome of the patients with divergence levels associated with somatic activity (<1% divergence against a L1Hs consensus [NCBI accession: L19088.1]) (Ebert et al. 2021). In order to mitigate the impact of ONT sequencing errors in homopolymers (Wang et al. 2021) which might lead to incorrect L1 assignments, single base substitutions are the only type of variants profiled by the current version of the Source Inference strategy.

To generate its assignments, Source Inference implements the following three steps (Figure 11):

I - Identification of diagnostic nucleotides

First, an initial alignment of L1 candidate source element sequences is made against the consensus L1 sequence with the long-read aligner minimap2 v2.24 (H. Li 2018) with default parameters. This alignment is required to determine the informative status of each variant according to the specific pool of candidate sources found in the cancer patient. After performing a pileup of the aligned candidate source sequences with Rsamtools v2.10.1 (Danecek et al. 2021), every substitution in the alignment is classified as diagnostic (private from a single candidate source element) or regular (shared by several candidate source elements). Both classes are relevant for our strategy, since they track sequence identity between insertions and candidate sources. However, only the diagnostic variants will be instrumental to trace specific elements-of-origin.

II - Annotation of L1 nucleotide variants

Second, the sequences of somatic L1 insertions and candidate source elements are aligned against the same L1Hs consensus in an iterative manner. This step is done with Biostrings v2.62.0 local pairwise alignment tool, which implements the Smith-Waterman algorithm (Smith and Waterman 1981). The alignment of L1 insertions is performed after a reconfiguration step where the inverted L1 sequences are returned to their primary orientation. After each pairwise alignment step, the variants from somatic L1 insertions and candidate sources are annotated with Biostrings v2.62.0 mismatchTable function, so each sequence is annotated with its collection of nucleotide substitutions which will be later used for assignment purposes.

III - Inference of source elements

Third, in the central classification step, every somatic L1 insertion is compared against the pool of candidate source elements to search for a potential element-of-origin. This comparison is always made across the homologous region between the insertion and each candidate source, which typically comprises the 3' end of the L1 sequence, due to the 5' truncation of most somatic derived copies in cancer (Rodriguez-Martin et al. 2020). For each comparison, L1

diagnostic and regular variants found across this region are profiled, and attribution of the insertion to a particular source is made only when the following conditions are met: 1) There is a 100% ratio of diagnostic variants shared between insertion and the potential element-of origin. 2) No other candidate source element has a 100% ratio of diagnostic variants shared with the insertion. 3) There is a ≥75% ratio of total variants shared between insertion and the potential element-of-origin. These criteria enable a specific determination of the most-likely L1 element-of-origin, although we cannot totally rule out cryptic sources represented by identical somatic FL-L1 copies retaining somatic activity, or candidate source elements undetected in the host genome. Finally, the set of somatic L1 insertions with successful source inference assignments is collected and exported for further analysis and visualization display on CIRCOS graphics (Gu et al. 2014).

The Source Inference method is written in R programming language. It was developed within the framework of the MEIGA project (Mobile Genomes & Disease Laboratory, CiMUS, USC). The program is available at: https://gitlab.com/mobilegenomesgroup/SourceInference

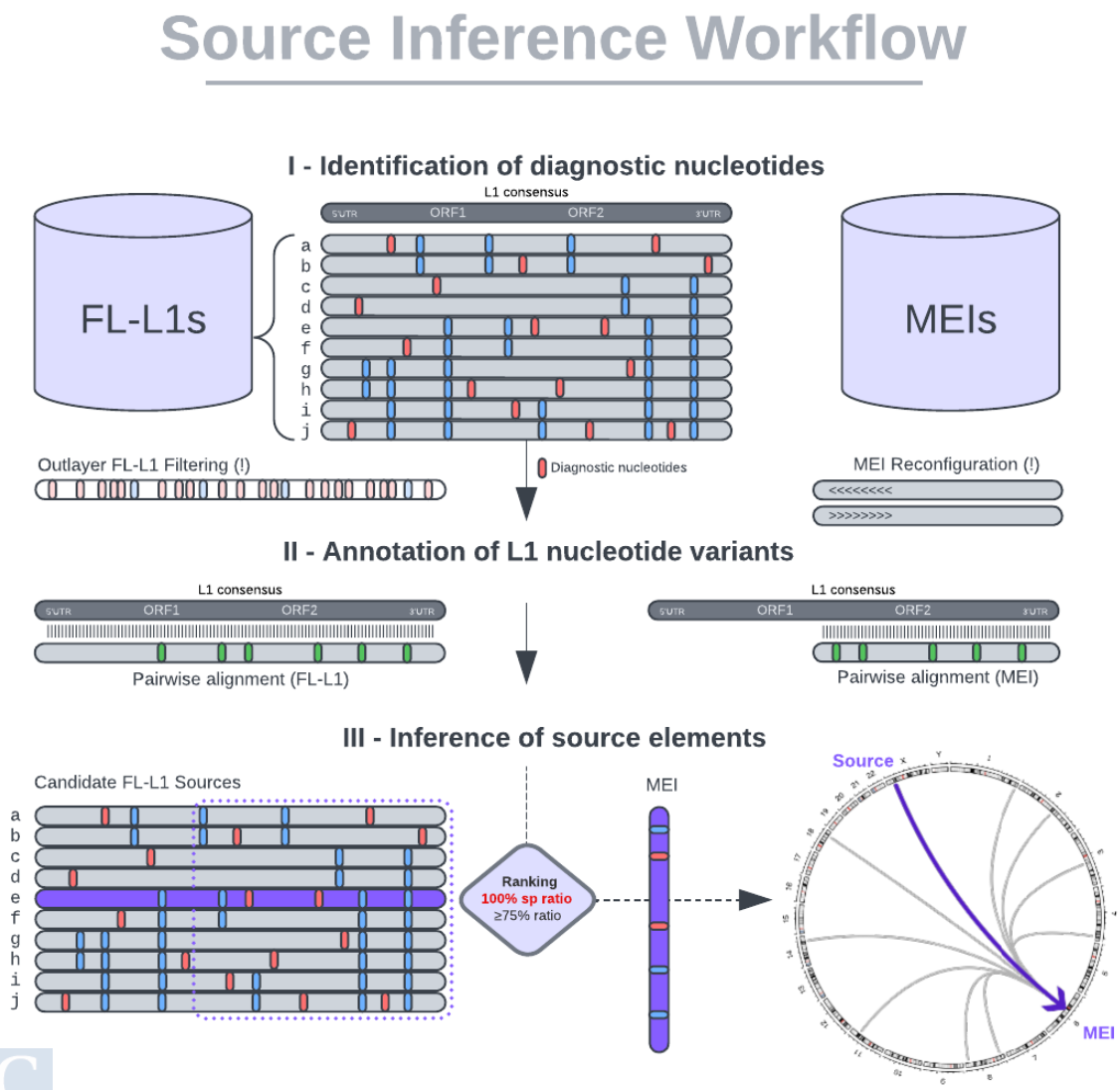


Figure 11: The workflow of the Source Inference method

Schematic representation of the Source Inference pipeline, which traces the element-of-origin for somatic L1 insertions based on the systematic profiling of diagnostic single nucleotide variants. In a first step, L1 diagnostic nucleotides are identified after aligning all candidate source FL-L1s in the patient genome. In a second step, somatic L1 insertions and candidate sources are aligned against consensus so each sequence is annotated with its specific set of variants. In a third step, every somatic L1 insertion is assessed against the pool of candidate source elements, and attribution of the element-of-origin is done only if there is specific concordance of diagnostic and regular variants across the retrotransposition sequence they share. Finally, traced events are represented in circos plots to visualize the L1 somatic mobilization patterns.

Evaluation of the Source Inference Method

To evaluate the specificity of the Source Inference method with real data, we employed a set of partnered transductions (n=388) detected in tumour genomes to analyse the concordance between source element inference derived from transductions and from solo-L1s (Figure 12). Partnered transductions are retrotranspositions that preserve both the L1 sequence and the downstream transduced region, making them a close-to-ideal scenario to test our algorithm. This analysis shows that 356 out of 388 partnered transductions were assigned to the same source element using both methods in an independent manner, being 91.8% the concordance between both strategies across the ten samples. We observed, however, that concordance varies with the length of the interrogated insertions, dropping off to 86% 117/136) when the analysis is restricted to partnered transductions with transduced regions below 200 bp, and increasing up to 94.8% (239/252) when delimiting the analysis to partnered transductions with transduced regions of at least 200 bp.

To further evaluate our method with in silico data, we used MEIsimulator to model the mobilization of all potential source elements (n=266) from the patient in the cohort with the highest L1 activity rate (PD0270), assuming a homogeneous activity of 20 mobilizations for all the elements and with the derived copies showing the same distribution of insertion size as in 2,954 cancer genomes from a previous somatic retrotransposition work (Rodriguez-Martin et al. 2020). The results demonstrated a high specificity of the pipeline (>99%) and an overall sensitivity of 47.6% (2,530/5,320). Sensitivity decreased when somatic L1s were derived from younger source elements, correlating with a reduction in genetic divergence (sensitivity=59.2% for *pre-Ta*; 47.0% for *Ta-0*; 36.8% for *Ta-1*). We also observed that the performance of the method is notably affected by the size of the derived copies, with a reduction in the sensitivity inferred from smaller derived copies, particularly in those copies derived from *Ta-1* source elements (Figure 13).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA


Source Inference Evaluation
Tracing Status
Concordant n=356 (91.8%)
Discordant n=32 (8.25%)
Partnered transductions

Figure 12: Evaluation of the Source Inference method with somatic retrotranspositions.
Top: Circos plot showing the tracing of somatic L1 mobilizations in tumours used to assess the specificity of the program. Green links: partnered transductions where both source inference and transduction analysis determine the same L1 element-of-origin (concordant events). Red links: partnered transductions where the source inference and transduction analysis determine different L1 elements-of-origin (discordant events). The ideogram represents chromosomes according to the human genome assembly GRCh38. Bottom: Raw table describing the represented partnered transductions events traced from both strategies (MEI indicated by arrows). Green squares denote concordant events (n=356 transductions), and red squares denote discordant events (n=32 transductions).

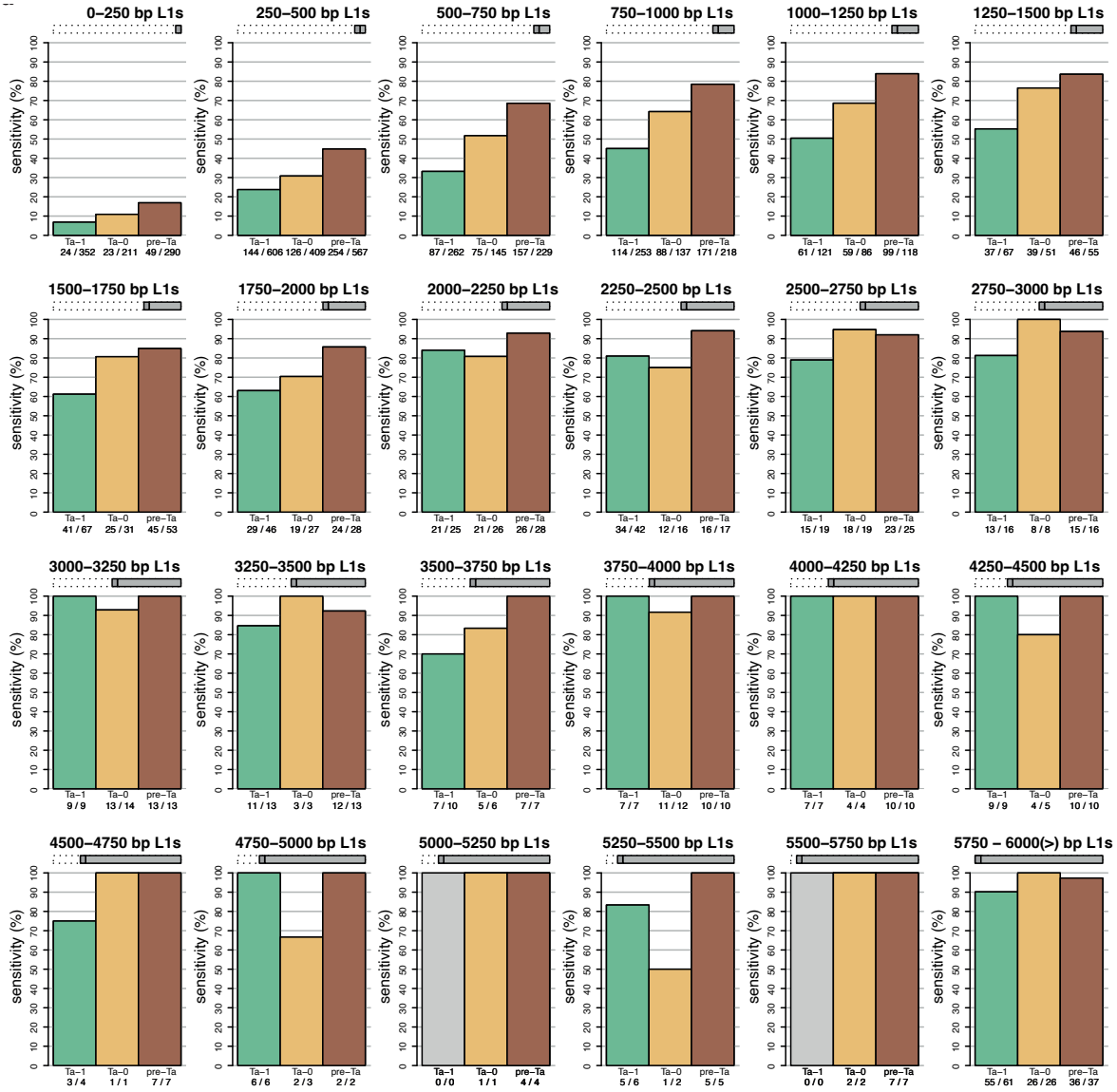


Figure 13: Evaluation of the Source Inference method with in-silico retrotranspositions.
The sensitivity of Source Inference is assessed, testing its ability to trace solo-L1 insertions to its original L1 source element as generated with MEIsimulator. Vertical bars indicate the fraction of events traced to its element-of-origin (true positive rate) under each specific insertion size range (5' truncation interval decreasing in 250 bp on each plot, as represented in the top L1 ideograms). The number of detected and simulated events for each category is indicated under each vertical bar. Bar colours indicate L1Hs subfamily: *Ta-1* (green); Orange: *Ta-0* (orange); *pre-Ta* (brown). Two size ranges without simulated insertions are denoted with grey vertical bars.

USC DE SANTIAGO DE COMPOSTELA

Collectively, these results demonstrate that Source Inference is a reliable tool to search the element-of-origin for somatic L1 insertions in cancer. This work demonstrates that the systematic profiling of internal L1 nucleotide variants have the potential to track a relevant number of solo-L1 insertions to their element-of-origin. While the approach faces intrinsic limitations, such as the common truncation of L1 somatic insertions (Rodriguez-Martin et al. 2020) and low sequence divergence across recently integrated FL-L1s (Chuang et al. 2021; Ebert et al. 2021), applying Source Inference to tumour sequencing data might offer a more comprehensive overview of retrotransposition dynamics in cancer.

C1.2.3 - Dynamics of somatic retrotransposition in the light of long reads

C1.2.3.1 - A novel panorama of source elements activity in the light of long reads.

The analysis of diagnostic nucleotides represents an opportunity for obtaining a better overview of the activity of L1 retrotransposition in cancer. We applied our Source Inference strategy to discover the element-of-origin of somatic L1 insertions in 10 cancer genomes (5 HNSC, 4 LUSC, 1 COAD) with high retrotransposition activity. These L1 events had been detected and characterized with MEIGA, taking advantage of the full-sequence reconstruction provided by this method.

Overall, Source Inference algorithm was able to trace 1,325 L1 events (including 899 L1-solos) to a particular source element in the genome of the patient (Figure 14). These events represented 44.5% (1,325/2,980) of the total number of traced retrotransposition events in MEIGA combining different approaches. The mobilizations traced by Source Inference had a substantial impact in the patterns of activity of many source elements. Remarkably, 688/899 (76.5%) of the newly traced L1-solos were attributed to source elements of known activity in cancer according to PCAWG studies, including 294 events linked to 14 source elements with known "hot" activity in cancer genomes (Rodriguez-Martin et al. 2020; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), highlighting its somatic activity. In addition to the 151 source elements identified in the MEIGA cohort by their derived transductions, another 46 cryptic elements emerged as donors of somatic solo-L1s under our diagnostic nucleotide's strategy. These cryptic source elements were typically *Ta-1* copies (25/46; 54.3%) and were elements often (33/46; 71.7%) present in the reference genome.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

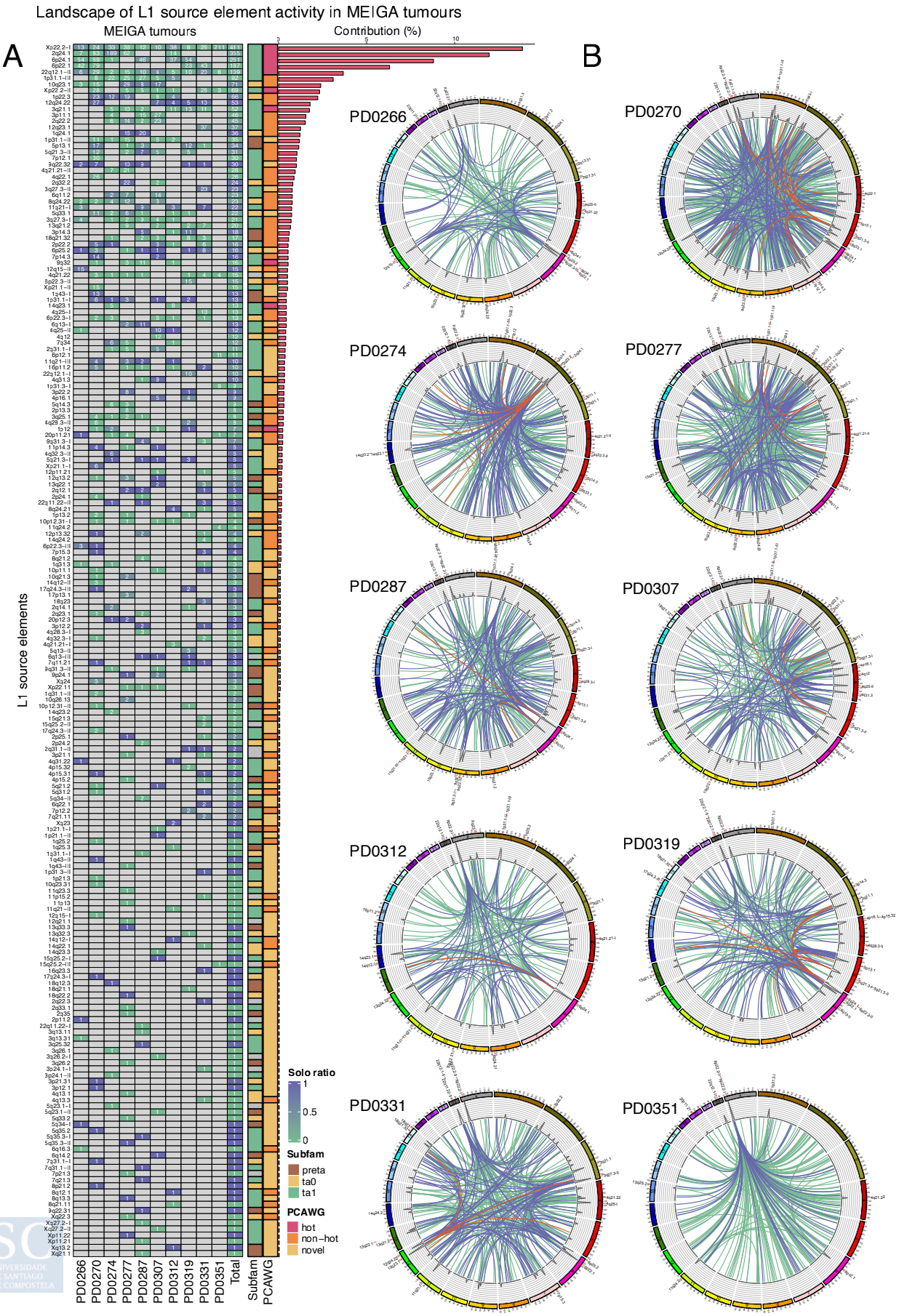

Landscape of L1 source element activity in MEIGA tumours
A
MEIGA tumours
Contribution (%)
L1 source elements
PD0266
PD0270
PD0274
PD0277
PD0287
PD0307
PD0312
PD0319
PD0331
PD0351
Total
Subfam
PCAWG
Solo ratio
1
0.5
0
Subfam
preta
ta0
ta1
PCAWG
hot
non−hot
novel
B
PD0266
PD0270
PD0274
PD0277
PD0287
PD0307
PD0312
PD0319
PD0331
PD0351

Figure 14: Overview of the L1 source element activity in MEIGA tumours.
A) Heatmap indicating the total number of somatic mobilizations traced to each active L1 source element (rows) on each MEIGA tumour (columns). The heatmap cell colour ranges according to the solo-ratio (ratio between activity entirely derived from solo-insertions (purple) and activity entirely derived from transductions (green)). The right annotation track indicates: L1Hs subfamily (green: *Ta-1*, orange: *Ta-0*, brown: *pre-Ta*); PCAWG somatic activity (red: PCAWG hot source element, orange: PCAWG non-hot source element, yellow: novel activity); Contribution: L1 source elements are sorted according to their relative contribution (%) to the total number of traced mobilizations in the cohort. B) Circos plots showing L1 source element somatic mobilizations in MEIGA tumours. Somatic mobilizations include canonical transductions (green links), ambiguous transductions (orange links) and solo-L1 insertions (purple links). The genomic locations of the 20 most-active source elements are indicated with cytoband labels. The ideograms represent chromosomes according to the human genome assembly GRCh38.

Interestingly, we discovered that certain active L1 loci generally propagated by generating somatic transductions, while others predominantly generate solo insertions. Despite our intrinsic limitation to trace short, derived insertions, we observed that in 37% (73/197) of the source elements the number of detected solos was higher than the number of detected transductions. For instance, in tumour PD0277a, the diagnostic nucleotide strategy identified up to 12 active sources that were missing in the transduction analysis, three of which (1q24.1, 9q22.32 and 3p22.2) rank within the 20 most active sources in the tumour after accounting for derived solos (Figure 15). Other source elements, although already discovered through the transduction analysis, notably increased their activity when derived solos were added. For instance, in PD0277a, the source at 2q32.2 increased its activity from 3 transductions to 22 mobilizations, after adding 19 derived solos, moving the source from the twentieth to the fifth position in the activity ranking of the tumour (Figure 15). Similar results were found in tumour PD0287a, were the diagnostic nucleotide strategy identified 15 active sources that were missing in the transduction-based analysis, four of which (1q24.1, 6q13-I, 9q22.32, 11q21-III) rank within the 20 most active sources after accounting for derived solos (Figure 15). Interestingly, the activity of the top source element in PD0287a (6p24.1) increased 1.6-fold after the inclusion of 18 new traced insertions, accentuating its mutagenic relevance in the tumour genome. Similar patterns were found on the dominant source elements of PD0274 (2q24.1), PD0312 (Xp22.2-I), PD0319 (6p24.1) and PD0351 (Xp22.2-I), with significant increases after accounting for solo-L1 insertions (Figure 15).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

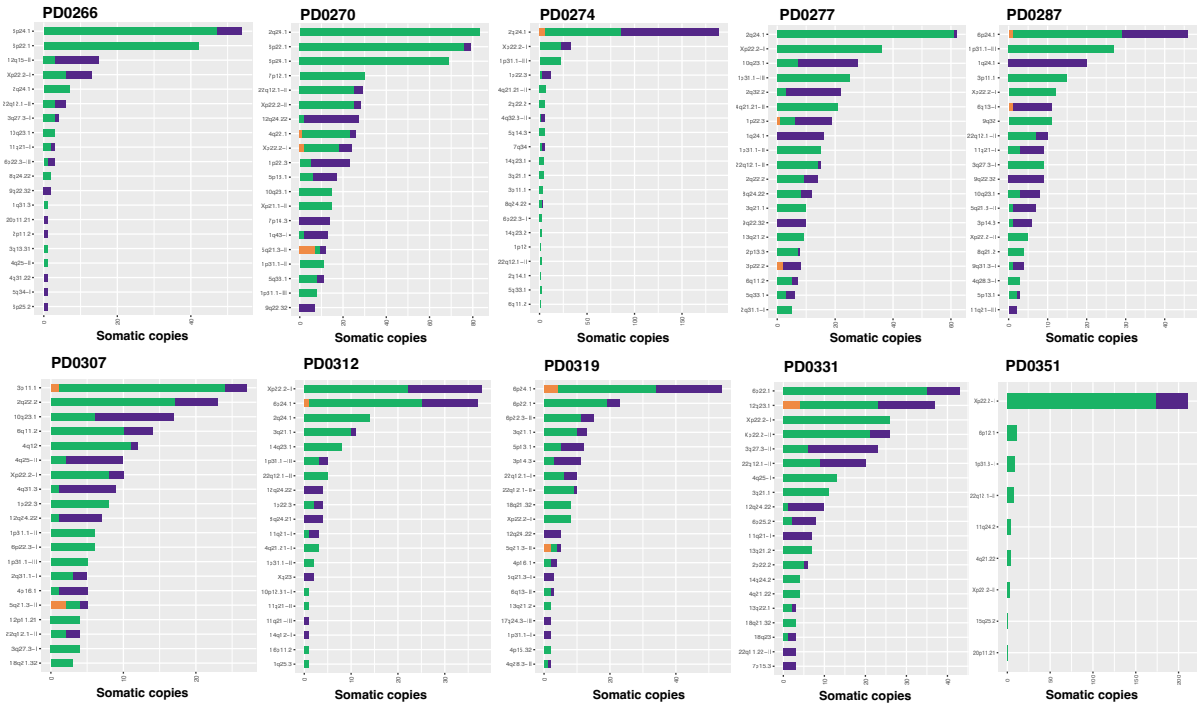


Figure 15: Patterns of mobilization of dominant source L1 elements in MEIGA tumours.
Horizontal bars indicate the number of somatic mobilizations detected for each active source element, including different categories of events: solo-L1 (purple), transductions (green) and ambiguous transductions (orange). Source elements are denoted with GRCh38 cytoband identifiers. The top 20 active source elements per tumour are shown (Y axis), ranked by activity level measured in terms of derived somatic copies (X axis).

During our analysis, we found 200 instances of partnered transductions in which the transduced sequences were too short to allow a specific identification of the source element. We termed these events "ambiguous transductions". Taking advantage of our new source inference strategy, we used the diagnostic nucleotides present in the companion L1 sequences to identify the element-of-origin from these insertions. Although in most of the events the 5' truncation of the insertions led to short L1 companion sequences, our inference method successfully assigned 19% (38/200) of the events to a particular element-of-origin (Figure 16). In one interesting example, our algorithm assigned 11 ambiguous transductions to a source element at 5q21.3-II, representing more than one-third (35%, 11/31) of the total derived retrotranspositions from that source in the cohort. These results highlight the potential of diagnostic nucleotide tracing to improve our understanding of dynamics of somatic retrotransposition in cancer.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

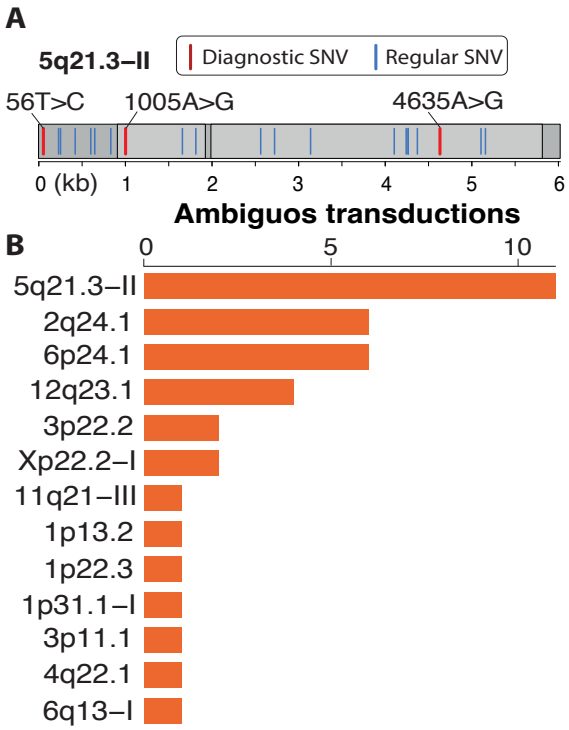


Figure 16: Tracing ambiguous transductions. A) L1 ideogram showing the regular and diagnostic variants used to trace somatic mobilizations to the source element located at the 5q21.3-II genomic region. In this source element, the following diagnostic single nucleotide variants were identified: 56T>C, 1005A>G and 4635A>G. B) Bar plot showing the total number of ambiguous transduction events (X axis) assigned to different L1 source elements (Y axis) in the cohort of MEIGA tumours. A total of 38 ambiguous transductions were traced to a particular element-of-origin under the Source Inference approach. Ambiguous transductions are represented by partnered transduction events bearing a short non-repetitive sequence on their 3' end which is insufficient to determine with confidence the L1 source element-of-origin of the somatic insertion, but now can be traced through the complete analysis of the internal L1 cognate sequence.

In view of these results, we took advantage of the addition of derived-solo copies to characterize the activity of the most active elements in the cohort: the so-called "hot-L1s" 1 (see Methods). Our analysis identified 27 hot-L1 sources (Figure 17), including 19 elements from the youngest subfamily *Ta-1*, eight *Ta-0* elements, and one element from the ancient subfamily *pre-Ta*. Two hot source elements, at Xp22.2-I and 2q24.1, stand out above all the others giving rise to 411 and 355 somatic insertions (Figure 17), respectively, accounting for 12% (766/6418) of the total somatic retrotransposition in the cohort. In these two sources, derived solos account for 19% (79/411) and 30% (105/355) of the copies assigned respectively to Xp22.2 and 2q24.1, representing a substantial contribution to the activity of both source elements in the cohort (Figure 17). Overall, the addition of solo-L1s increases in average 1.37-fold the activity of hot-L1s relative to a transduction strategy, highlighting their relevance as mutagenic elements in the cancer genome.

Notably, the hot-L1 dataset contains eight elements in which the number of assigned solos exceeds the number of assigned transductions (Figure 17). Notably, a highly active element at 1q24.1 generated up to 36 solo-L1s in the cohort, 16 in tumour PD0277a and 20 in PD0287a, and no transductions at all. Other hot-L1 sources, although already supported by transductions, significantly increased their activity when accounting for derived solos. For instance, in the hot source at 12q24.22, solo-L1s represent 93% (49/53) of the total derived copies, placing this element within the 10 FL-L1s with the highest activity in the cohort. Similarly, 67% (44/66) of the copies from a hot element at 1p22.3 are attributed to solo-L1s. Of note is the source element at 1q43, a fixed *pre-Ta* element which represents one of the oldest hot active FL-L1s in the MEIGA cohort. This source is active in tumour P0270a giving rise to 13 derived copies of which 11 are solo-L1s.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

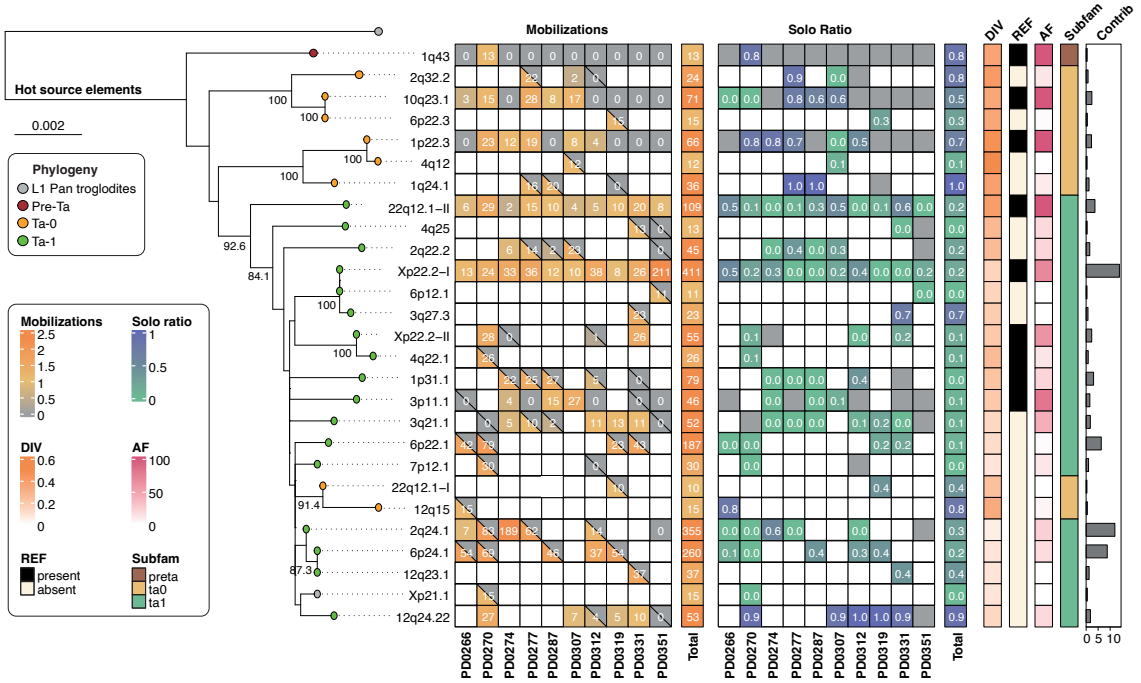


Figure 17: The landscape of hot L1 source element activity in MEIGA tumours.
Maximum-likelihood phylogenetic tree for 27 MEIGA "hot" FL-L1s annotated by subfamily designation. Bootstrap values ≥ 80% are shown. Tree branch lengths are scaled according to the average number of substitutions per base position. Dashed lines map each L1 cytoband identifier to its corresponding branch on the tree. Pan troglodytes (L1Pt) is included as an outgroup. The left heatmap shows the number of mobilizations found in each MEIGA tumour genome (columns). Genotyping information is included, with white cells indicating homozygous absence (0/0), split cells indicating heterozygous presence (0/1), and coloured cells indicating homozygous presence (1/1). The right heatmap represents the solo ratio, ranging from activity entirely derived from solo-L1 insertions (purple) to activity entirely derived from transductions (green). Hot source elements are annotated according to their sequence divergence, presence in the reference genome, allele frequency, L1 subfamily and relative contribution (%) to the total traced somatic mobilizations in the MEIGA cohort.

Collectively, these results demonstrate that the systematic profiling of L1 diagnostic nucleotides can trace a substantial fraction of the somatic retrotransposition activity in cancer whose origin was unknown. We found that certain L1 sources tend to mobilize primarily generating derived solo insertions, while others tend to preferentially generate transductions. We explored a biological aspect that might explain this pattern: the intensity and distribution of polyadenylation signals (PAS), which are biochemically involved in the termination of L1 transcription (Moran, DeBerardinis, and Kazazian 1999; Victoria P. Belancio, Whelton, and Deininger 2007).

We conducted a PAS analysis with APARENT (Bogard et al. 2019) neural networks in order to detect these sites in the reconstructed L1 source elements and their downstream regions. Despite more extensive analysis being required, we found clear instances where the distribution of the PAS was determining the transduction rates of the associated L1 source elements (Figure 18). In one remarkable example in de HNSC tumour PD0270, the source element located at 2q24.1 generated 83 derived copies, all of which were transductions with no solo insertions. Notably, most of these transduction events terminated 234 bp downstream from the L1 locus. Examination of the APARENT predictions revealed the absence of a canonical PAS at the end of the source L1 element. Instead, the first alternative PAS was predicted to be 234 bp downstream from the 3' end of the L1, matching the endpoint of the derived transductions. In another example from the same patient, the source element located at 1p22.3 predominantly

resulted in solo insertions (n=18) compared to transductions (n=5), with transductions ending 64 bp downstream the source element. Interestingly, the APARENT analysis in this case revealed a robust PAS at the canonical end of the source element, followed by a weaker PAS at the transduction end. These observations suggest that the strength of canonical and alternative PAS plays a relevant role determining the transduction rate of L1 source elements. However, other factors, such as the limitation to trace shorter, less divergent L1 and downstream sequences, probably leads to significant biases when measuring the somatic activity of specific L1 copies.

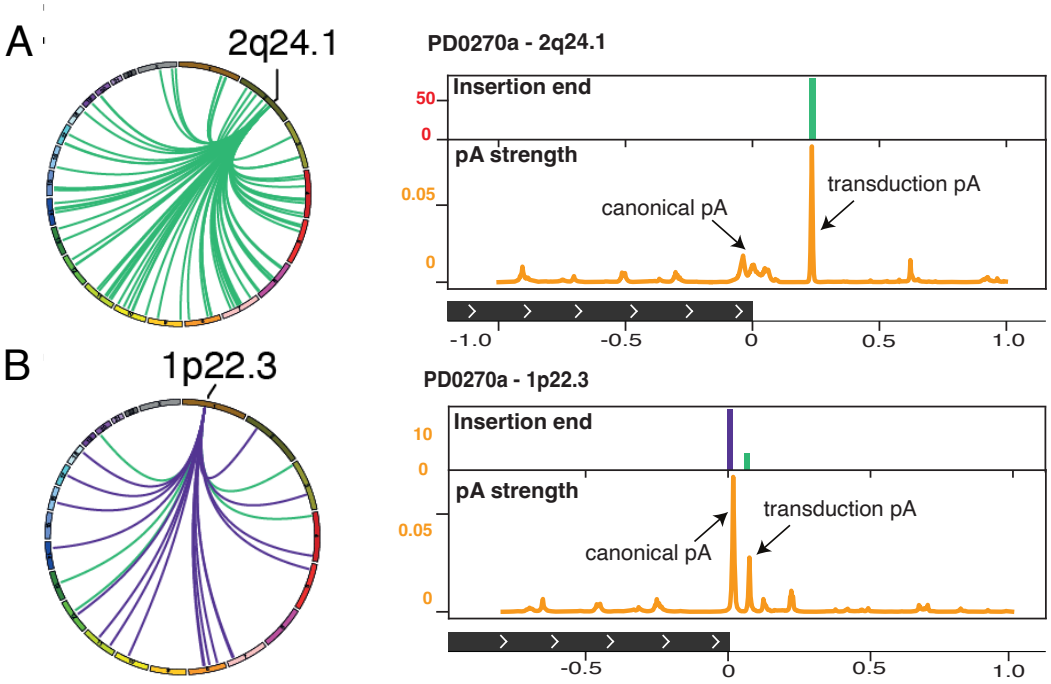


Figure 18: The role of polyadenylation signals in L1 source element transduction rates.
A) Circos plot showing the somatic mobilizations of the source 2q24.1 in the tumour PD0270, where it exclusively mediates transductions (n=83) including 234 bp of downstream sequence. The analysis of PAS identified a weak signal at the end of the L1 source element, but a strong alternative PAS 234 bp downstream from the 3' end of the L1, aligning with the endpoint of the derived transductions. B) Circos plot showing the somatic mobilizations of the source 1p22.3 in the tumour PD0270, where it predominantly mediates solo insertions (n=18) compared to transductions (n=5). The analysis of PAS found a stronger canonical termination signal at the end of the L1 element compared with the first alternative PAS predicted. Interestingly, the alternative PAS aligned with the endpoint of the derived transductions. Credit: Sonia Zumalave.

C1.2.3.2 - Instances of allelic heterogeneity in FL-L1 source elements

The Source Inference strategy is primarily designed to trace L1 insertions to specific L1 loci in the cancer genome. However, it can be also adapted to discriminate between alleles in those homozygous L1 source elements where somatic activity was detected. Apart from nucleotide substitutions that are source-locus-specific, our strategy can trace a second type of diagnostic sites called allele-informative substitutions. These are substitutions that discriminate between two L1 alleles from a given locus for a given patient. Here, our strategy first assigns solos to specific source elements using the locus-specific diagnostic substitutions, and then can be re-applied to assign those solos to one of the alternative alleles based on the set of allele-informative substitutions.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Taking advantage of the phased long-read sequencing data, we searched for instances where we could trace L1 somatic mobilizations not only to a particular locus, but also to a specific allele. To ensure the specificity of the analysis, we focused on homozygous loci with somatic activity confirmed by two complementary strategies (transduction and diagnostic nucleotide analyses).

Although the resolution of our analysis is limited by the short size of most derived solos (median size=850 bp), we found several instances in which our algorithms could unambiguously identify the specific allele whence they derive. For example, in patient PD027, the source locus at 1p22.3 is characterized by the source-specific substitution 5847G>A. Here, our analysis found two informative substitutions that discriminate between the two alleles (Figure 19). 5973 A>G in 1p22.3_allele1 and 5929A>T in 1p22.3_allele2. These diagnostic sites (and the evaluation of additional regular variants) allowed the unequivocal assignment of 20 derived L1 insertions (18 solo and 2 partnered events) exclusively to the first allele and none to the second. Remarkably, the analysis of the depth of coverage at the relevant region in the tumour suggests copy number loss involving 1p22.3_allele2 (Battenberg total CN = 3, minor CN = 0), being the most likely explanation for the absence of activity from the second allele. In another example, the source at 10q23.1 in patient PD0287 showed several differential substitutions: 5903G>T, 5468T>C, 4917C>G and 4084C>T in 10q22.3_allele1 and 6012T>G, 5048G>T, 4887T>C and 4574G>A in 10q22.3_allele2. Here, our pipeline assigned 6 derived L1 insertions (5 solo and 1 partnered event) exclusively to 10q23.1_allele2 and none to 10q23.1_allele1(Figure 19).

Overall, our method allowed us to trace 52 somatic L1 insertions to their specific allele-of-origin across 9 active L1 source elements in the MEIGA cohort (Table 3**,** Figure 20). Notably, we did not find instances of significant somatic activity derived from both L1 alleles from the same L1 locus in a cancer genome. However, since our strategy probably bias detection toward one particular allele at the source inference stage, we do not rule out this scenario might occur in some active source elements in the cancer genome. Interestingly, we found different haplotypes from the same source element showing somatic activity in different cancer patients (e.g. chr5_39788000 and chr10_85356000). Gaining insights into the allele-specific genetic and epigenetic regulation of these source elements might be relevant for improving our understanding of L1 biology and their activation dynamics in cancer.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

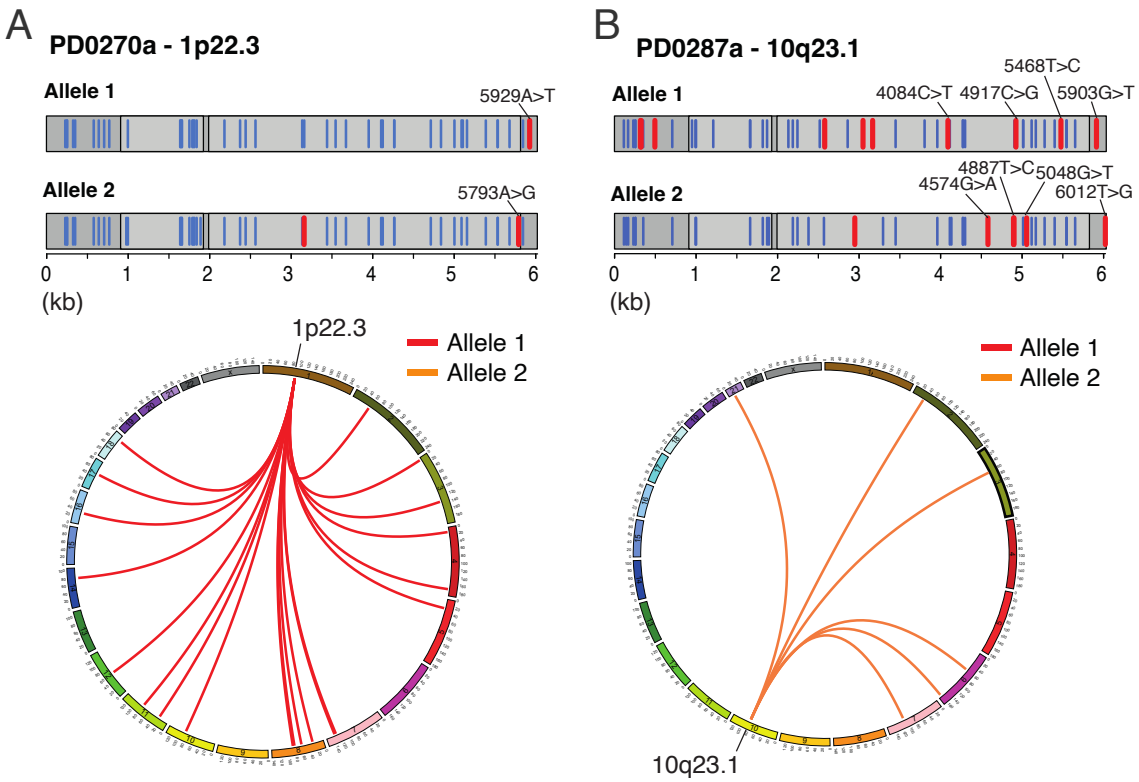


Figure 19: Active source elements with L1 mobilizations traced at allelic resolution.
A) On the top, L1 ideograms comparing the alternative alleles for the source element 1p22.3 in patient PD0270. Blue bars indicate regular variants, red bars indicate allele-diagnostic variants. On the bottom, Circos plot showing somatic mobilizations (n=20) traced exclusively to the 1p22.3 allele 1 in PD0270. B) On the top, L1 ideograms comparing the alternative alleles for the source element 10q23.1 in patient PD0287. Blue bars indicate regular variants, red bars indicate allele-diagnostic variants. On the bottom, Circos plot showing somatic mobilizations (n=6) traced exclusively to the 10q23.1 allele 2 in PD0287.

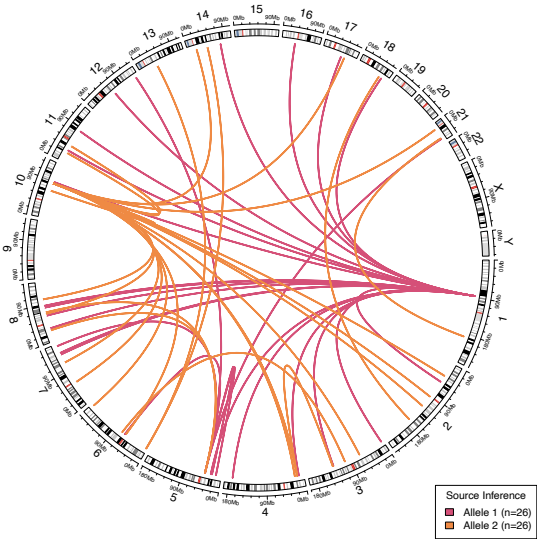


Figure 20: Inference of L1 somatic activity traced at allelic resolution.
Circos plot representing the complete set of somatic mobilizations confidently traced at allele resolution (n=52) on homozygous L1 source elements genotyped in the golden set of 10 MEIGA tumours. Red links represent somatic mobilizations traced to the first allele (haplotype 1) of the L1 source element, orange links represent somatic mobilizations traced to the second allele (haplotype 2) of the L1 source element. The ideograms represent chromosomes according to the human genome assembly GRCh38.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Table 3: Inference of L1 somatic activity traced to source loci at allelic resolution.

MEI	Sample	Type	Source transduction	Source Inference	Allele
chr10:89157939	PD0270	solo	Not applicable	chr1_86679000	hp1
chr11:29327508	PD0270	solo	Not applicable	chr1_86679000	hp1
chr11:83650896	PD0270	solo	Not applicable	chr1_86679000	hp1
chr12:63522224	PD0270	solo	Not applicable	chr1_86679000	hp1
chr14:79926929	PD0270	solo	Not applicable	chr1_86679000	hp1
chr16:32059392	PD0270	solo	Not applicable	chr1_86679000	hp1
chr17:47474054	PD0270	solo	Not applicable	chr1_86679000	hp1
chr18:64586964	PD0270	solo	Not applicable	chr1_86679000	hp1
chr2:51529118	PD0270	solo	Not applicable	chr1_86679000	hp1
chr3:4176867	PD0270	partnered	chr1_86679000	chr1_86679000	hp1
chr3:133211385	PD0270	solo	Not applicable	chr1_86679000	hp1
chr4:11971382	PD0270	partnered	chr1_86679000	chr1_86679000	hp1
chr4:171738538	PD0270	solo	Not applicable	chr1_86679000	hp1
chr5:25604956	PD0270	solo	Not applicable	chr1_86679000	hp1
chr7:132160445	PD0270	solo	Not applicable	chr1_86679000	hp1
chr7:136271674	PD0270	solo	Not applicable	chr1_86679000	hp1
chr8:31698836	PD0270	solo	Not applicable	chr1_86679000	hp1
chr8:63399368	PD0270	solo	Not applicable	chr1_86679000	hp1
chr8:83292770	PD0270	solo	Not applicable	chr1_86679000	hp1
chr8:87729385	PD0270	solo	Not applicable	chr1_86679000	hp1
chr12:126170012	PD0319	partnered	chr5_39788000	chr5_39788000	hp1
chr21:39874728	PD0319	solo	Not applicable	chr5_39788000	hp1
chr5:12866848	PD0319	solo	Not applicable	chr5_39788000	hp1
chr5:22613072	PD0319	solo	Not applicable	chr5_39788000	hp1
chr6:64187632	PD0319	solo	Not applicable	chr5_39788000	hp1
chr7:154206080	PD0319	solo	Not applicable	chr5_39788000	hp1
chr13:42526888	PD0270	partnered	chr6_2418000	chr6_2418000	hp2
chr14:47633584	PD0270	solo	Not applicable	chr6_2418000	hp2
chr18:57881181	PD0277	partnered	chr2_169249000	chr2_169249000	hp2
chr21:37919828	PD0287	partnered	chr1_187344000	chr1_187344000	hp2
chr2:34592377	PD0287	solo	Not applicable	chr10_85356000	hp2
chr21:13289097	PD0287	solo	Not applicable	chr10_85356000	hp2
chr3:64957231	PD0287	solo	Not applicable	chr10_85356000	hp2
chr6:62129503	PD0287	solo	Not applicable	chr10_85356000	hp2
chr6:165868460	PD0287	partnered	chr10_85356000	chr10_85356000	hp2
chr7:110750467	PD0287	solo	Not applicable	chr10_85356000	hp2
chr6:72229176	PD0287	partnered	chr3_103557000	chr3_103557000	hp2
chr8:26803203	PD0287	partnered	chr5_39788000	chr5_39788000	hp2
chr8:71501764	PD0287	solo	Not applicable	chr5_39788000	hp2
chr10:66712386	PD0307	partnered	chr10_85356000	chr10_85356000	hp2
chr11:22926962	PD0307	partnered	chr10_85356000	chr10_85356000	hp2
chr11:41866723	PD0307	solo	Not applicable	chr10_85356000	hp2
chr14:20873601	PD0307	solo	Not applicable	chr10_85356000	hp2
chr17:55605384	PD0307	solo	Not applicable	chr10_85356000	hp2
chr2:95423378	PD0307	solo	Not applicable	chr10_85356000	hp2
chr2:124233324	PD0307	solo	Not applicable	chr10_85356000	hp2
chr4:16304042	PD0307	solo	Not applicable	chr10_85356000	hp2
chr4:21773228	PD0307	solo	Not applicable	chr10_85356000	hp2
chr7:56503269	PD0307	solo	Not applicable	chr10_85356000	hp2
chr8:68361896	PD0307	solo	Not applicable	chr10_85356000	hp2
chr8:103630288	PD0307	solo	Not applicable	chr10_85356000	hp2
chr3:134826872	PD0319	partnered	chr4_15842000	chr4_15842000	hp2

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C1.3 – CONCLUSIONS OF CHAPTER I

I. Long-read sequencing technologies implemented by PacBio and ONT enable a comprehensive characterization of L1 source elements and their derived insertions. This advanced characterization allows to determine some relevant aspects of L1 biology, such as the integrity of their ORFs, their allelic variation and their diversification patterns, at unprecedented level of resolution.

II. Source Inference is a reliable bioinformatic method to trace L1 somatic retrotranspositions to their element-of-origin based on internal diagnostic nucleotides. While insertion length limits the sensitivity of the approach, our method successfully traces a significant fraction of somatic solo-L1 insertions back to their original source elements in tumours.

III. L1 activity in human cancer is driven by a diverse set of source elements belonging to multiple L1Hs sublineages. Notably, while some of these source elements predominantly generate solo-L1 derived insertions, others preferentially generate transductions, likely influenced by local sequence variation in their termination signals.

C1.4 - PUBLICATIONS

The results of the first section of this chapter ("*Analysis of FL-L1 Sources in the Human Germline*") have been published as a modest but significant contribution to an article in Science [1]. This publication represents a collaborative effort to advance our understanding of the impact of structural genomic variation in the human population, enabled by haplotype-resolved genome assemblies derived from long-read sequencing. Specifically, I contributed to the section of the manuscript concerning mobile element insertions (MEIs). My major contribution consisted in the estimation of L1Hs phylogenies based on the complete sequences of active FL-L1 elements. I generated multiple sequence alignments based on polished FL-L1 sequences and compared alternative substitution models to optimize ML tree inferences. I developed and implemented code to systematically assess the integrity of L1 open reading frames (ORF1 and ORF2), determine L1Hs subfamilies (*pre-Ta*, *Ta-0* and *Ta-1*) and obtain L1 age estimations based on sequence divergence. I performed statistical analyses assessing L1Hs subfamily age distributions. I analysed the phylogenetic distribution of L1 source elements with somatic activity in cancer. I created circos plots to visualize the genomic distribution of retrotransposition events. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work. Contents were adapted to align with the format of the doctoral thesis. Further detail regarding author's contributions is provided in the appendix, alongside the full list of authors and their affiliations.

References:

1. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537): eabf7117. doi:10.1126/science.abf7117

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER II: THE IMPACT OF GENOMIC VARIATION IN THE EVOLUTION OF BIVALVES´ TRANSMISSIBLE NEOPLASIA

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER II

The impact of genomic variation in the evolution of Bivalves' transmissible neoplasia

C2.1 - BACKGROUND

C2.1.1 - Transmissible cancer in bivalves

C2.1.1.1 – Clonal transmissible tumours: A new paradigm in cancer biology

Traditionally, the cancer paradigm implies that tumours are transient lineages of somatic cancer cells, fuelled by genetic mutations that drive their expansion confined within the host organism. However, in transmissible cancer this paradigm is broken, since certain cancer cells developed the ability to overcome the lifespan of their host, metastasizing not only to alternative tissues and organs, but also to an entirely new host in the population (Table 4) (Ostrander, Davis, and Ostrander 2016). According to this new perspective, transmissible cancers can be defined as potentially immortal somatic cell lineages that are transmitted between individuals via the direct transfer of living cancer cells.

Transmissible cancer biology has now been studied for decades, with focus on the two mammalian species affected: the domestic dog *(Canis lupus familiaris),* and the Tasmanian devil (*Sarcophilus harrisii).* The first species is affected by an unique millenary cancer clone known as Canine Transmissible Venereal Tumour (CTVT), which has remarkably achieved global distribution (Baez-Ortega et al. 2019). The second is affected by two independent, aggressive cancer clones known as DFTD-1 and DFTD-2, which are currently responsible for severe population decline in the host species (Stammnitz et al. 2023). Additionally, exceptional cases of cancer transmission have been documented in humans, including the dissemination of metastatic cells from the tapeworm *Hymenolepis nana* in an HIV-infected man (Muehlenbachs, Mathison, and Olson 2016), several instances of transplacental cancer transmission (Tolar and Neglia 2003), and documented cases of cancer propagation during surgical operations and other medical interventions (Gandhi and Strong 2007; Gärtner et al. 1996).

Table 4: General features of transmissible cancers identified in nature.

Cancer	Organism	Transmission	Precursor	Clone(s)	Age	Reference
CTVT	*Canis lupus familiaris*	Sexual contact	Myeloid cell	1	4,000 – 8,500 years	*Baez-Ortega et al. 2019*
DFTD	*Sarcophilus harrisii*	Biting	Schwann cell	2	> 20 years	*Stammnitz et al. 2023*
BTN	Bivalve species	Water filtration	Haemocyte	>8	> 40 years	*Metzger et al. 2016*

Abbreviations: CTVT=canine transmissible venereal tumour; DFTD=devil facial tumour disease; BTN=bivalve transmissible neoplasia.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.1.1.2 - Bivalve transmissible neoplasia (BTN)

Research using mammalian transmissible cancer models has led to significant advances in evolutionary genomics, immunology and molecular oncology (Baez-Ortega et al. 2019; Stammnitz et al. 2023; 2018; Rebbeck, Leroi, and Burt 2011; Ostrander, Davis, and Ostrander 2016), and holds promise for deepening our understanding of metastatic cancer biology. However, further research is required into a rare and fascinating form of transmissible cancer: that which affects bivalve molluscs in the marine environment. In these organisms, transmissible cancer has evolved independently multiple times, and cancer clonal lineages have been found affecting cockles, mussels, oysters and clam organisms (Figure 21). Affected species include common cockles (*Cerastoderma edule*), soft-shell clam (*Mya arenaria*), bay mussel (*Mytilus trossulus*), golden carpet shell clam (*Polititapes aureus*) and Baltic clam (*Macoma balthica*) among others (Table 5) (Metzger et al. 2016; Yonemitsu et al. 2019; Garcia-Souto et al. 2022; Michnowska et al. 2022; Hammel et al. 2022; Metzger et al. 2015). The disease affecting these bivalves is termed "haemic neoplasia" (HN), "disseminated neoplasia" (DN) or "bivalve transmissible neoplasia" (BTN). Although this term represents different conditions across bivalve species, these diseases have traditionally been grouped together because they share morphological characteristics and common patterns of disease development (María J. Carballal et al. 2015).

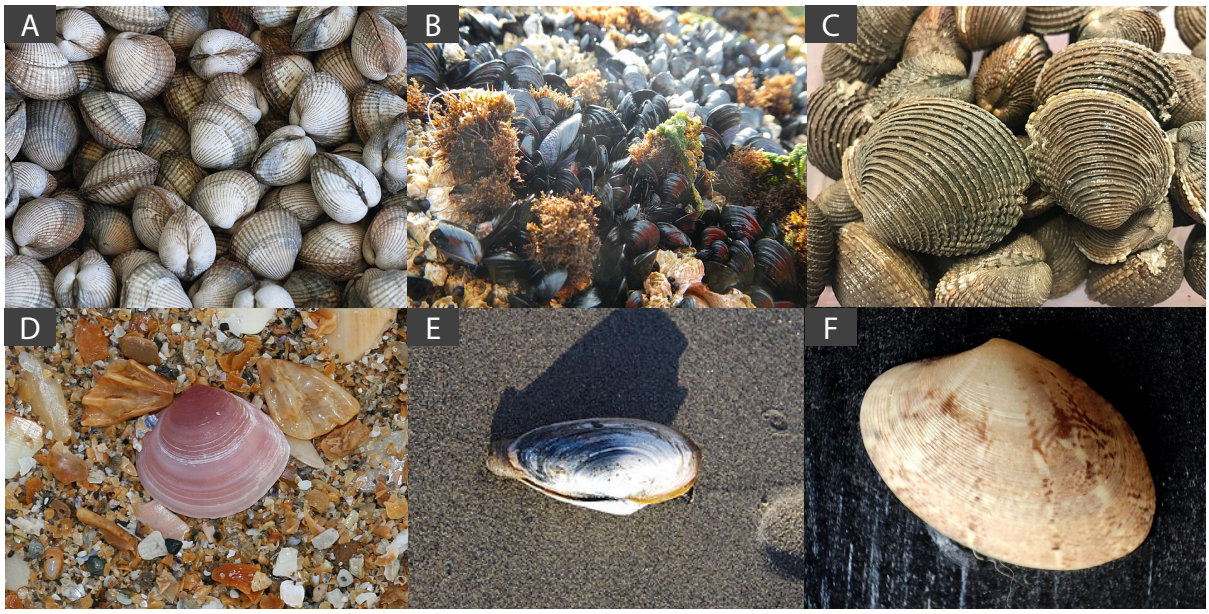


Figure 21: Bivalve species studied for being affected by transmissible cancer.
Transmissible cancers are a singular form of cancer that is spread between individuals through the direct transfer of tumour cells. Bivalve species affected by this type of cancer are displayed: A) *Cerastoderma edule* B) *Mytilus trossulus*; C) *Venus verrucosa*; D) *Macoma balthica*; E) *Mya arenaria*; F) *Polititapes aureus*. Sources: Wikimedia Commons and iNaturalist.org.

The first description of BTN was made in 1969 in oysters from the genus *Crassostrea (Crassostrea virginica* and *Crassostrea gigas),* being reported as a *"probable neoplastic disease of the hematopoietic system"* (Fanley 1969). The disease is considered a leukaemia-like condition, probably derived from the anomalous proliferation of haemocytes (Baez-Ortega and Murchison 2021). This cell type plays essential roles in the open circulatory system from these animals, including primary immune response, tissue repair and nutrient transport (Allam and Raftos 2015). Under physiological conditions, bivalve haemocytes are restricted to the internal media of the animal, not being able to survive in the seawater. However, in BTN, neoplastic

haemocytes physically escape from their host, spread through the marine water column and eventually reach and invade a new bivalve (Burioli et al. 2021; Giersch et al. 2022). The subsequent clonal expansion of tumour cells within the haemolymph and connective tissues of multiple organs often leads to the death of the affected animal, although spontaneous remission occasionally occurs (Garcia-Souto et al. 2022). In bivalve transmissible cancers, invasion is considered to occur during the water filter-feeding stage characteristic of these animals (Giersch et al. 2022), in contrast to the close physical contact required for transmission in mammalian transmissible cancers. The recurrence of this transmission process leads to the propagation of this transmissible cancer lineage across the bivalve populations, often reaching significant prevalence (María J. Carballal et al. 2015).

Table 5: Bivalve species known to be affected by transmissible bivalve neoplasia (BTN).

Species	Common name	Family	Genome	BTN	Source	Reference
Cerastoderma edule	Common cockle	*Cardiidae*	19 chr (0.8GB)	CedBTN1 CedBTN2	Same	*Metzger et al. 2016*
Mya arenaria	Soft shell clam	*Myidae*	17 chr (1.2GB)	MarBTN	Same	*Metzger et al. 2015*
Polititapes aureus	Carpet shell clam	*Veneridae*	Not available	VcoBTN	*Venerupis corrugata*	*Metzger et al. 2016*
Venus verrucosa	Warty venus clam	*Veneridae*	Not available	CgaBTN	*Chamelea gallina*	*Garcia-Souto et al. 2022*
Macoma balthica	Baltic clam	*Tellinidae*	Not available	LbaBTN	Same	*Michnowska et al. 2022*
Mytilus trossulus	Foolish mussel	*Mytilidae*	Not available	MtrBTN1 MtrBTN2	Same	*Metzger et al. 2016*
Mytilus edulis	Blue mussel	*Mytilidae*	14 chr (1.8GB)	MtrBTN2	*Mytilus trossulus*	*Yonemitsu et al. 2019*
Mytilus chilensis	Chilean mussel	*Mytilidae*	Not available	MtrBTN2	*Mytilus trossulus*	*Yonemitsu et al. 2019*
Mytilus galloprovincialis	Mediterranean mussel	*Mytilidae*	14 chr (1.7GB)	MtrBTN2	*Mytilus trossulus*	*Hammel et al. 2021*

C2.1.1.3 - Breaking the barriers of cancer transmission

The extent to which BTN demands or promotes an immunosuppressive host condition during its transmission cycle is currently unknown. Remarkably, both CTVT and DFTD have developed molecular strategies in order to surpass the recognition by the host immune system (for example, CTVT through somatic mutation, DFTD through epigenetic downregulation) (Ostrander, Davis, and Ostrander 2016). It must be noted that bivalves display less complex and specialized immune systems compared with vertebrates, and their adaptive responses against infection are much less efficient than the ones displayed by mammals. However, potential mechanisms of lectin-mediated allogeneic immune recognition have been already identified in marine bivalves (Allam and Raftos 2015), which raises the possibility that active molecular mechanisms of immune evasion might have evolved in BTN in order to overcome the bivalve host defence.

Remarkably, some forms of marine BTN transmissible cancer can spread not only between geographically distant populations, but also between different bivalve species (even across different genus). This is the case for a transmissible tumour derived from the pullet carpet shell,

Venerupis corrugata, whose cancer cells seem to be adapted to recurrently infect a new host species (*Polititapes aureus*), but not their founder organism (Metzger et al. 2016). Other examples of marine interspecific transmission have been recently reported, such as the one described in clams between warty venus (*Venus verrucosa*) and striped venus (*Chamelea gallina*) (Garcia-Souto et al. 2022), or in mussels between the blue mussel (*Mytilus trossulus*) and other mussel species (*Mytilus chilensis* and *Mytilus edulis*) (Yonemitsu et al. 2019). Altogether, these findings suggest that cancer transmission might be a more common process than previously considered -at least in the marine environment-.

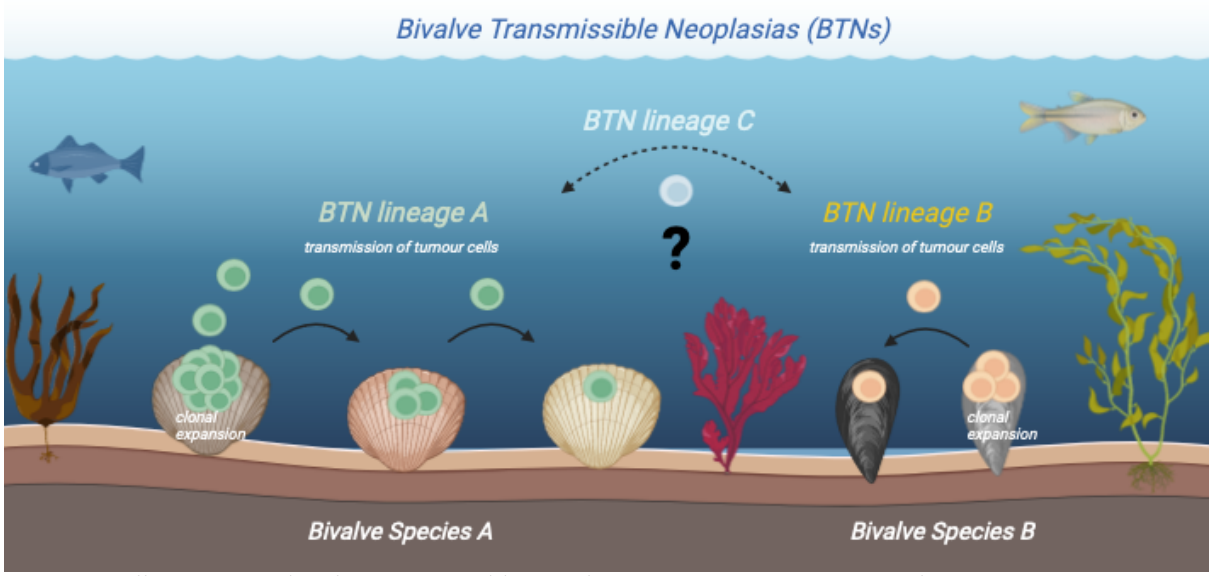


Figure 22: Illustration of bivalve transmissible neoplasias (BTNs) transmission cycle.
In these tumours (e.g. BTN lineage A and BTN lineage B), neoplastic haemocytes experiment rounds of clonal expansion in their host tissues before being released into the seawater and finally reaching and infecting another bivalve in the population, where a new clonal expansion is seeded. Sometimes, several BTN lineages can coexist in the same species, and transmission between different bivalve species (e.g. BTN lineage C) has also been reported. Sources: Metzger et al. 2016; Murchison 2016. Created with BioRender.com.

C2.1.2 - *Cerastoderma edule* bivalve transmissible neoplasia (CedBTN)

C2.1.2.1 - The common cockle *Cerastoderma edule*

One of the most remarkable bivalves threatened by BTN is the common cockle or edible cockle, *Cerastoderma edule (*family *Cardiidae,* class *Bivalvia).* This marine cockle is distributed across the Atlantic coast of Europe and north-west Africa. It is typically found in the intertidal zones which constitute its primary natural habitat, where it can comprise 60% of the benthic biomass (Peteiro et al. 2018). Adult common cockles, which can live up to 10 years, usually measure between 3 to 5 centimetres, and remain buried in the sediment, where they filter the seawater in search for plankton and organic detritus. *Cerastoderma edule* is characterized by a defensive oval shell, inhalant and exhalant siphons, mantle and gills, which serve both nutritional and respiratory functions, and a foot muscle that is used for movement in the seabed. As other bivalves, the common cockle has an open circulatory system; nutrients, oxygen and waste

USC DE COMPOSTELA

products are transported across the organism through the haemolymph. *Cerastoderma edule* is a dioic species with fast sexual maturation and high fecundity. Its larval stage is planktonic and typically remains in the water column for a month, which promotes dispersal by sea currents driving gene flow and connectivity between cockle populations (Vera et al. 2022).

C2.1.2.2 – Identification of cancer transmissibility in the common cockle

The common cockle is affected by a neoplastic disease which is characterized by the unrestricted proliferation of tumour cells disseminating through the circulatory system and infiltrating the connective tissue of various organs (María J. Carballal et al. 2015). These tumour cells share characteristic features with hemic neoplasias found in other bivalves, including large size, round or oval shape, high nucleus-to-cytoplasm ratio, prominent nucleolus and a high frequency of mitotic figures.



Figure 23: Common cockles are affected by the Cerastoderma edule BTN.
CedBTN is a transmissible tumour which affects the haemolymph and different organs of these bivalves. Image obtained within the framework of the Scuba Cancers ERC project, corresponding to the haemolymph extraction from a common cockle individual. Credit: Alicia L. Bruzos and Seila Díaz.

Despite the first description of bivalve hemic neoplasia being made in the late-sixties, it was not until 1984 when it was first reported in the common cockle *Cerastoderma edule* (Twomey and Mulcahy 1984). The disease was first detected in Cork Harbour, Ireland. Since then, it has been reported in several European countries, including the Netherlands, France, Portugal, and Spain, where it is associated with significant cockle mortalities. In the Spanish region of Galicia, where *Cerastoderma edule* has significant socioeconomic value, hemic neoplasia was first reported in 1997 and has caused recurrent epidemic outbreaks since then, resulting in substantial losses for commercial fisheries (Vera et al. 2022). Notably, this disease has been recently categorized as a transmissible cancer (Metzger et al. 2016), being referred as *Cerastoderma edule bivalve transmissible neoplasia* (CedBTN).

Similar to other transmissible tumours, the prevalence of CedBTN can be remarkably high and is subject to dynamic variation over time and space (Stammnitz et al. 2023; María J. Carballal

et al. 2015). The overall prevalence was estimated to be high (>20%) in cockles from the European Atlantic coast, ranging from disease-free to heavily affected populations, as found in Ireland, France and Spain (María J. Carballal et al. 2015). In Galician seabeds, hemic neoplasia prevalence was reported to be high (>20%) in several cockle populations from southern and northern estuaries (M. J. Carballal et al. 2001). This pattern might be consistent with founder events followed by rapid tumour transmission across susceptible cockle individuals. As observed in other bivalve neoplasias, the disease prevalence experiences seasonal variations, likely influenced by an interplay of biotic and abiotic factors (e.g. salinity drops, sea temperature changes, reproductive burden, pathogen exposure) (Peteiro et al. 2018; M. J. Carballal et al. 2001; María J. Carballal et al. 2015). To which extent these stressor agents interplay with the cockle physiology in order to promote CedBTN transmission and progression, remains unclear.

The genetic diversity and population structure of the host cockle populations is another relevant factor which might influence the emergence and evolution of CedBTN, as occurs in other transmissible tumours (Baez-Ortega et al. 2019; Stammnitz et al. 2023). Previous work analysing mtDNA *COI* sequences (Krakau et al. 2012) found that European cockle populations are differentiated in two major groups: one conformed by populations from the South-Western Atlantic coast and another conformed by populations from the Northern Atlantic coast. Another study based on microsatellite DNA markers (Martínez et al. 2015) identified a third differentiated group in the British Islands and the southern North Sea, placed in a relevant contact zone between major metapopulations in the proximity of the English Channel. Both studies showed that *Cerastoderma edule* populations have considerable genetic diversity and genetic differentiation, especially at intermediate and northern latitudes. These observations support that at least some of these European cockle populations are ancient and might have survived in northern cryptic refugia during Pleistocene glaciations. Recently, the existence of two major cockle metapopulations (Northern and Southern) was confirmed through an extensive analysis of nuclear DNA SNV markers (Vera et al. 2022). This study also identified diversity patterns compatible with the existence of important glacial refuges in the south of Europe. Additionally, it reported that cockles from the genus *Cerastoderma* probably have limited genetic diversity compared with other bivalve molluscs.

Over the last decades, various aetiological agents have been proposed to explain the origin of bivalve haemic neoplasia: genetic factors, viral infections, environmental contamination and the presence of biotoxins and other stressors (María J. Carballal et al. 2015). Notably, both retroviral infections (Romalde et al. 2007) and the effect of coastal pollution (Twomey and Mulcahy 1988) have been considered to explain the emergence of haemic neoplasia in *Cerastoderma edule*. However, while some contribution from these factors cannot be ruled out, a relevant study by Meztger and colleagues demonstrated that the aetiology of *Cerastoderma edule* hemic neoplasia is indeed the clonal transmission of cancer cells between cockle individuals (Metzger et al. 2016). This conclusion was demonstrated through the analysis of mtDNA and microsatellite DNA haplotypes, which proved that 1) the genotype from tumour cells does not correspond to the genotype of the host individual, as happens in non-transmissible cancer types; and 2) tumour genotypes from different cancer cases clustered together, being indicative of a shared clonal origin. Interestingly, this study also found evidence suggesting that *Cerastoderma edule* haemic neoplasia may have a polyphyletic origin (Figure 24), identified by the presence of two likely independent cancer lineages exhibiting different histological features.

USC DE COMPOSTELA

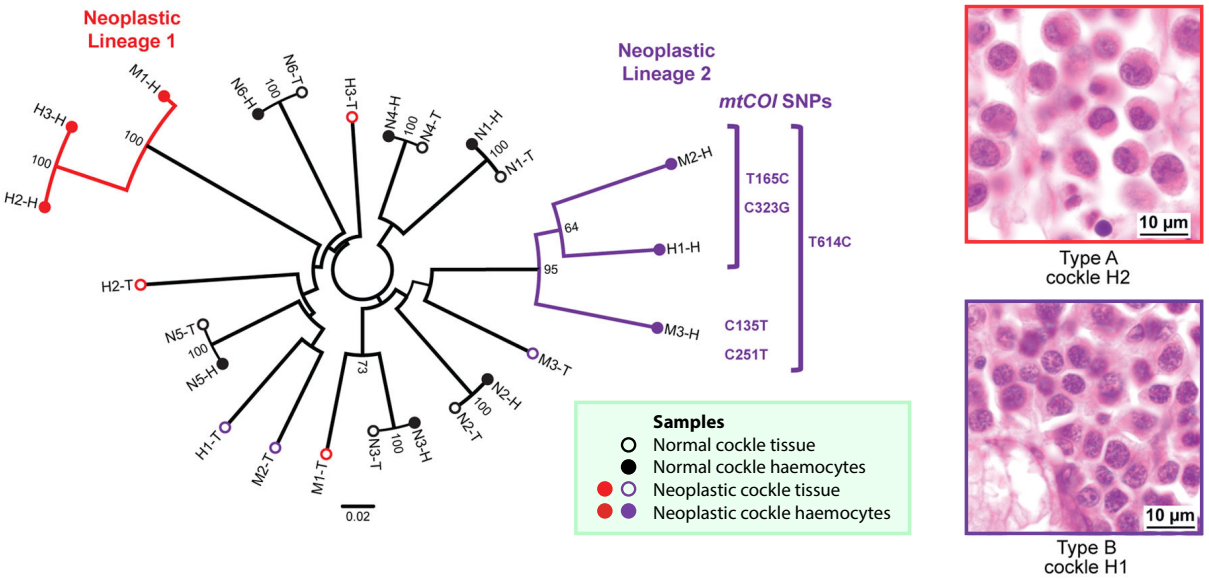


Figure 24: CedBTN NJ phylogenetic tree estimated from mtDNA microsatellite markers.
The analysis from Metzger et al. supported a polyphyletic origin of CedBTN (neoplastic lineage 1 in red, neoplastic lineage 2 in purple). Specific mtDNA *COI* SNPs from neoplastic lineage 2 are displayed on the right. Bootstrap values over 50 are shown. Neoplastic lineages 1 and 2 were shown to be significantly differentiated at the histological level (type A and type B neoplasia, respectively). Figure adapted from Metzger et al. 2016 with permission from Springer Nature.

C2.1.3 - Genetic characterisation of bivalve transmissible neoplasia

C2.1.3.1 - Molecular alterations

The identification of the genetic molecular alterations driving the evolution of bivalve transmissible neoplasias remain to be achieved, despite valuable research efforts have been made in the last two decades. The investigation of the molecular basis of BTN has generally focused on the genetic profiling of p53-family proteins (C. W. Walker et al. 2011), coupled with analyses of their expression patterns at both proteomic and transcriptomic levels.

Among the affected species, the soft-shell clam *Mya arenaria* has become a relevant model. Seminal investigation conducted in this species led to the discovery of a potentially oncogenic mechanism: p53 proteins are sequestered in the cytoplasm of neoplastic haemocytes by the action of mortalin-like proteins (Hart et al. 2023; C. Walker, Böttger, and Low 2006). This molecular process results in the functional impairment of wild-type p53, as observed in several human cancer types (e.g. undifferentiated neuroblastoma, retinoblastoma, colorectal and hepatocellular carcinomas, and glioblastoma) (Shankaranarayana et al. 2023). Recent research in MarBTN found further evidence supporting overexpression of heat shock proteins supporting this model of p53 dysregulation in marine transmissible cancers (Hart et al. 2023).

In mussels belonging to the genus *Mytilus*, researchers have proposed an oncogenic role for the truncated p53-family isoform ΔNp63/74 (Muttray, Schulte, and Baldwin 2008, 53). Additionally, an observed up-regulation of an MDM2-like protein suggests its potential function as a negative regulator of the p53-family pathway in bivalves (Muttray et al. 2010).

Recent gene expression analysis in MtrBTN identified genes associated with the malignant state, including *CYP11A1, THBS1, SELP, ALOX5, BMP2* and *ADAMTS1*. Notably, some of these genes may play a role in the metastatic process and the immune system response. Disruption of universal oncogenic routes was also reported, including the alteration of Hippo, Notch, Wnt, Myc and PI3K pathways. Additionally, missense SNV mutations were identified in the *AKT*, *TP53*, *MDM2*, and *MYC* genes, although the impact of these variants remains unclear (Burioli et al. 2023).

In the common cockle *Cerastoderma edule*, individuals with moderate and advanced neoplasia are characterized by the elevated expression of a 53 kDa mutant p53 protein (Díaz et al. 2010). This condition was not observed in disease-free cockles. Potential upregulation of heat shock proteins (Hsp70 and Hsp90) was also reported. Concurrently, there is a high transcriptional expression of the *RAS* proto-oncogene, although confined to certain disease stages (Ruiz et al. 2013). Over the last years, despite the development of sequencing technologies and the discovery of the transmissible nature of CedBTN, no systematic profiling of cancer-associated gene variants was performed in cockles affected by this disease.

Collectively, these studies explore some molecular alterations associated with bivalve transmissible neoplasia. Some of these alterations, such as p53 dysregulation, appear to be convergent across different bivalve species and potentially conserved across bilaterians (Burioli et al. 2023). These pathways are frequently implicated in human cancers, suggesting shared mechanisms of tumorigenesis. Nevertheless, these studies also highlight the need for further research to elucidate specific molecular drivers involved in BTN evolution, especially those involved in the transition toward a somatic transmissible cell lineage.

C2.1.3.2 - Chromosomal alterations

Bivalve species generally present stable karyotypes, usually comprising diploid organisms with an haploid number of chromosomes ranging from 10 to 23, and DNA contents ranging from 0.65 to 5.4 picograms (Saavedra and Bachère 2006). Remarkably, the genome of bivalve transmissible neoplasia seems to be more variable in terms of chromosome structure and genetic content. Chromosomal alterations and ploidy changes related to haemic neoplasia have been described in several bivalve species, and are considered to be a characteristic feature of these transmissible tumours (María J. Carballal et al. 2015). However, the specific molecular mechanisms driving their genetic instability patterns remain unknown.

In *Mya arenaria,* flow cytometry and karyotype analysis showed that neoplastic cells have multiple large-scale genomic alterations (Reno, House, and Illingworth 1994). Tumour cells presented systematically higher DNA contents than normal clam cells, and chromosomal numbers ranged from 44 to 80 (mode=69) almost duplicating the characteristic range of the host species (26 - 39, mode =34). Tumour cells presented acrocentric and telocentric chromosome configurations which were not detected in normal cell karyotypes. As expected, neoplastic cells presented higher mitotic indexes than normal clam cells. These findings were confirmed and extended by a recent analysis by Hart et al. (2023) analysing whole-genome sequences of MartBTN (Hart et al. 2023). This study identified widespread copy-number changes and structural rearrangements in MartBTN. Tumour genomes were predominantly tetraploid (4N), but a diverse range of copy number states was reported. Tumour chromosomes

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

were impacted by multiple structural variants, including deletions, duplications, inversions, TE insertions and telomere length alterations.

In mussels from the genus *Mytilus*, a wide range of aneuploidy has been described as characteristic from tumour cells. Lowe and Moore (1978) conducted genetic analyses on *Mytilus edulis* in southern Britain, revealing two distinct neoplastic cell types (A and B) with different DNA content (Lowe and Moore 1978). Similar results were obtained by Elston et al. (1990), who reported the sequential development of bivalve neoplasia with high ploidy values in mussels from the genus *Mytilus*. In this study, flow cytometry analyses allowed the detection of two populations of tumour cells, one with pentaploid (5n) and decaploid cells (10n), and other less abundant with tetraploid (4n) and octoploid (8n) cells. González-Tizón et al. (2000) performed karyotype analysis of *Mytilus trossulus* reporting chromosomal alterations associated with neoplasia (González-Tizón et al. 2000). Interestingly, the observed chromosome number in tumour cells ranged from 48 to 92, while the normal diploid state was 28, as found for other mussels. A more recent study by Vassilenko et al. (2014) in *Mytilus trossulus* showed that late-stage mussels affected by neoplasia displayed a wider spectrum of aneuploidy peaks ranging from 1.4n to 5.5n (Vassilenko and Baldwin 2014). Collectively, these studies evidence a complex relationship between aneuploidy and neoplasia progression in mussels from the genus *Mytilus* (María J. Carballal et al. 2015).

In the common cockle *Cerastoderma edule*, a wide spectrum of ploidy variation has also been reported, leading to diverse karyotypes marked by the abundance of small chromosomes (María J. Carballal et al. 2015). Le Grand et al. (2010) studied aneuploid patterns and changes in cell ploidy relative to disease development in cockles from Arcachon Bay, France (Le Grand et al. 2010). Flow cytometry revealed that neoplastic cells presented a broad continuum of ploidy distribution ranging from 1.6n to 9.6n. Ploidy levels increased according to disease progression, with highly diseased cockles presenting more alterations. Interestingly, coexistence of several aneuploid peaks within the same individual was reported in 40% of the cockles analysed. Similar results were obtained by Díaz et al. (2013) studying cockles from Galician estuaries (Díaz et al. 2013). In this study, tumour aneuploid peaks ranged from 1.3 to 8.9, with higher ploidies found in advanced neoplasias. Karyotyping revealed that chromosome number was extremely variable in tumour cells, with chromosome numbers ranging from 41 to 145. Interestingly, metaphase inspection revealed that increments in the chromosome number might be due to the occurrence of micro-chromosomes. Later, Matías et al. (2014) reported wider ranges of chromosome number in cockle tumour cells (57 to 317), and found that advanced tumour stages were associated with an increase of the percentage of double minute-like chromosomes (Matias et al. 2014). Altogether, these findings suggest that *Cerastoderma edule* haemic neoplasia might present one of the most complex and rearranged cancer genomes among bivalve transmissible cancers.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

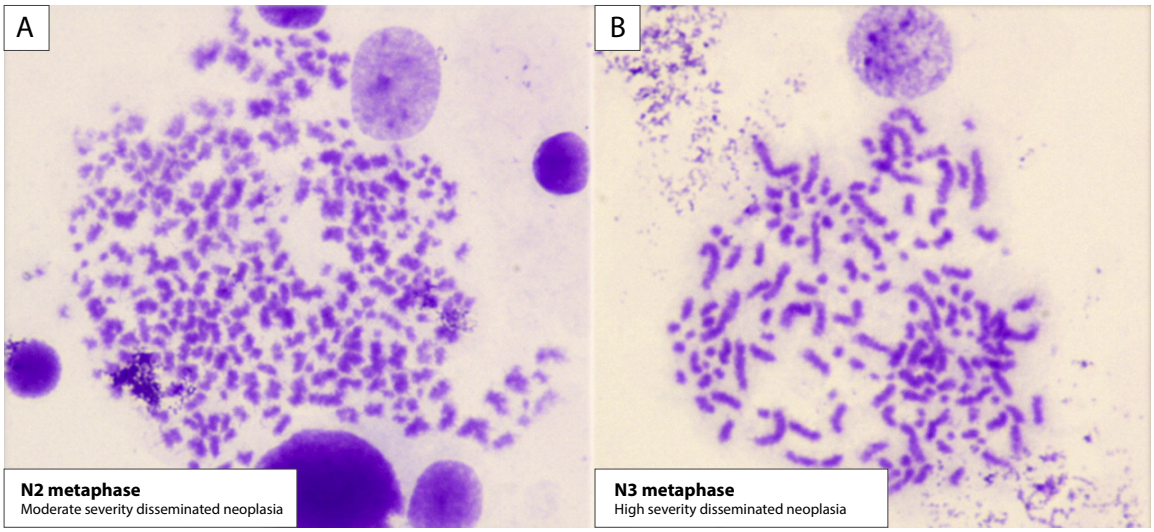


Figure 25: Neoplastic metaphases in the cockle Cerastoderma edule.
Examples of neoplastic metaphases in the common cockle. A) Metaphase of a cockle diagnosed with moderate severity disseminated neoplasia (N2), with 317 chromosomes; B) Metaphase of a cockle diagnosed with high severity disseminated neoplasia (N3), with 169 chromosomes. Figure adapted from: Matias et al. 2014.

C2.1.3.3 - Bivalve genomics in the Era of next-generation sequencing technologies

Our knowledge of the structure and gene content of bivalve genomes is less extensive than for other invertebrate groups, and certainly less than for vertebrates. With the advent of next-generation sequencing technologies, large-scale genomic profiling of bivalve genomes has finally become feasible. For decades, research on bivalve genetics and bivalve cancer was constrained to techniques like flow cytometry and karyotype analysis, which, while providing valuable insights into the genetic features of these organisms, offered a limited scope. Additionally, our capacity to detect and characterize genetic alterations was primarily based on the analysis of short mitochondrial DNA sequences (e.g. *COI* haplotypes) and specific nuclear DNA sequences, such as satellites, microsatellites and specific marker genes (e.g. histone H3, ribosomal DNA genes). These earlier, classical approaches contributed to our understanding of the genetic characteristics of bivalves (for example, doubly uniparental inheritance patterns), and the identification of potential disease markers, also in relation to the origin of BTN. However, the emergence of advanced genomic technologies such as whole-genome sequencing has transformed our ability to investigate bivalve genomes comprehensively. This technological leap now allows for systematic, genome-wide exploration of specific mutations, gene expression profiles, and the underlying molecular mechanisms driving transmissible cancers in these aquatic organisms.

In the past decade, more than 40 bivalve species have been sequenced to chromosome level according to the NCBI genomes database (Table 6), including at least four species affected by hemic neoplasia. This represents an unprecedent opportunity to study the genetic basis and evolutionary dynamics of bivalve transmissible cancer. However, by the start of this research, there was no reference genome available for the common cockle *Cerastoderma edule*, challenging the study of the genetic basis of CedBTN. Previous studies have shown that common cockles have a diploid chromosome number of 38, with a variable occurrence of supernumerary chromosomes, and a karyotype conformed by 12 submetacentric, 4 subtelocentric and 3 telocentric chromosome pairs (Insua and Thiriot-Quiévreux 1992). During the development of this PhD thesis, a reference genome sequence consistent with this haploid

karyotype was assembled and annotated by our research group within the framework of the ERC Scuba Cancers project. This work yielded a key resource to systematically identify and characterize genetic alterations in the common cockle *Cerastoderma edule*. These alterations include short variants (SNVs, MNVs and Indels), SVs, and CNVs, among other types of genetic alterations expected to be somatically acquired during the evolution of CedBTN.

Table 6: Bivalve species with published reference genomes.

Assembly accession	Bivalve genome	Size	N	Assembly level	Release
GCA_032173915.1	*Ostrea edulis*	0,9 GB	10	Complete	29/9/23
GCA_963383655.1	*Bathymodiolus septemdierum*	1,4 GB	15	Chromosome	26/8/23
GCA_963210305.1	*Tridacna derasa*	1,1 GB	18	Chromosome	19/8/23
GCA_025765685.3	*Crassostrea gigas*	0,6 GB	10	Chromosome	7/7/23
GCA_025765675.3	*Crassostrea angulata*	0,6 GB	10	Chromosome	7/7/23
GCA_029721355.1	*Tegillarca granosa*	0,9 GB	19	Complete	18/4/23
GCA_029582155.1	*Panopea generosa*	1,5 GB	19	Chromosome	6/4/23
GCA_029237695.1	*Conchocele bisecta*	1,9 GB	17	Chromosome	20/3/23
GCA_028554795.1	*Sinohyriopsis cumingii*	3,4 GB	19	Chromosome	9/2/23
GCA_028015205.1	*Mytilisepta virgata*	1,0 GB	14	Chromosome	31/1/23
GCA_948146395.1	*Fragum whitleyi*	1,2 GB	18	Chromosome	18/1/23
GCA_027627225.1	*Congeria kusceri*	1,6 GB	17	Complete	13/1/23
GCA_947846245.1	*Cerastoderma edule* ♦	0,8 GB	19	Chromosome	5/1/23
GCA_028142955.1	*Pinctada fucata*	0,9 GB	14	Chromosome	19/12/22
GCA_947623455.1	*Mimachlamys varia*	1,0 GB	19	Chromosome	19/12/22
GCA_028253585.1	*Pinctada fucata*	0,9 GB	14	Chromosome	19/12/22
GCF_026914265.1	*Mya arenaria* ♦	1,2 GB	17	Chromosome	15/12/22
GCA_947247005.1	*Spisula solida*	0,9 GB	19	Chromosome	30/10/22
GCA_922989275.2	*Gari tellinella*	1,6 GB	19	Chromosome	25/10/22
GCA_025267735.1	*Mactra quadrangularis*	1,0 GB	19	Chromosome	25/9/22
GCA_025277285.1	*Mytilus galloprovincialis* ♦	1,7 GB	14	Chromosome	23/9/22
GCA_025276775.1	*Mytilus edulis* ♦	1,8 GB	14	Chromosome	23/9/22
GCA_945859785.2	*Tridacna gigas*	1,2 GB	17	Chromosome	22/9/22
GCA_946902895.1	*Fragum fragum*	1,2 GB	19	Chromosome	22/9/22
GCA_946811185.1	*Hippopus hippopus*	1,1 GB	19	Chromosome	15/9/22
GCA_944474755.1	*Limnoperna fortunei*	1,3 GB	15	Chromosome	13/7/22
GCA_943736015.1	*Tridacna crocea*	1,0 GB	18	Chromosome	24/6/22
GCF_021730395.1	*Mercenaria mercenaria*	1,9 GB	19	Chromosome	1/2/22
GCA_021292105.1	*Anadara kagoshimensis*	1,1 GB	19	Chromosome	28/12/21
GCA_021229015.1	*Solen grandis*	1,3 GB	19	Chromosome	19/12/21
GCA_020567875.1	*Crassostrea ariakensis*	0,6 GB	10	Chromosome	4/11/21
GCF_020536995.1	*Dreissena polymorpha*	1,8 GB	16	Chromosome	21/10/21
GCA_020458035.1	*Crassostrea ariakensis*	0,7 GB	10	Chromosome	15/10/21
GCA_017311375.1	*Mytilus coruscus*	1,6 GB	14	Chromosome	10/3/21
GCA_015776775.1	*Crassostrea hongkongensis*	0,6 GB	10	Chromosome	8/12/20
GCA_014843695.1	*Archivesica marissinica*	1,5 GB	19	Chromosome	7/10/20
GCA_009762815.1	*Sinonovacula constricta*	1,3 GB	19	Chromosome	18/12/19
GCF_902652985.1	*Pecten maximus*	0,9 GB	19	Chromosome	26/11/19
GCA_009026015.1	*Ruditapes philippinarum*	1,1 GB	19	Chromosome	15/10/19
GCF_002022765.2	*Crassostrea virginica*	0,7 GB	10	Complete	1/9/17
GCA_002216045.1	*Pinctada imbricata*	1,0 GB	14	Chromosome	20/7/17

This table collects bivalve species with either complete or chromosome-level bivalve genome assemblies. Bivalve species known to be affected by bivalve transmissible neoplasia are denoted with a diamond. Source: https://www.ncbi.nlm.nih.gov/datasets/genome/ (consulted by 20/10/2023).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.2 - RESULTS AND DISCUSSION

C2.2.1 - Impact of structural variation in the genome of CedBTN

To study the genome-wide impact of somatic structural variation in the CedBTN, we applied a combination of algorithms (Manta, Lumpy and Delly) to call structural variants in our golden set of 10 high-purity tumours. After removing calls with low support and discriminating germline polymorphisms through variant genotyping in a diverse pool of 455 non-neoplastic cockle genomes, we identified a conservative set of structural variants acquired along CedBTN evolution. Our screening identified a total of 18,272 structural variants which likely represented somatic alterations (Figure 28**)**. On average, CedBTN tumour genomes presented 5,654 somatic structural variants (4,477 deletions, 558 duplications, 102 inversions and 518 breakends), with a standard deviation of 730 events, and a range of 4,923 to 6,861 alterations. Most of the detected structural variants (12,016/18,272; 65%) were shared by several CedBTN tumours, suggesting they corresponded to relatively antique somatic mutations, acquired by older states of the transmissible cancer genome. However, only 74 structural variants were totally shared across all the CedBTN tumours of our golden set, some of which might still correspond to ancient germline variants. A total of 6,256/18,272 (35%) structural variants were found to be private from individual tumour genomes. This pattern is consistent with a high degree of ongoing genomic instability in CedBTN, despite we cannot rule out that a residual fraction of these variants still correspond to rare germline polymorphisms derived from cockle host genomes.

We explored in further detail this set of somatic alterations to assess the prevalence, features and distribution of different categories of structural variation in our CedBTN golden set:

Deletions represented the most common type of somatic structural variation in the genome of CedBTN (79.8%, 14,589 out of 18,272 events, Figure 28). Deletions affected every chromosome of the CedBTN genome, with a higher density observed on chromosome 9 (24.7 deletions / MB), chromosome 3 (23.6 deletions / MB) and chromosome 10 (23.5 deletions / MB). Deletions generally represented the smaller type of structural variant, with a median size of 301 bp across detected events. However, multiple instances of large-scale deletions were also identified, including 463 deletions greater than 100 KB and 55 deletions greater than 1 MB. Notably, a total of 6,402 deletions (43.8%) were found impacting annotated *Cerastoderma edule* genes, while 8,187 deletions (56.1%) were entirely intergenic. Overall, deletions caused the removal of 337.9 MB of genetic material in CedBTN, representing a major source of somatic structural variation at the intrachromosomal level.

Duplications constituted the second most common type of somatic structural variation in the genome of CedBTN (9.2%, 1,687 out of 18,272; Figure 28). These duplications affected all chromosomes but showed higher density on chromosome 10 (3.1 duplications / MB), chromosome 2 (3.0 duplications / MB) and chromosome 9 (2.8 duplications / MB). Duplications exhibited a median size of 5,171 bp, with several notable instances of large-scale duplications identified, including 196 duplications that exceeded 100 KB and 57 that surpassed 1 MB. Of the total duplications detected, 987 (58.5%) impacted annotated *Cerastoderma edule* genes, while 700 (41.5%) were intergenic. Collectively, these duplications resulted in the

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

amplification of 286.9 MB of genetic material, highlighting their significance as a major source of somatic structural variation at the intrachromosomal level.

Breakends, here associated with translocations and other interchromosomal rearrangements, (9.1%) represented the third most common category of structural variation in the genome of CedBTN (Figure 28). With 1,655 events, they impacted on every chromosome of the CedBTN genome, with higher density of events on chromosome 9 (2.5 breakends / MB), chromosome 16 (2.5 breakends / MB) and chromosome 13 (2.4 breakends / MB). The higher frequency of interchromosomal rearrangements was found between chromosomes 3 and 5 (29 breakends), between chromosomes 1 and 3 (28 breakends) and between chromosomes 3 and 4 (28 breakends). A total of 679 breakends (41.0%) were found impacting annotated *Cerastoderma edule* genes, while 976 breakends (60.0%) were intergenic. Collectively, breakends supported the formation of multiple derivative chromosome configurations, indicating a substantial level of interchromosomal reorganization in the CedBTN genome.

Inversions (1.9%) represented the less common type of structural variation identified in the genome of CedBTN (Figure 28). With 341 events, inversions still impacted on every chromosome of the CedBTN genome, with higher density of events on chromosome 12 (0.7 inversions / MB), chromosome 3 (0.6 inversions / MB) and chromosome 7 (0.6 inversions / MB). Inversions generally represented the larger type of structural variants, with a median size of 6647 bp. Notably, we identified 91 inversions greater than 100 KB, and 38 inversions greater than 1 MB. A total of 217 inversions (63.6%) were found impacting annotated *Cerastoderma edule* genes, while 124 events (36.4%) were intergenic. Collectively, inversions led to the reorganization of 161.0 MB of genetic material in CedBTN, still representing a relevant source of somatic structural variation at intrachromosomal level.

Overall, these results are indicative of a significant degree of somatic structural reorganization in the genome of CedBTN, consistent with aneuploidy states previously described for this neoplasia (Le Grand et al. 2010; Ruiz et al. 2013; Matias et al. 2014), and supported by cytological evidence generated within the framework of the ERC Scuba Cancers project (Supp. Figure 7). Our genome-wide screening demonstrated the widespread impact of structural variation across multiple chromosomes. According to our results, at least of 786 MB of nuclear DNA sequence were impacted by intrachromosomal genomic rearrangements caused by somatic deletions, duplications and inversions, being chromosome 3 (1,512 events, 9.1%), chromosome 2 (1,472 events, 8.9%) and chromosome 4 (1,202 events, 7.2%) the genomic regions showing higher incidence.

These results can be compared with recent findings obtained for *Mya arenaria* BTN (Hart et al. 2023). In this work, significant levels of structural variation are also reported for the transmissible cancer MarBTN after whole-genome analyses. Deletions also represented the most frequent type of structural variant in the MarBTN tumour genome (48,978 detected SVs per cancer genome on average), followed by duplications (2,861 detected SVs per cancer genome on average), breakends (2,071 detected SVs per cancer genome on average), and finally, inversions (1,569 detected SVs per cancer genome on average). Further work is needed to determine if shared patterns of structural genetic variation can be found along *Cerastoderma edule* and *Mya arenaria* germline and somatic BTN evolution; this research may require examining polymorphisms in host species. Interestingly, a similar size distribution was found for somatic duplications both in CedBTN and MarBTN lineages, which might suggest some of

USC DE COMPOSTELA

them could have been originated by similar molecular processes (e.g. alterations in TP53 and BRCA1 pathways) (Yi and Ju 2018).

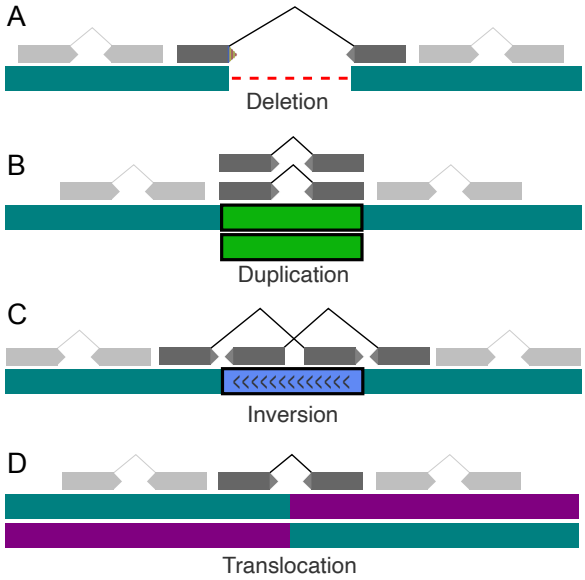


Figure 26: Structural variation detection from short-read paired end sequencing data.
Detection arises from specific patterns of clusters of discordant reads (reads that do not align to the reference genome as expected), clusters of breakpoint-supporting clipped reads (reads partially aligned to the reference) and coverage variation associated with each type of genomic rearrangement. A) Deletions are characterized by a significant drop of coverage and by some discordant read pairs aligned at longer distances than expected; B) Duplications are characterized by a significant increase of coverage and by some discordant reads aligned at longer distance than expected C) Inversion are characterized by some discordant read pairs aligned in abnormal orientation and spacing; D) Translocations involve the presence of some discordant read pairs aligning to different chromosomes. Schematic figure adapted from Cosenza et al. 2022.

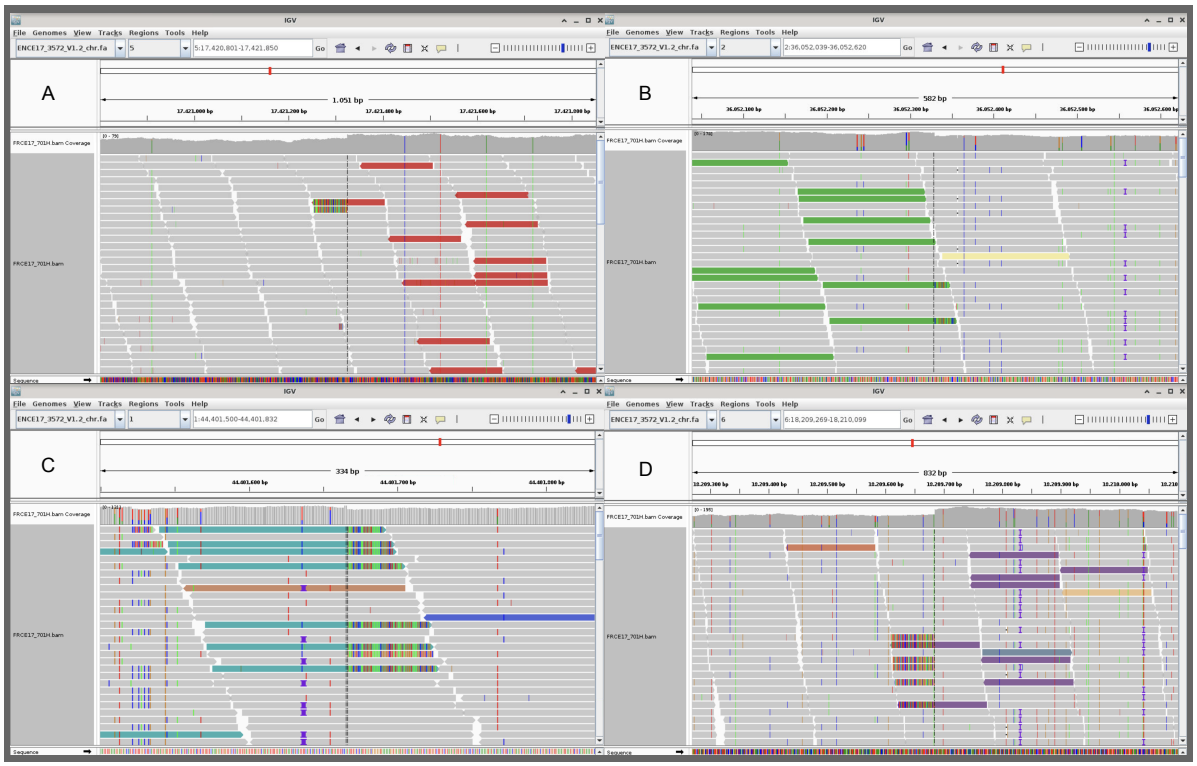


Figure 27: IGV visualization of CedBTN somatic structural variant breakpoints.
Variant support arises from specific patterns of I) clusters of discordant reads; II) clusters of clipped reads; and III) coverage variation. A) Deletion found in chromosome 5; B) Duplication found in chromosome 2; C) Inversion found in chromosome 1; D) BND indicative of a rearrangement between chromosome 6 and chromosome 19. Tumour: FRCE17_701H.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

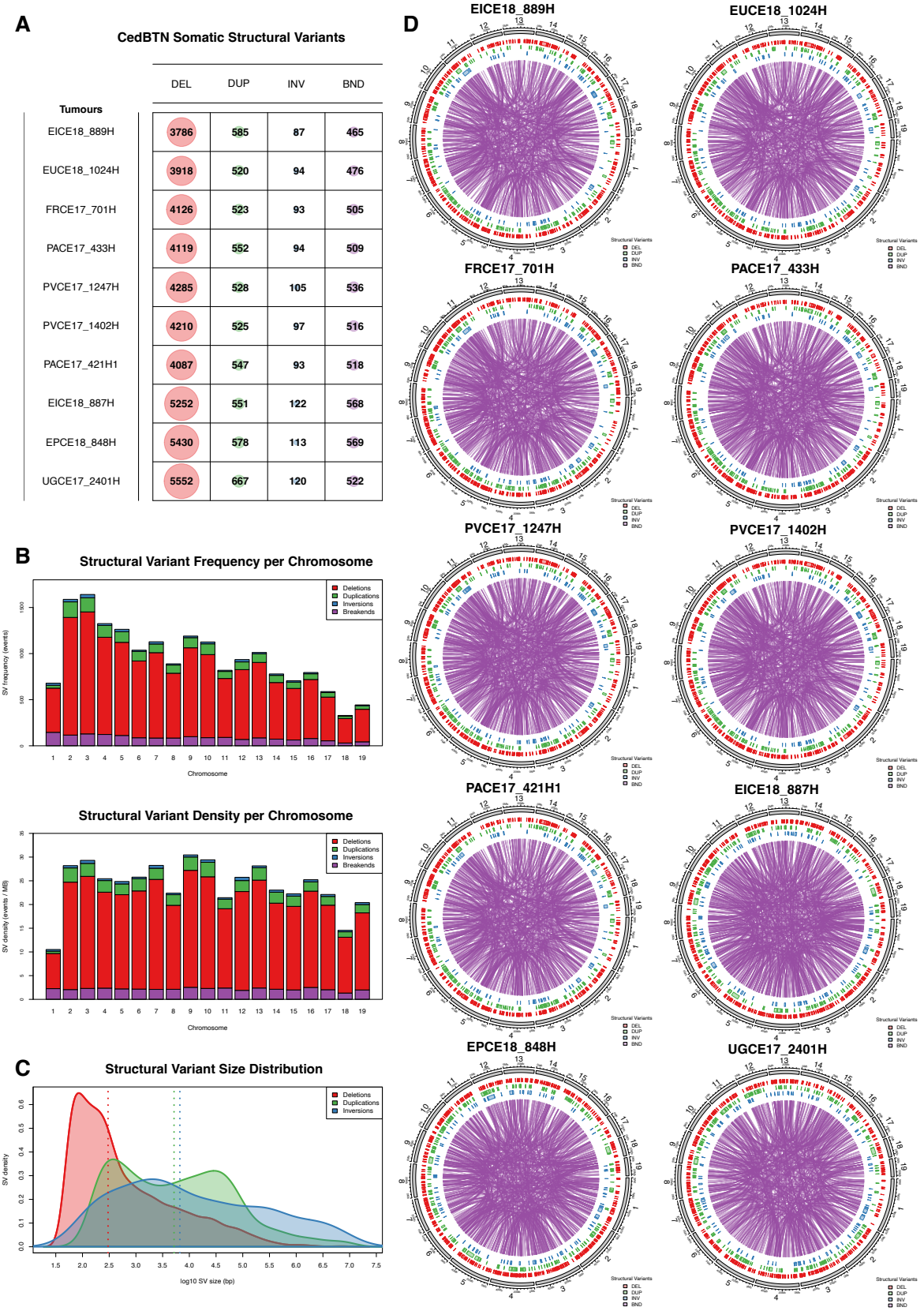


Figure 28: Landscape of somatic structural variants across the CedBTN genomes.
A) Summary of somatic structural variant events found in the golden set of 10 CedBTN genomes. The number of deletions (DEL), duplications (DUP), inversions (INV) and breakends (BND) identified on each tumour (rows) is indicated, with circle size proportional to the number of events. B) Structural variant frequency (top; number of

events) and density (bottom; number of events / MB) as identified for each *Cerastoderma edule* chromosome (1-19). C) Structural variant size distribution for deletions, duplications and inversions. Median values for each type of structural variation are indicated with a coloured dashed line. D) Circos plots showing the genomic distribution for deletions (red), duplications (green), inversions (blue) and breakend (purple links) events on each CedBTN tumour. *Cerastoderma edule* chromosomes (1-19) are represented as external bars in the ideogram.

Table 7: Size distribution of detected somatic structural variants in the CedBTN genomes.

Structural variant	Median	Mean	Minimum	Maximum	Standard deviation
Deletions	0.30 kb	23.2 kb	0.05 kb	9,790 kb	256.7 kb
Duplications	0.52 kb	170.0 kb	0.09 kb	9,970 kb	881.2 kb
Inversions	0.66 kb	472.3 kb	0.07 kb	9,779 kb	1,393.5 kb

Abbreviations: kb=kilobase pairs.

C2.2.2 – Clonal structure and phylogenetic composition of CedBTN

Our genome-wide screening of nuclear variants enables to track the nuclear DNA evolution of CedBTN lineages and explore its underlying clonal structure. Previous work by Metzger et al. based on classical markers showed the presence of two different lineages of the neoplasia, differentiated at histological level (Metzger et al. 2016). Similar results were recently found on other bivalve species, such as *Mytilus trossulus* (Skazina et al. 2023), where multiple clones might coexist. However, no systematic analyses were made in *Cerastoderma edule* considering variants profiled at genome scale, and after extensive sampling across the species distribution. To investigate the origins and evolution of CedBTN, we performed genome-wide phylogenetic analyses based on two major types of nuclear DNA genetic variants (SNVs and SVs).

C2.2.2.1 – Phylogenetic inference from SNVs

First, we generated Maximum Likelihood trees based on CedBTN SNVs (Figure 29). After accounting for host DNA contamination and common polymorphisms in *Cerastoderma edule*, we identified a set of 4.3 million SNVs that represented "BTN-specific" variants, and thus could be informative to trace its evolutionary trajectory. These SNVs were detected on the whole-genome sequences of tumours, and comprised the following changes on the reference sequence: 233,270 A>C, 584,151 A>G, 380,421 A>T, 219,563 C>A, 143,321 C>G, 610,906 C>T, 611,243 G>A, 143,057 G>C, 219,662 G>T, 380,288 T>A, 581,819 T>C and 233,012 T>G substitutions (55% transitions, 45% transversions). The most frequent type of these mutations corresponded to cytosine-to-timine substitutions (28.2%). Detailed information regarding SNV calling and underlying mutational signatures can be consulted in the main publication linked to this chapter (Bruzos et al. 2023), being a work performed by Adrián Baez-Ortega.

This phylogenetic tree split the tumours into two divergent lineages, which consistently matched the two histological types of CedBTN identified to date (Metzger et al. 2016). The first lineage (CedBTN1), matching the histological type A neoplasia (characterized by larger, isolated tumours cells), is delimited by a large basal branch supported by 1,081,955 shared SNVs (24.9%). The second lineage (CedBTN2), matching the histological type B neoplasia (characterized by smaller, clustered tumour cells), is delimited by a shorter basal branch supported by 524,139 shared SNVs (12.1%). Internal branches were shorter and were found to

USC DE COMPOSTELA

be associated with a total of 130,423 SNVs in CedBTN1 (3.0%), and 612,575 SNVs in CedBTN2 (14.1%). Notably, CedBTN tumours from the same sampling location clustered together on the phylogenetic tree (PACE17_433H - PACE17_421H and PVCE17_1247H - PVCE17_1402H monophyletic clades), but each major clone comprised tumours from a broad geographic distribution. A large basal tree branch associated with 1,541,008 SNVs (35.5%) split all tumours from the reference genome, suggesting significant divergence between tumours and non-neoplastic cockle genomes.

The analysis of dN/dS ratios performed by Adrian Baez-Ortega showed that lineage-specific basal branches were still enriched in ancestral germline variation (dN/dS values of 0.58, indicative of variants preserved under negative selection). In addition, the basal tree branch displayed a stronger signal (dN/dS value of 0.24), which is compatible with large-scale germline purifying selection expected for most protein coding sequences. Conversely, internal tumour branches were found to be enriched in cancer somatic variation (dN/dS values of 0.96 in CedBTN1 and 0.94 in CedBTN2). These values agree with large-scale neutral evolution within the CedBTN cancer genomes. Recent studies in human and transmissible cancer genomes have found that cancer genome evolution is usually driven by large-scale neutral evolution, with only a small fraction of genes and somatic variants undergoing strong negative selective pressure (Martincorena et al. 2017; Baez-Ortega et al. 2019). Altogether, these observations support an independent origin for both cancer lineages, arising from two different individuals belonging to divergent cockle populations.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

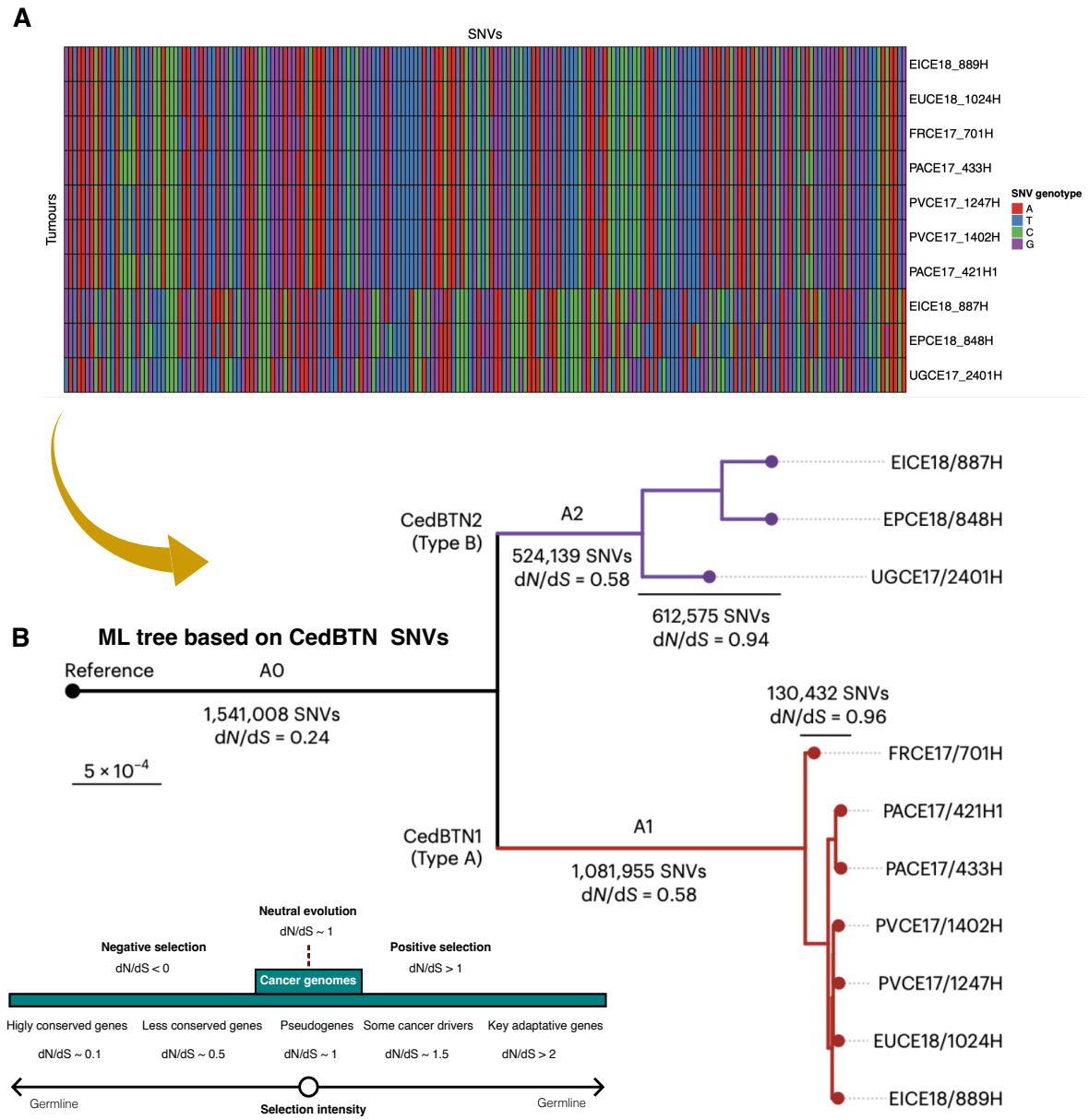


Figure 29: Phylogenetic analysis from CedBTN single nucleotide variants.
A) Alignment of tumour genotypes for 200 randomly selected BTN-Specific SNVs (sampled for illustrative purposes). B) Phylogenetic tree estimated from 4340713 BTN-Specific SNVs in ten golden set samples, showing two monophyletic lineages matching histological cancer subtypes A and B. The number of SNVs and dN/dS ratios associated with each tree branch(es) are indicated. An Illustrative scheme of dN/dS ratio as measurement of selection and typical values found on different genetic sequences under different scenarios is included.

The observation that CedBTN may consist of two primary nuclear DNA clones of independent origin is further supported by a principal component analysis of germline variation (Figure 30). For this analysis, we genotyped 100K SNVs across 10 CedBTN tumours and 100 non-neoplastic cockles from eleven countries across the Atlantic coast. The distribution of the samples across the two principal components placed the tumours as two divergent clusters differentiated from the non-neoplastic sample genomes, which were split into two main groups corresponding to known northern-Europe and southern-Europe cockle populations (Vera et al. 2022). Both tumour clusters matched in composition with the two main clades found in our CedBTN phylogenetic analyses, being represented by the same CedBTN1 and CedBTN2

genomes. These observations, together with the impossibility to remove a significant fraction of cockle germline variation according to the dN/dS ratios, support that CedBTN1 and CedBTN2 are indeed ancient clones which emerged in a different genetic background from the modern *Cerastoderma edule* populations.

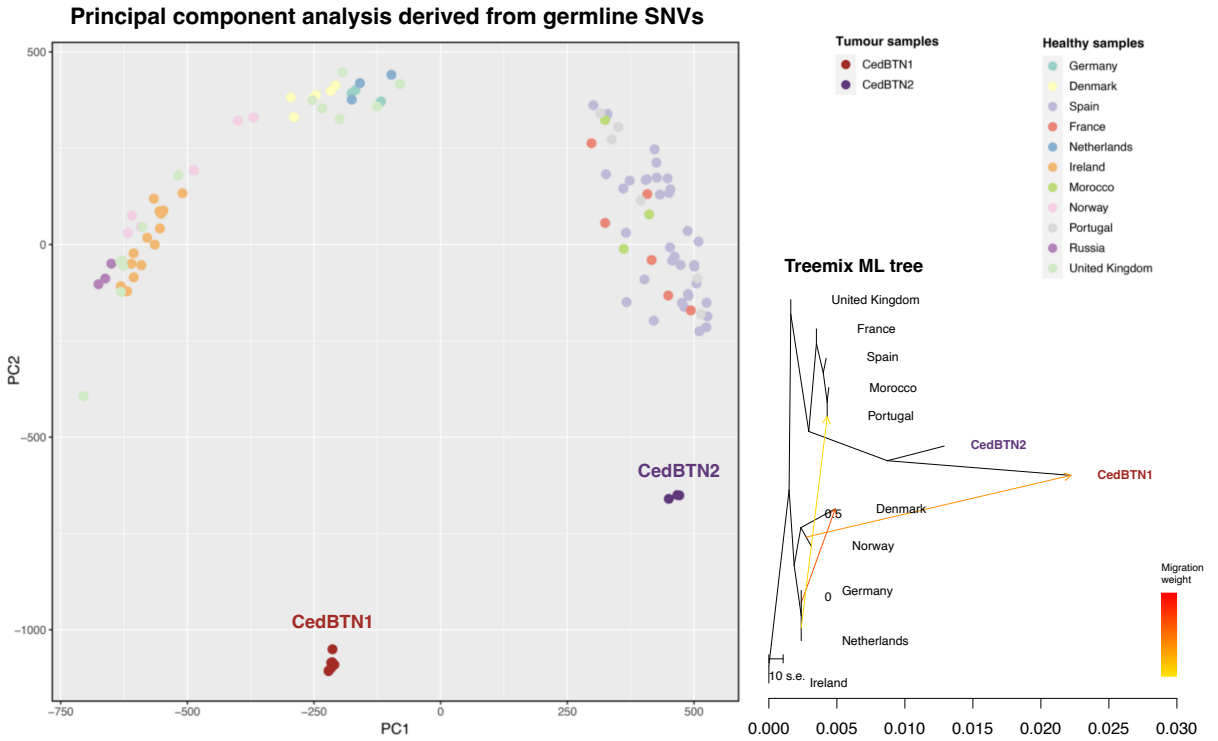


Figure 30: Analysis of germline CedBTN SNVs.
On the right, a principal component analysis of germline polymorphisms in CedBTN and non-neoplastic *Cerastoderma edule* samples. Logistic PCA was performed on a randomly selected subset of 100K germline exonic SNVs, genotyped across 100 non-neoplastic cockles (covering all sampling locations) and 10 golden set tumours. On the left, a Treemix ML analysis based on the covariance of allelic frequencies from a random selection of 30K SNVs on populations of neoplastic and non-neoplastic cockle genomes (10 populations) is shown.

C2.2.2.2 - Phylogenetic inference from SVs

We generated a Maximum Parsimony tree based on CedBTN using structural variants (Figure 31). After accounting for common polymorphisms in *Cerastoderma edule*, we identified 18,272 high-confidence BTN-specific SVs. These SVs were detected on the whole-genome sequences of tumours, and comprised the following changes: 14,589 deletions, 1,687 duplications, 341 inversions and 1,655 breakends (See section '*Impact of Structural Variation in the Genome of CedBTN'*). Tree evaluation under parsimony testing showed that identified variants were associated to moderate homoplasy levels (Consistency index of 0.81 for deletions, 0.91 for duplications, 0.77 for breakends and 0.79 for inversions), being indicative of accuracy in SV genotyping.

The SV-based phylogenetic tree split the tumours into two major divergent lineages, which matched in composition with the ones previously identified through SNV-based analyses. The first lineage (CedBTN1), matching the histological type A neoplasia, is delimited by a large basal branch supported by 2,825 shared SVs (15.5%). On average, our ten CedBNT1 tumours

USC DE SANTIAGO DE COMPOSTELA

presented 5,214 somatic structural variants (4,076 deletions, 540 duplications, 95 inversions and 503 breakends), with a standard deviation of 186 events. The second lineage (CedBTN2), matching the histological type B neoplasia, is delimited by another basal branch supported by 2,582 shared SVs (14.1%). On average, the three CedBTN2 tumours presented 6,681 somatic structural variants (5,411 deletions, 599 duplications, 118 inversions and 553 breakends), with a standard deviation of 185 events. Internal branches were shorter and were found to be associated with a total of 4,091 SVs in CedBTN1 (14.1%), and 8,343 SVs in CedBTN2 (22.4%). As occurred in the case SNV-based analyses, CedBTN tumours from the same sampling location clustered together on the phylogenetic tree (PACE17_433H - PACE17_421H and PVCE17_1247H - PVCE17_1402H monophyletic clades). Notably, the number of structural variants shared by every CedBTN tumour is low (76 events, 0.4%), suggesting there might be limited ancestral germline variation in the dataset represented by structural variation polymorphisms.

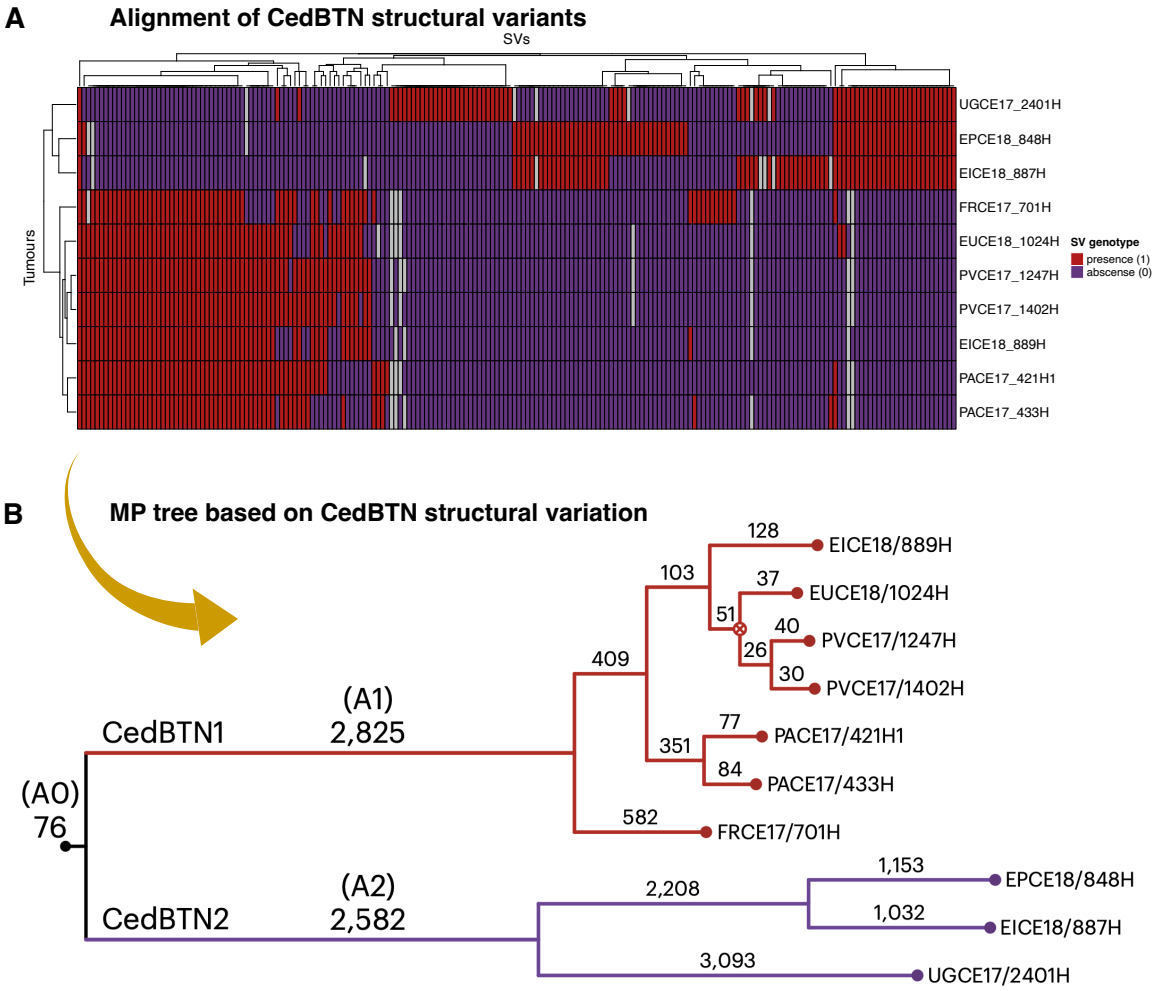


Figure 31: Phylogenetic analysis from CedBTN structural variants.
A) Alignment of tumour binary genotypes (presence vs absence) for 200 randomly selected BTN-Specific SVs (sampled for illustrative purposes). B) MP phylogenetic tree estimated from 18272 BTN-Specific SVs (including deletions, duplications, inversions and breakends) in ten golden set samples, showing correspondence to histological cancer lineages A and B. The number of SVs associated with each phylogenetic tree branch is shown.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.2.2.3 - Two main transmissible cancer lineages propagate in cockles

Altogether, our findings suggest that transmissible cancer in *Cerastoderma edule* is represented by at least two genetically divergent cancer clones, CedBTN1 and CedBTN2. These clones exhibit phenotypic differences at the histological level and have accumulated a significant mutational burden through different molecular mechanisms, including different classes of structural and punctual variation. According to our results, CedBTN1 and CedBTN2 appear to be remarkably ancient clones which emerged in a different genetic background compared with the extant Atlantic *Cerastoderma edule* populations. Analysis of sequencing mapping data showed that the fractions or sequence reads aligning against the *Cerastoderma edule* reference genome in BTN tumours (97 – 98%) are comparable to those for 462 non-neoplastic cockles (interquartile interval, 97 – 98%), and substantially higher than fractions for cockles of the closest known species, *Cerastoderma glaucum* (48 – 60 %, six samples). This is consistent with both lineages having arisen from *Cerastoderma edule* founder individuals genetically distant from extant common cockles. Our results therefore confirm and extend previous findings from Metzger et al. (2016), which initially identified a polyphyletic origin of BTN in *Cerastoderma edule*.

To which extent the likely independent origin of CedBTN1 and CedBTN2 leads to differences in the evolution of this transmissible cancer is a question that remains to be addressed. Motivated by our results, which support the existence of two ancient and likely independent BTN lineages, we developed a PCR-based molecular assay to identify each of these clones based on specific structural variant genotypes (Figure 32, Table 8). This assay might be a valuable resource to identify and discriminate both cancer lineages in a cost-effective manner. Recurrent, independent clonal origin of transmissible cancer have been shown to occur in other transmissible cancers (DFTD), and in other bivalve transmissible neoplasias (MtrBTN), leading to the development of specific molecular phenotypes, which include differences in host ranges and differential mutation rates and patterns of genomic instability (Hammel et al. 2022; Stammnitz et al. 2023).

In *Cerastoderma edule*, the two CedBTN lineages identified are characterized by different histological features in terms of tumour cell size and interaction. However, both lineages seem to share major key molecular features. According to gene expression patterns analysed in the ERC Scuba Cancers framework, the cellular origin of both CedBTN lineages might correspond to haemocyte precursors. Interestingly, both tumour lineages have acquired mitochondrial DNA haplotypes from its hosts (up to six captures in CedBTN1, and three captures in CedBTN2) (López Bruzos 2022). This phenomenon has also been observed in both mammalian (CTVT) and bivalve (MtrBTN) transmissible cancers and may represent an adaptive feature of the tumour that helps prevent genomic decay (Strakova et al. 2020; Yonemitsu et al. 2019). Additionally, similar to patterns seen in human cancers, whole-genome duplication events were observed in CedBTN1 and CedBTN2, which may also represent an adaptive strategy to maintain the fitness of somatic cells in the context of high mutation burden and genomic instability (Merlo et al. 2010). Importantly, given the extended age of these transmissible cell lineages, this mechanism may be essential for maintaining the viability of tumour cells and could represent a convergent molecular adaptation in BTN to cope with high mutation rates during prolonged somatic evolution.

USC UNIVERSIDADE

According to our somatic screening, small but significant differences in structural variation burden might be found when comparing the genomes of CedBTN1 and CedBTN2 tumours

(Wilcoxon rank sum test p-value=0.017). Notably, all CedBTN2 tumour presented more alterations that CedBTN1 tumours (CedBTN2 range= 6,493 – 6,861 SVs, CedBTN1 range=4,923 – 5,454 SVs). Deletions (p-value = 0.01667), inversions (p-value = 0.02184) and breakends (p-value = 0.03333) were slightly more frequent in the genome of CedBTN2, but no significant differences were found regarding the number of duplications (p-value = 0.1167). These results might be conditioned by factors such as the relative age of each tumour lineage, differential mutation rates, but also by a technical difference in our ability to detect and filter ancestral germline variants from each founder population. Overall, we did not find substantial differences in the incidence of specific structural variant types, which suggests that shared ongoing mutational processes might have promoted structural variation in the genomes of CedBTN1 and CedBTN2.

An important limitation for our phylogenetic study is the challenge to fully discriminate between somatic and germline variation within the CedBTN1 and CedBTN2 cancer genomes, especially for older variant sets. This limitation is inherent to the ancient nature of these transmissible tumours, whose founder genome are expected to retain germline variants (SNVs, SVs and other polymorphisms) which might be underrepresented in modern populations because of the effect of genetic drift or natural selection. This situation has represented a challenge in the field when studying the genome of ancient transmissible cancer lineages, such as CTVT, estimated to be 4,000 – 8,500 years old, and MarBTN, estimated to be > 200 years old (Baez-Ortega et al. 2019; Hart et al. 2023). This situation is particularly complex to address under certain phylogenetic scenarios, such as population bottlenecks leading to allele loss. If founder population faces severe decline or extinction, discrimination can be extremely difficult, even when ancient host genomes are available (Ní Leathlobhair et al. 2018). Fine grain analysis of the *Cerastoderma edule* phylogeography and paleogenomic analyses might help to overcome these limitations, and shed more light into CedBTN genome evolution, but they are beyond the scope of this thesis.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

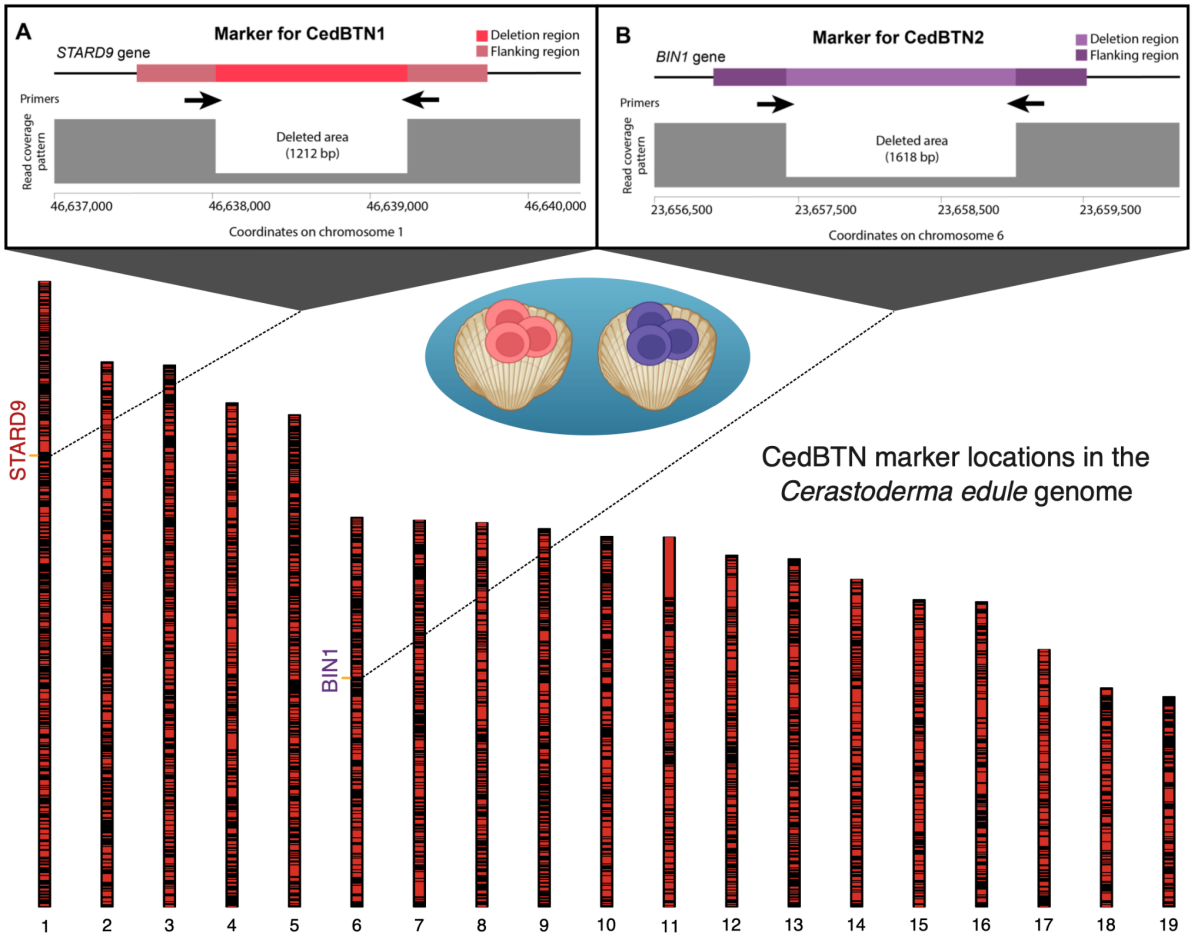


Figure 32: Genomic locations of the marker variants included in our diagnostic assay.
A) Marker deletion impacting on the gene *STARD9* (CedBTN1 lineage) on chromosome 1. B) Marker deletion impacting on the gene *BIN1* (CedBTN2 lineage) on chromosome 6. Primer locations are represented with illustrative arrows around the deletion breakpoints.

Table 8: Primer sequences designed for the detection of CedBTN1 and CedBTN2.

Primer ID	Coordinates	Sequence (5' - 3')	Amplicon size (Deletion +)	Amplicon size (Deletion -)
CedBTN1-F	chr1: 46637779-46637800	TTATGTGGTGCTAGGTGGGGAT	404 bp	1.6 kb
CedBTN1-R	chr1: 46639438 -46639415	GGAGAAATGCAAGACACCAGTAAG		
CedBTN2-F	chr6: 23657304-23657326	TCACTGAACCTTTGAATAGCTCA	160 bp	1.8 kb
CedBTN2-R	chr6: 23659125-23659105	ACCTTTGCTCATCTCAAGACA		

Abbreviations: CedBTN1/2=*Cerastoderma edule* bivalve transmissible neoplasia 1/2; F=Forward; R=Reverse; chr=chromosome; bp=base pairs; kb=kilobase pairs; chr=chromosome.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.2.3 - Candidate drivers of CedBTN genome evolution

We leveraged our genome-wide screening of nuclear variants to search for somatic alterations which might have been involved in early evolutionary stages of CedBTN (Figure 33). We focused on pre-divergence variants (those found in most tumours from each lineage, thus corresponding to early-onset mutations) impacting conserved gene sequences with homology with well-established human cancer genes. Some of these alterations, likely acquired during the first stages of CedBTN1 and CedBTN2 evolution, might comprise molecular driver events involved in the evolution of these transmissible cancer clones.

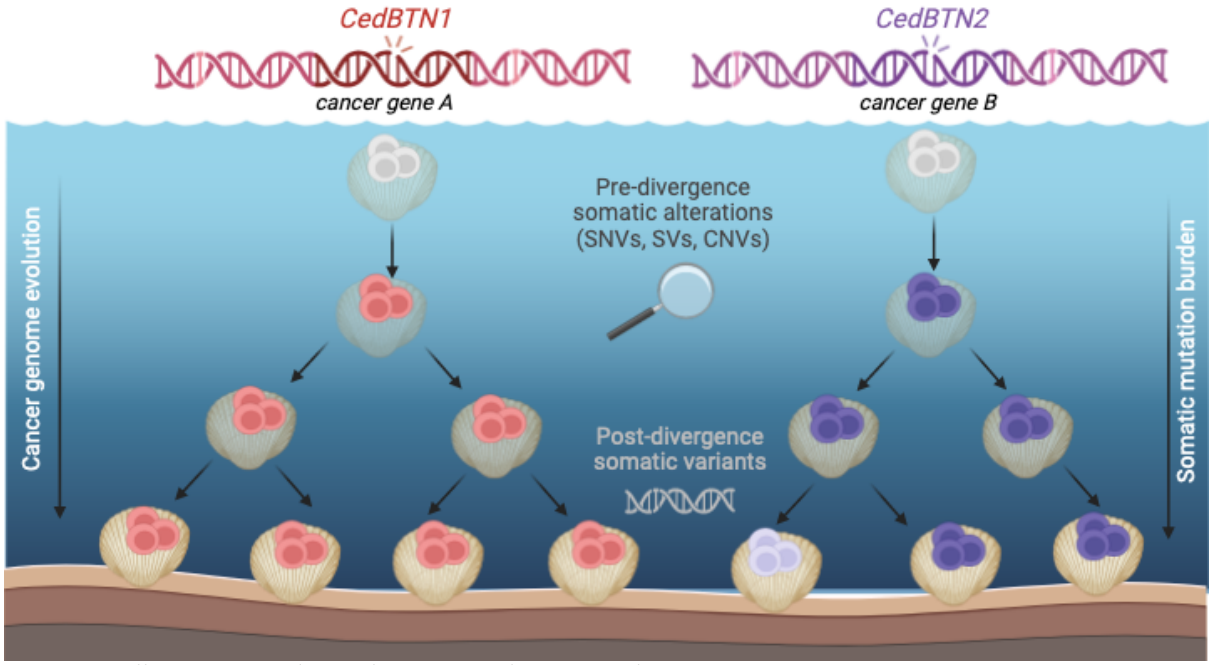


Figure 33: Illustration of the Scuba Cancers driver search strategy.
We leveraged our genome-wide screening of nuclear variants to search for somatic genetic alterations which might have presented a significant impact in the first stages of CedBTN1 and CedBTN2 cancer genome evolution. For this purpose, we screened for pre-divergence, early-onset tumour variants (SNVs, SVs, CNVs) expected to generate high functional impact on annotated well-established cancer genes. Created with BioRender.com.

C2.2.3.1 - Impact of CedBTN short variation on cancer genes

Our genome-wide screening of short-variants identified a total of 20,023 candidate somatic, pre-divergence mutations impacting *Cerastoderma edule* cancer genes (Figure 34). In CedBTN1, this variant set comprised a total of 11,635 short variants (10,792 SNVs, 690 MNVs and 153 Indels) impacting 200 cancer genes. In CedBTN2, this variant set comprised a total of 8,388 short variants (7,754 SNVs, 548 MNVs and 86 Indels) impacting 196 cancer genes. As expected, only a small fraction of these variants (888 / 20,023; 4.4%) impacted on coding regions of the target genes, while remaining alterations affected intronic (71.3%), UTR (5.2 %), upstream (9.1%) or downstream (9.7%) cancer gene regions.

We identified a total of 547 variants with a predicted significant (moderate to high) impact in cancer gene function (300 in CedBTN1, and 247 in CedBTN2). These alterations were typically missense variants (484 mutations), but also included variants of greater effect, such as 14 frameshift variants, 19 splicing variants and 25 stop-gained variants. We focused on those genes for which the same type of significant alteration had been already described for the same gene

in human cancer (COSMIC). We found 56 cancer genes in CedBTN1 and 55 cancer genes CedBTN2 impacted by these kinds of alterations. Cancer genes affected by higher-impact variants are reported in Table 9. Affected genes by candidate variants include relevant loci involved in DNA repair and chromosomal stability (e.g. *NBN*, a gene involved in DNA double-strand break repair and DNA damage-induced checkpoint activation), cell cycle control (e.g. *RB1*, a gene involved in G1/S transition and chromatin regulation), and immune response (e.g. *KEAP1*) ((Sondka et al. 2024)). Interestingly, *STAG2*, a gene involved in the proper separation of sister chromatids during cell division and a potential driver of aneuploidy, was found to be impacted in both CedBTN1 and CedBTN2.

Unfortunately, the integration of pre-divergence short variants in CedBTN1 or CedBTN2 with associated copy number profiles did not allow to identify high-confidence candidate cancer drivers. For most of the affected genes, local copy number profiles associated with BTN aneuploidy supported the presence of reference alleles for potentially impacted tumour suppressors (Table 9, Supp. Table 3). Additionally, the analysis of dN/dS ratio yielded no evidence of positive selection for post divergence cancer gene SNVs in either lineage. For this reason, even when this thesis reports these short variant alterations, these events cannot be catalogued as confident drivers of CedBTN1 and CedBTN2 evolution. Their potential relevance for bivalve transmissible cancer development and (or) progression must be further explored in future investigations.

Table 9: Candidate Cancer Genes impacted by high-impact short variants in CedBTN.

Gene	Role	Variant consequence	Variant coordinates	Lineage	Gene CN
MEN1	TSG	stop gained frameshift variant	chr1:60325745 G>A (SNV) chr1:60325762 T>TC (Indel)	CedBTN2	4.7
MAP3K1	oncogene TSG	stop gained	chr2:47406335 C>T (SNV)	CedBTN2	5.7
TSC2	TSG	stop gained stop gained	chr3:40972285 A>T (SNV) chr3:40976941 C>T (SNV)	CedBTN1	7.5
KMT2D	oncogene TSG	stop gained	chr6:2005095 C>A (SNV)	CedBTN1	4.8
CIC	oncogene TSG	spliced donor variant	chr6:34253431 A-T>A (Indel)	CedBTN1	2.7
RECQL4	oncogene TSG	stop gained	chr7:9192850 TA>AT (Indel)	CedBTN1	3.4
NBN	TSG	frameshift variant	chr7:23938706 T>TC (Indel)	CedBTN1	3.7
TBX3	oncogene TSG	frameshift variant	chr8:20417234 TC>T (Indel)	CedBTN2	5.4
PPM1D	oncogene	stop gained	chr8:32727221 GG>AA (MNV)	CedBTN2	6.6
DNM2	TSG	splice acceptor variant	chr15:12433191 G>A (SNV)	CedBTN1	4.1
RB1	TSG	splice donor variant	chr16:13768100 T>A (SNV)	CedBTN1	4.2
STAG2	TSG	frameshift variant splice donor variant	chr16:18521827 GA>G (Indel) chr16:18523654 G>A (SNV)	CedBTN1 CedBTN2	7.5 4.5
KEAP1	TSG	frameshift variant frameshift variant	chr17:15474229 GC>G (Indel) chr17:15474231 AC>A (Indel)	CedBTN1	2.2
PTPRT	TSG	stop gained	chr18:2134066 G>T (SNV)	CedBTN1	2.3
SETD2	TSG	stop gained	chr19:19031036 G>T (SNV)	CedBTN1	3.9

Abbreviations: TSG=tumour suppressor gene; SNV=single-nucleotide variant; MNV=multi-nucleotide variant; Indel=insertion-deletion; CN=copy number; chr=chromosome.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

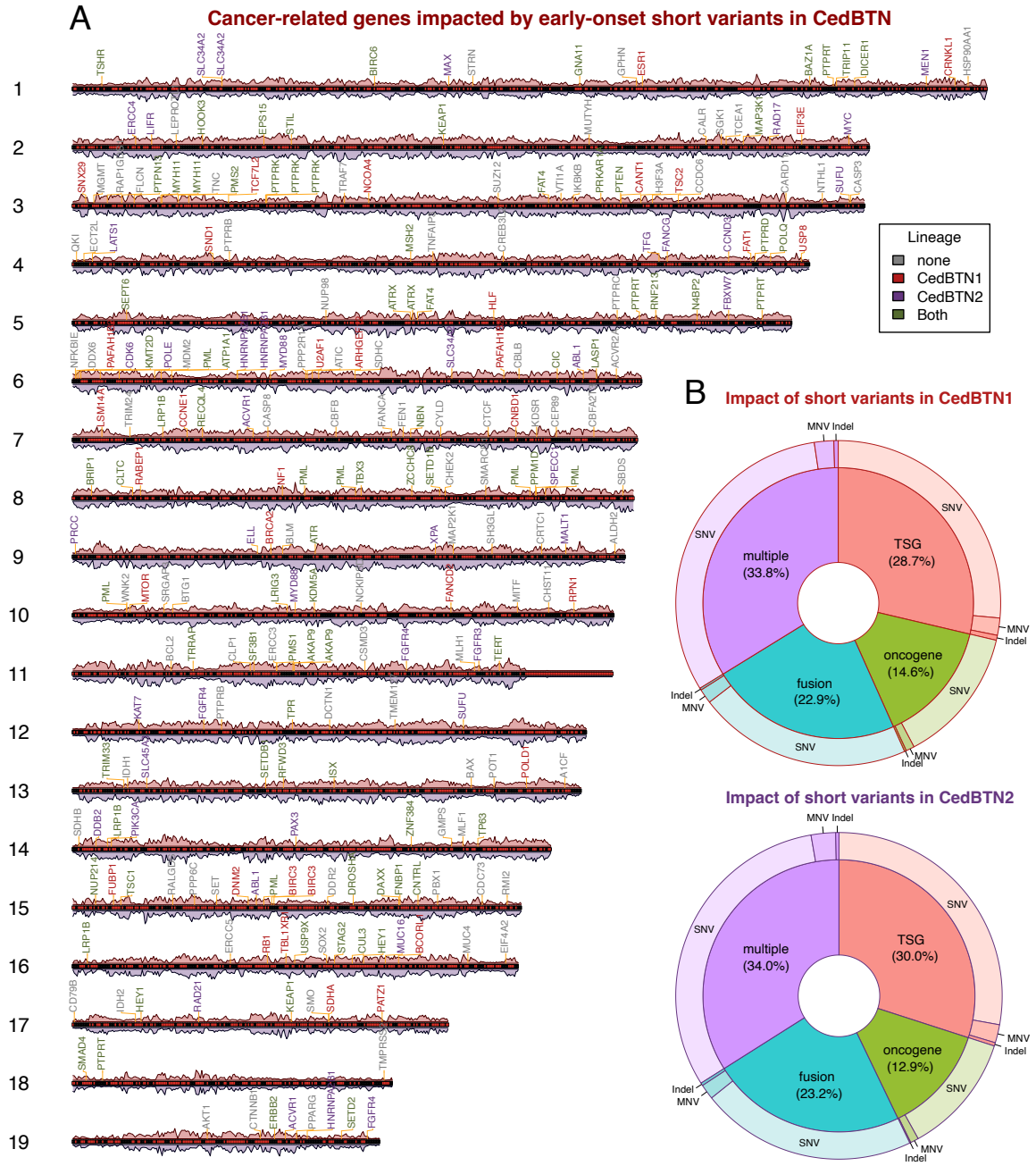


Figure 34: Impact of CedBTN short variation in cancer genes.
A) Genomic distribution of cancer genes impacted by early-onset short variants with significant impact in CedBTN1 (red), CedBTN2 (purple) or in both lineages (green). Density tracks indicate the distribution of early-onset short variants on each lineage. B) Circle charts showing cancer gene roles (tumour suppressor genes, oncogenes, fusion genes and cancer genes with multiple roles) impacted by early-onset short variant types (SNVs, MNVs and Indels) in CedBTN1 (upper) and CedBTN2 (lower).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.2.3.2 – Impact of CedBTN structural variation on cancer genes

Our genome-wide screening of structural variation identified a total of 221 somatic, early-onset alterations impacting *Cerastoderma edule* candidate cancer genes (Figure 35). In CedBTN1, this variant set comprised a total of 97 structural variants (52 deletions, 28 duplications, 7 inversions and 10 breakends) impacting 95 cancer genes. In CedBTN2, this variant set comprised a total of 124 structural variants (85 deletions, 27 duplications, 7 inversions and 5 breakends) impacting 98 cancer genes.

As expected, only a small fraction of these variants caused a significant impact on protein coding gene sequences. We focused on those genes for which the same type of alteration (amplification, ablation, etc) had been described for the same gene in human cancer. We found 2 cancer genes in CedBTN1 and 6 cancer genes in CedBTN2. Affected genes are reported on Table 10 and include: *MEN1* (CedBTN2), *RAD17* (CedBTN1), *CASP3* (CedBTN2), *PTEN* (CedBTN2), *SUFU* (CedBTN2), *NF1* (CedBTN2), *BRCA2* (CedBTN1 and CedBTN2), *TSC1* (CedBTN2). We found large deletions of cancer genes involved in DNA repair, chromosomal stability and apoptosis, such as *BRCA2*, *RAD17* and *CASP3* (full transcript ablation) (Sondka et al. 2024). Interestingly, *BRCA2* was found to be altered both in CedBTN1 and CedBTN2 genomes. *BRCA2* alterations are associated with an increased burden of small somatic deletions in human cancer (Yi and Ju 2018).

Unfortunately, the integration of pre-divergence structural variants (typically deletions) in CedBTN1 or CedBTN2 with associated gene copy number data (often exceeding the tetraploid state, consistent with multiple copies of each tumour suppressor) did not lead to the identification of any high-confidence candidate cancer driver (Table 10, Supp. Table 3). Therefore, even when this thesis reports these alterations, they cannot be catalogued as confident drivers of CedBTN1 and CedBTN2 evolution. Their potential relevance for bivalve transmissible cancer development and (or) progression requires further investigation.

Table 10: Candidate Cancer Genes impacted by high-impact structural variants in CedBTN.

Gene	Gene coordinates	Role	Variant type	Coordinates	Lineage	Gene CN
MEN1	chr1: 60325668-60340565	TSG	SV (deletion)	chr1: 60294218-60373198	CedBTN2	4.7
RAD17	chr2: 49263095-49284342	TSG	SV (deletion)	chr2: 41968381-51112158	CedBTN1	4.5
CASP3	chr3: 54988740-55001415	TSG	SV (deletion)	chr3: 52944272-55821235	CedBTN2	6.9
CASP3	chr3: 54988740-55001415	TSG	SV (deletion)	chr3: 54995168-54997051	CedBTN2	6.9
PTEN	chr3: 38721178-38749689	TSG	SV (deletion)	chr3: 38730971-38731423	CedBTN2	5.3
SUFU	chr3: 54854750-54919360	TSG	SV (deletion)	chr3: 52944272-55821235	CedBTN2	8
NF1	chr8: 14799580-14896067	TSG	SV (deletion)	chr8: 14575309-14827624	CedBTN2	2.9
BRCA2	chr9: 14026508-14037315	TSG	SV (deletion)	chr9: 6645348-15565967	CedBTN1	3
BRCA2	chr9: 14026508-14037315	TSG	SV (deletion)	chr9: 14018716-14080689	CedBTN2	6
TSC1	chr15: 3417855-3446737	TSG	SV (deletion)	chr15: 3422532-3428683	CedBTN2	4.7

Abbreviations: TSG=tumour suppressor gene; SV=structural variant; CN=copy number.

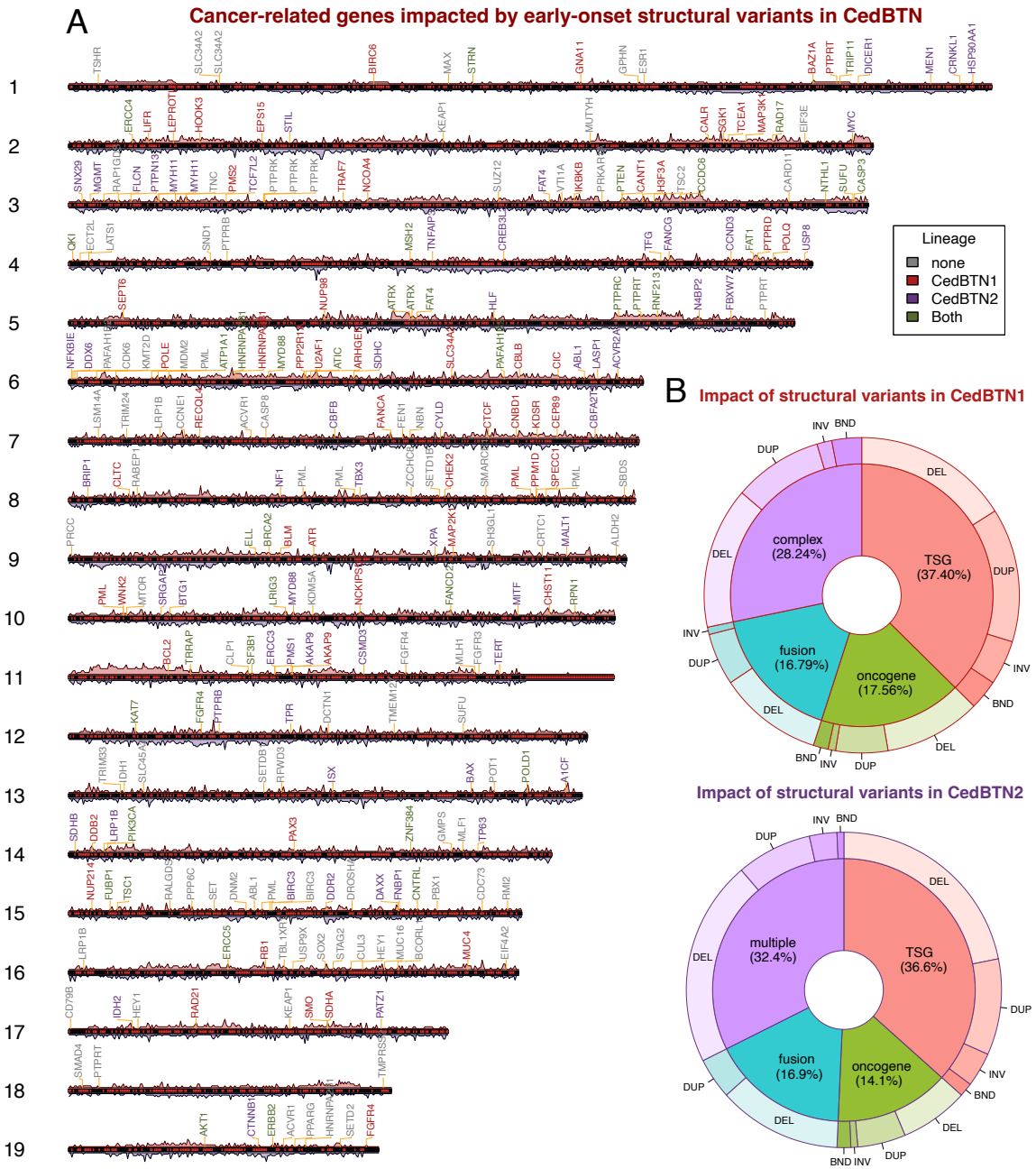


Figure 35: Impact of CedBTN structural variation in cancer genes.
A) Genomic distribution of cancer genes impacted by early-onset structural variants with significant impact in CedBTN1 (red), CedBTN2 (purple) or in both lineages (green). Density tracks indicate the distribution of early-onset structural variants on each lineage. B) Circle charts showing cancer gene roles (tumour suppressor genes, oncogenes, fusion genes and cancer genes with multiple roles) impacted by early-onset structural variant types (deletions, duplications, inversions and breakends) in CedBTN1 (upper) and CedBTN2 (lower).

C2.2.3.3 - Impact of CedBTN copy number variation on cancer genes

The availability of copy number data offered another opportunity to identify potential early-driver CN alterations in CedBTN. We systematically screened for significant gains and losses

of regions containing oncogenes and tumour suppressor genes, respectively (Figure 36). This analysis identified 30 cancer genes in CedBTN1 (8 TSGs and 22 oncogenes) and 30 cancer genes in CedBTN2 (10 TSGs and 20 oncogenes) with potentially relevant copy number alterations (+/-2 CN variation) (Figure 37). Tumour suppressor genes with hemizygous losses were explored in search for other high-impact SNV and SVs which might disrupt remaining alleles.

Our analysis identified likely ancestral amplifications involving two canonical oncogenes in CedBTN1 (Table 11): *MDM2* (10-13 copies; mean CN = 10.9, gene CN percentile 98.3), encoding an important inhibitor of tumour suppressor proteins, and *CCND3* (8-18 copies; mean CN = 10.7, gene CN percentile = 98.2), encoding a cyclin that promotes G1/S cell cycle transition. Recurrent amplification of these genes has been observed in multiple cancer types, and it is thought to prevent cell cycle arrest and apoptosis under conditions of genomic instability. In CedBTN2, we found evidence of an ancestral *MYC* amplification (7-11 copies; mean CN = 9.2, gene CN percentile = 96.3). *MYC* amplification often leads to cell cycle alterations in several human tumour types, and its upregulation has also been proposed as an early driver of a mammalian transmissible cancer in CTVT (Baez-Ortega et al. 2019). Notably, we also identified an ancestral homozygous deletion of *MGMT* in CedBTN1 (Figure 38). The enzyme encoded by this gene -a DNA methyltransferase - is essential for the repair of alkylated DNA bases, and its inactivation results in hypersensitivity to the toxic and mutagenic effects of alkylating agents. Interestingly, in the context of the ERC Scuba Cancers project, a CedBTN1-specific mutational signature was identified that closely resembles human COSMIC SBS11, which is associated with alkylating agents such as temozolomide. According to these results, it seems likely that this ancestral loss of *MGMT* in CedBTN1 is the key event responsible of the accumulation of unrepaired DNA damage caused by alkylated DNA bases in this transmissible cancer lineage.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

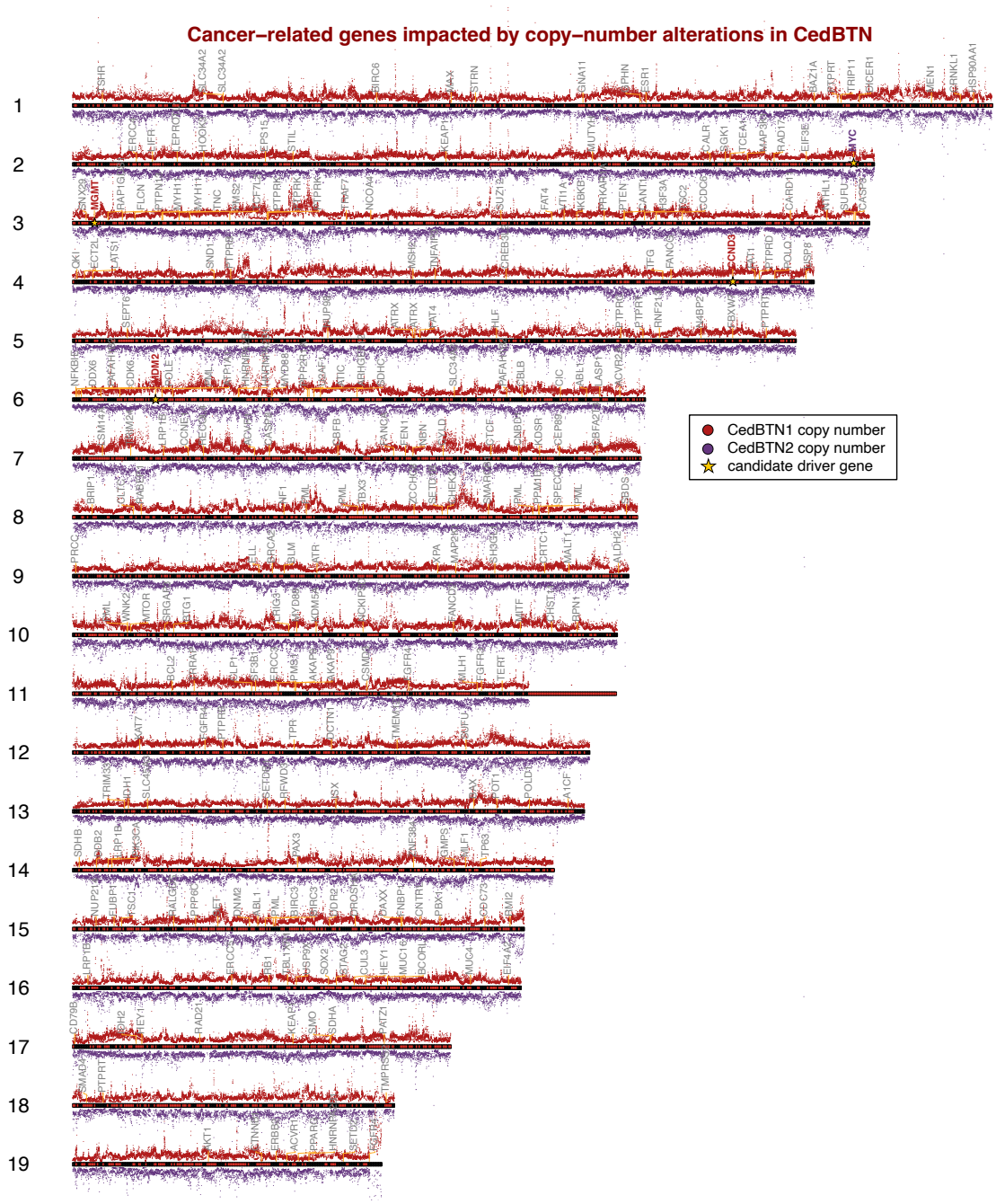


Figure 36: Analysis of cancer genes impacted by copy-number alterations in CedBTN.
The upper red dot plot represents the copy-number profiles found in CedBTN1 genomes. The lower purple dot plot represents the copy-number profiles found in CedBTN2 genomes. Cancer genes with somatic alterations are denoted with labels on their genomic location across the *Cerastoderma edule* chromosomes (central ideogram).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

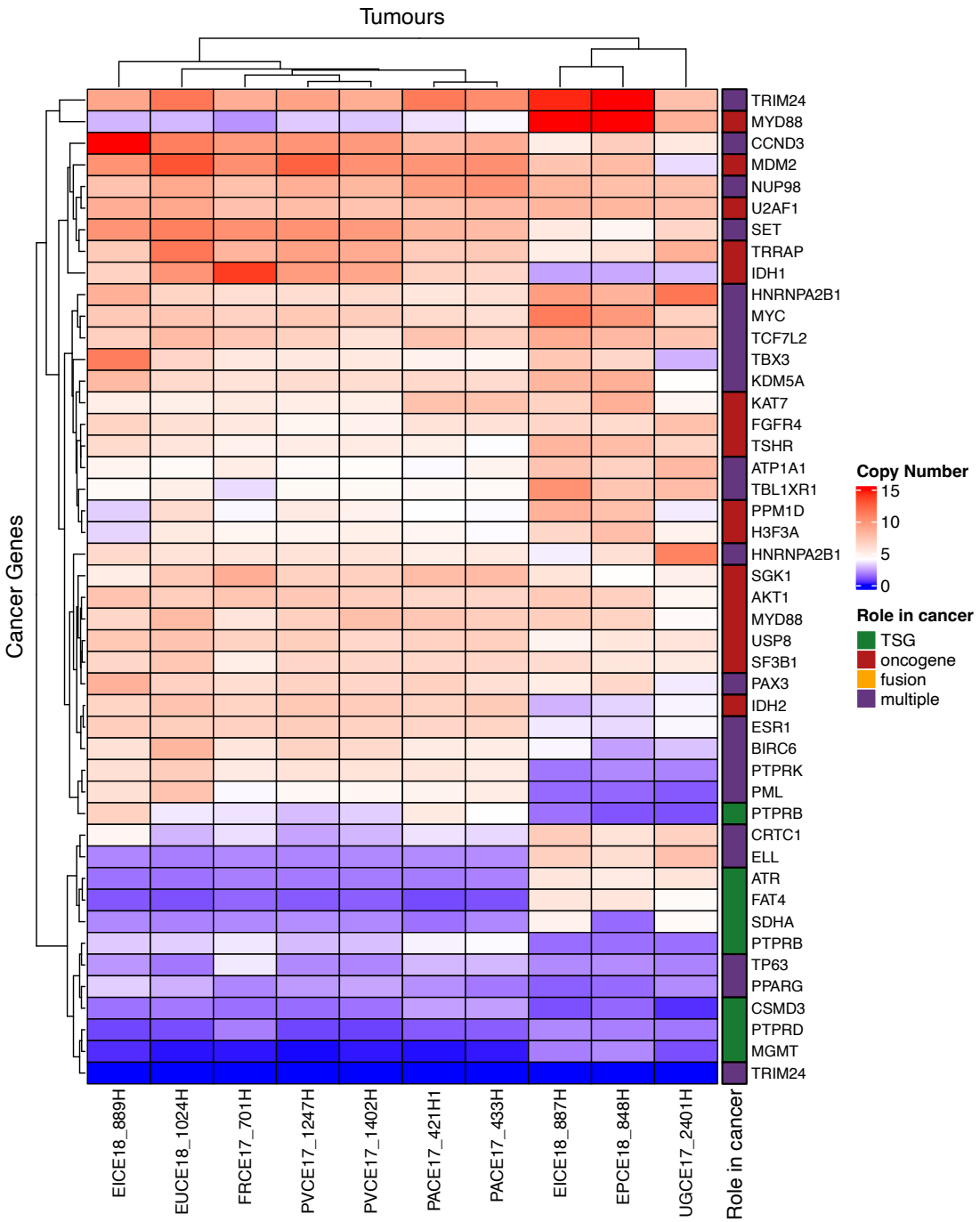


Figure 37: Analysis of CedBTN copy number alterations on cancer genes.
Gene-wise copy number estimates for relevant cancer genes in CedBTN1 and CedBTN2 tumours. Only relevant genes (TSG and oncogenes) with average copy number variation of +/- 2 copies from basal tumour ploidy (4n) are displayed. Role in cancer associated with each gene (TSG, oncogene, fusion or multiple) is indicated in the right column. Both samples (columns) and cancer genes (rows) are grouped according to a hierarchical clustering.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

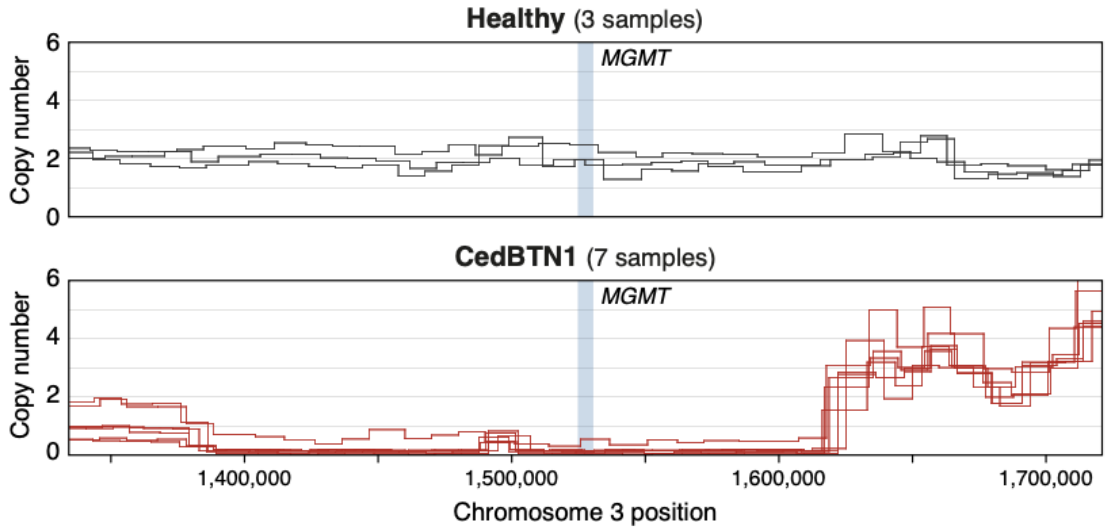


Figure 38: Copy number profiles of the *MGMT* gene.
Copy number profiles of the *MGMT* tumour suppressor gene locus and surrounding genomic regions (500 KB) in non-neoplastic (n=3) and CedBTN1 (n=7) tumour samples. Each sample is represented by a solid line. The highest copy number estimate for *MGMT* (0.4) corresponds to EICE18_889H. Credit: Adrián Baez-Ortega.

We examined cancer gene expression in relation to normal tissues for these early driver candidates and found evidence of increased expression of amplified genes *CCND3*, *MDM2* and *MYC* in the relevant CedBTN lineage, as well as absence of MGMT expression in CedBTN1 (Figure 39). Remarkably, we also observed overexpression of *MDM2* in CedBTN2 relative to normal tissues, perhaps related to moderate amplification of these genes in CedBTN2 (3-8 copies). These findings support the conclusion that CN alterations of *CCND3*, *MDM2*, *MGMT* and *MYC* are likely drivers of early CedBTN evolution and raise the possibility that upregulation of *MDM2* has been independently selected in both cancer lineages. Despite the high completeness of our *Cerastoderma edule* genome assembly, we cannot exclude the possibility that further early driver events have escaped detection due to the lack of homology between the set of cancer genes in humans and bivalves.

Table 11: Candidate ancestral drivers.

Gene Symbol	*CCND3*	*MDM2*	*MGMT*	*MYC*
Lineage	CedBTN1	CedBTN1	CedBTN1	CedBTN2
Coordinates	chr4: 46361343-46371948	chr6: 5818996-5829375	chr3: 1524727-1530540	chr2: 54860017-54866217
Cancer role	Oncogene	Oncogene	Tumour suppressor	Oncogene
Event	Amplification	Amplification	Homozygous deletion	Amplification
Percentile	98.2%	98.3%	Not applicable	96.3%
Mean CN	10.7	10.9	0.2	9.2
Median CN	9.8	10.3	0.1	9.8
Minimum CN	8	10	0	6.7
Maximum CN	18	13.1	0.4	11.2

Abbreviations: CedBTN1/2=*Cerastoderma edule* bivalve transmissible neoplasia 1/2; chr=chromosome.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

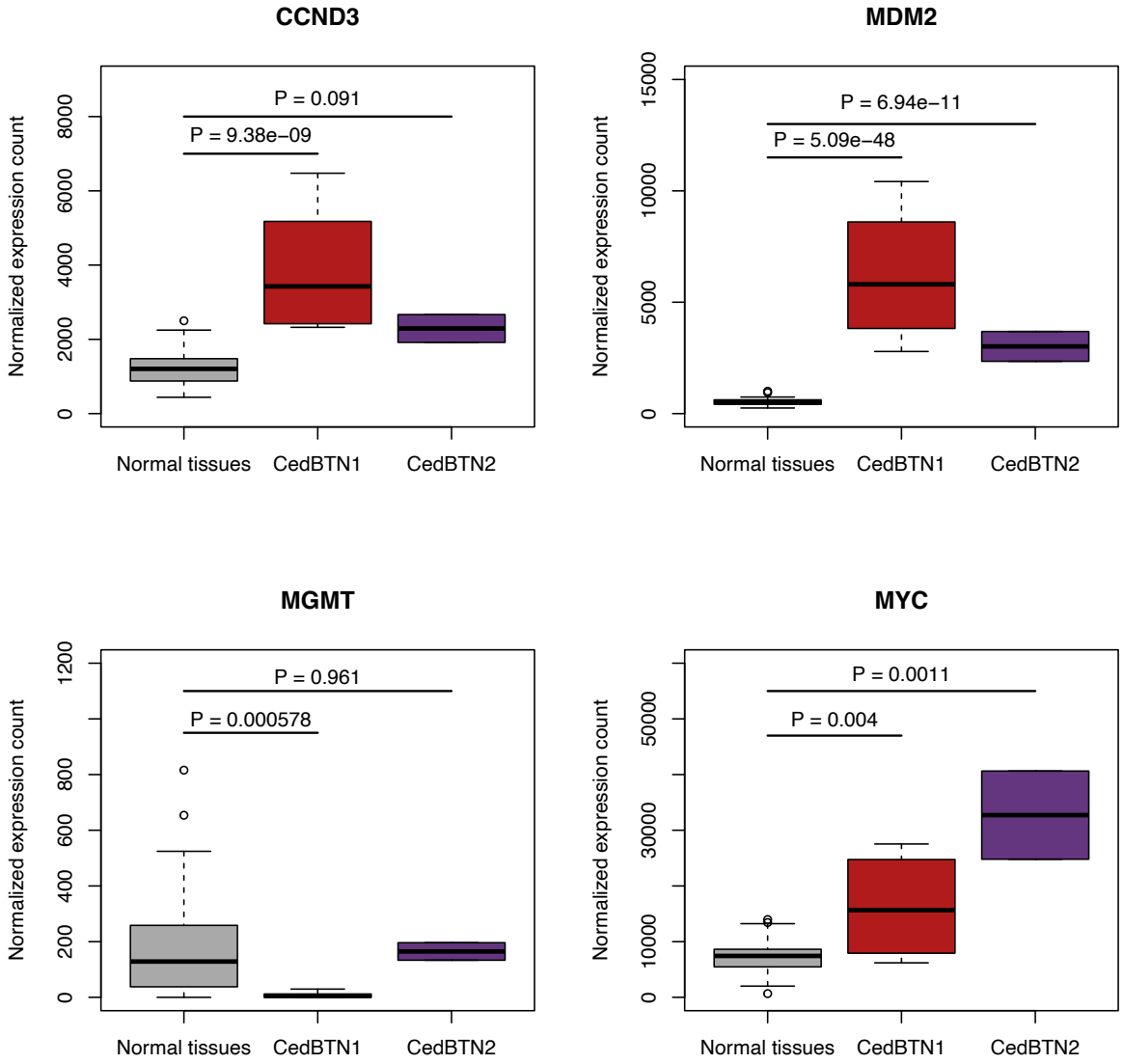


Figure 39: Expression of cancer genes with potential driver copy number alterations.
For each cancer gene, normalized expression counts are shown for normal tissue samples (n=28), CedBTN1 samples (n=6) and CedBTN2 samples (n=2). Adjusted p-values were obtained through differential expression analysis (two-sided Wald test with Benjamini-Hochberg correction) based on normalized RNA-seq read counts. Normalized count values are comparable across samples for the same gene but are not comparable across genes. Analysis performed by Adrián Baez-Ortega, after the identification of somatic copy number alterations in CedBTN.

According to these results, the ancestral amplification of cancer genes such as *MDM2* and *CCND3* (both in terms of copy number and gene expression) probably represent CedBTN adaptations to survive in a context of long-term genomic instability. This might occur through the disruption of *TP53* responses against aneuploidy and DNA damage, as found in other vertebrate and invertebrate species. Remarkably, the *MDM2* binding site in its transcriptional activation domain has been shown to be conserved across p53 proteins of several bivalve species (C. W. Walker et al. 2011). Alterations in the p53 pathway at protein levels had been previously informed for *Cerastoderma edule* neoplasia (Díaz et al. 2010). Upregulation of mutant p53 proteins has been described for multiple bivalves affected by disseminated neoplasia. In addition, p53 disruption through cytosolic sequestration has been identified as a potential mechanism of cancer evolution in MarBTN (Hart et al. 2023; C. Walker, Böttger, and Low 2006), being a process also described on several human cancers. Other ancestral molecular

alterations of genes involved in DNA repair and cell cycle and mitotic control, such as *MYC* upregulation, *MGMT* downregulation and *STAG2* somatic mutations might potentially have contributed to the carcinogenic phenotype in *Cerastoderma edule*. Interestingly, different *MYC* alterations have been reported both in bivalve (Burioli et al. 2023) and mammalian transmissible cancer (Baez-Ortega et al. 2019). Collectively, all these genetic alterations may have promoted, or at the very least created a tolerance for, the widespread genomic imbalance that appears to characterize CedBTN tumour genomes.

Remarkably, we were not able to fully characterize high-confidence driver mutations shared by CedBTN1 and CedBTN2 lineages. However, the process of whole-genome doubling, reported in diverse transmissible cancers (Hart et al. 2023; Stammnitz et al. 2023), might represent a key genomic adaptation independently selected on both tumour lineages. This phenotype, potentially driven by chromosome missegregation during mitosis or spontaneous cell fusions (Miroshnychenko et al. 2021), likely plays a critical role in the long-term evolution of these cancer lineages buffering the impact of somatic mutation burden (Bramwell et al. 2024). Selection of whole genome doubling events in cancer has been associated with tumours showing extensive loss of heterozygosity and ongoing chromosomal instability (López et al. 2020). Further investigation is needed to characterize late-onset CedBTN driver mutations in the context of redundant gene content and uncertain dosage compensation. These alterations might include additional variants crucial for the survival of tumour cells in the ocean environment, and for promoting immune evasion when infecting host bivalves.

C2.3 – CONCLUSIONS OF CHAPTER II

I. The genome of the bivalve transmissible neoplasia of *Cerastoderma edule* has a notable burden of somatic structural variation. Deletions are the most common type of somatic structural alterations, followed by duplications and potential interchromosomal rearrangements. These variants are widely distributed across all the transmissible CedBTN chromosomes.

II. Two divergent lineages of the bivalve transmissible neoplasia coexist in the marine environment: CedBTN1 and CedBTN2, as supported by complementary genetic variants (SNVs and SVs). The large fraction of private candidate somatic variants, along with their divergence from existing cockle populations, suggests that both lineages may be extremely old.

III. Oncogene amplification (*MDM2*, *CCND3*, and *MYC*) and tumour suppressor losses (*MGMT*) might be among the driver genetic alterations driving CedBTN1 and CedBTN2 early evolution. The complex copy number alterations profiles found in CedBTN tumours, including whole genome doubling events, hinder conclusive findings regarding additional specific driver alterations.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C2.4 - PUBLICATIONS

The main contents of this chapter have been published as an article in Nature Cancer [1]. This publication represents a collaborative effort aimed at advancing our understanding of the long-term somatic evolution of transmissible cancers. My contributions to this work were focused on the investigation of structural variation patterns and clonal somatic evolution of the *Cerastoderma edule* bivalve transmissible neoplasia. I conducted a genome-wide structural variation calling in CedBTN, genotyping the identified alterations in both tumours and a diverse panel of common cockles. I discriminated between germline and somatic variation in CedBTN and analysed the genomic distribution of somatic alterations. To further characterize the clonal structure of the neoplasia, I estimated ML and MP phylogenies based on large sets of SNVs and SVs and explored CedBTN variation in the context of the host species genetic diversity. I performed a comprehensive screening of high-impact somatic mutations – including SNVs, INDELs, SVs and CNVs - in CedBTN1 and CedBTN2, identifying candidate driver alterations in several cancer genes. I identified molecular markers to discriminate CedBTN1 and CedBTN2 tumours. I examined the distribution of repetitive content within the reference genome of the common cockle. I created circos plots to visualize genomic rearrangements affecting CedBTN chromosomes. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work. The contents of the publication have been adapted to align with the format of this thesis. Further detail regarding author's contributions is provided in the appendix, alongside the full list of authors and their affiliations.

References:

1. Bruzos AL, Santamarina M, García-Souto D, et al. Somatic evolution of marine transmissible leukemias in the common cockle, Cerastoderma edule. *Nat Cancer*. 2023;4(11):1575-1591. doi:10.1038/s43018-023-00641-9

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER III: PREDICTION OF MICROSATELLITE INSTABILITY IN COLORECTAL CANCER

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CHAPTER III

Prediction of microsatellite instability in colorectal cancer

C3.1 - BACKGROUND

C3.1.1 - Microsatellite DNA sequences

Microsatellites, also known as short tandem repeats (STRs), are a particular type of short repetitive DNA sequences which are spread throughout the genome (Ruiz-Bañobre and Goel 2019). They are composed of tandems of small genetic units (e.g. AAT, CGG, CA motifs) typically ranging from one to six nucleotides (Bagshaw 2017). The number of repetitive units in each microsatellite can be highly variable (usually 10 - 60 motifs), leading to microsatellite sizes which generally span from only a few bases up to several hundred base pairs (Baretti and Le 2018). Notably, microsatellite DNA elements comprise some of the genomic locations with higher mutation rates, and are important drivers of genome evolution both at genic and intergenic chromosome regions (Bagshaw 2017; Schlötterer 2000).

Over the course of evolution, microsatellite DNA elements have become very abundant in both prokaryotic and eukaryotic genomes. It is estimated that they might represent around 3-5% of nuclear genetic content in mammalian genomes (Shikano et al. 2010). In the human genome, they cover 3% of the genomic sequence, represented by more than 700.000 annotated copies in the GRCh38 human reference genome (Nassar et al. 2023; Sawaya et al. 2013). Interestingly, these elements constitute a relevant cause of genetic variation across individuals and populations. Many of them constitute highly polymorphic *loci* due to variation in the number of repeat units, being valuable markers in molecular biology for phylogenetic tracing and DNA fingerprinting (Haddrill 2021). In addition, different microsatellite DNA elements are involved in important aspects of gene regulation (e.g. transcription factor binding, nucleosome packaging, alternative splicing) (Bagshaw 2017; Sawaya et al. 2013). For this reason, the alteration of these elements also play a significant role in the development of genetic diseases, such as cancer and neurological disorders (Bagshaw 2017).

C3.1.2 - Sources of microsatellite instability (MSI)

The origin and evolution of microsatellite elements is intimately associated with their repetitive nature. Microsatellites often arise as a consequence of DNA replicating errors, specifically, due to backward or forward polymerase slippage during the synthesis of a new DNA strand (Schlötterer 2000). When this process occurs and it is not corrected by the cell machinery, it can lead to the addition or removal of repeat units, driving to a microsatellite expansion or contraction, respectively (Figure 40). For these reasons, microsatellite *loci* are particularly prone to the accumulation of mutations, specially under contexts of disrupted DNA mismatch repair and genetic instability (Baretti and Le 2018).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

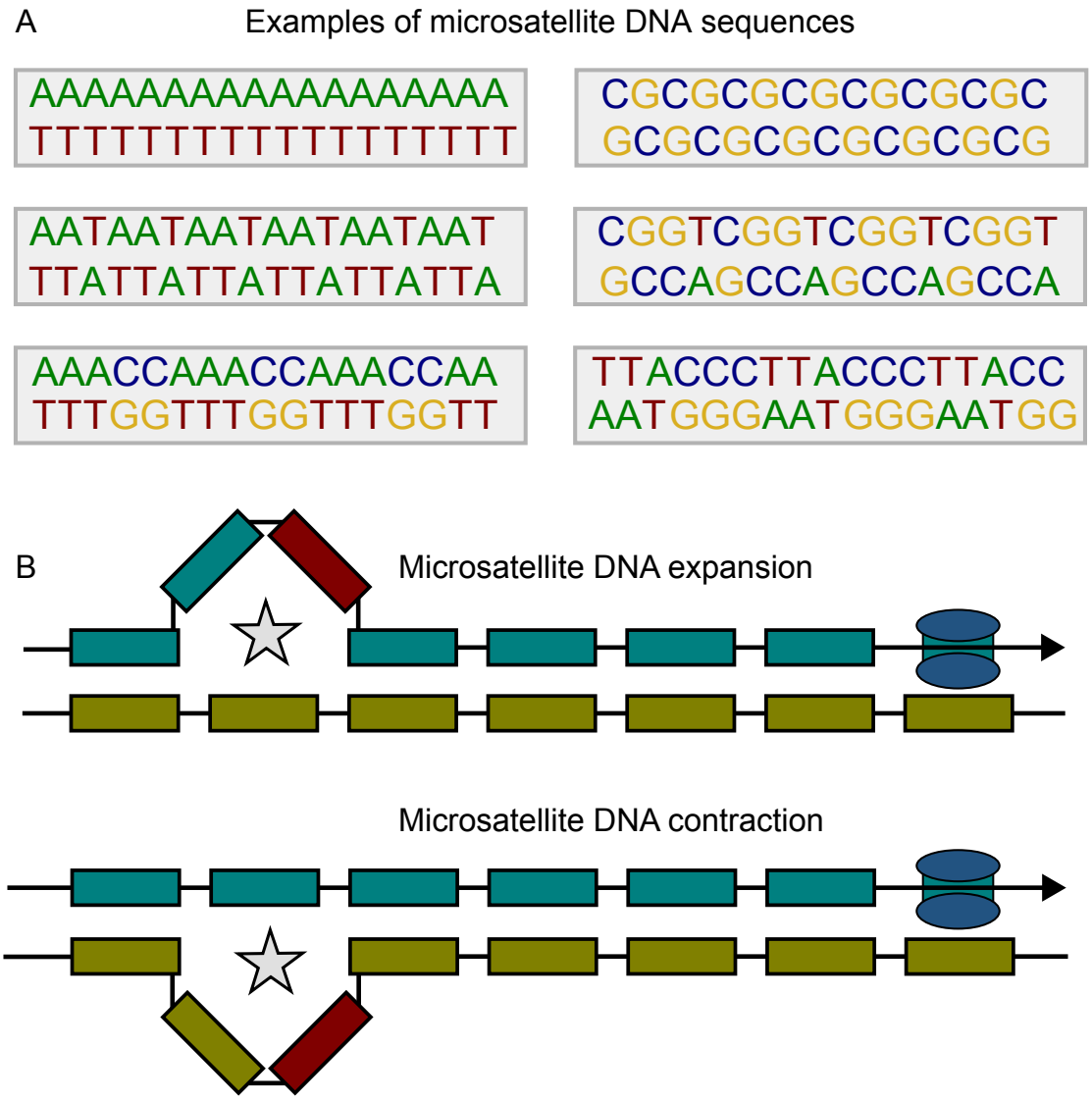


Figure 40: Microsatellite DNA overview.
A) Examples of typical microsatellite DNA configurations: mononucleotide repeats (I), dinucleotide repeats (II), trinucleotide repeats (III), tetranucleotide repeats (IV), pentanucleotide repeats (V), hexanucleotide repeats (VI). B) Model of microsatellite DNA instability generated during DNA replication. Top: polymerase backward slippage leading to one motif gain in the microsatellite. Bottom: polymerase forward slippage leading to one motif loss in the microsatellite.

Deficiency in DNA mismatch repair (MMR) activity usually leads to a hypermutator phenotype, termed microsatellite instability (MSI), which is characterized by the systematic accumulation of indels and single nucleotide substitutions within these DNA microsatellite regions (Ruiz-Bañobre and Goel 2019; Jiricny 2006). It is estimated that the MMR system is responsible for increases in the replication fidelity of 100 to 1000 fold (Martín-López and Fishel 2013), and as consequence of its disruption, MSI phenotype usually is characterized by drastic increases in the somatic mutation rates across the human genome. This process often results in altered signalling transduction, DNA repair, transcriptional regulation, protein translocation and modifications, and immune recognition (Martín-López and Fishel 2013). These alterations can

compromise cell function and commitment, and for this reason, control mechanisms such as cell cycle arrest and apoptosis might be initially triggered by this phenotype. However, resistance mechanisms can be acquired during the evolution of somatic cell lineages. For example, when this condition arises in the context of cancer cells, resistance to apoptosis is frequently observed, and the additional mutation burden provided by MSI can sometimes confer a selective advantage to certain neoplastic clones.

The MMR deficiency can be originated from germline and somatic mutations or epigenetic alterations in any of the canonical MMR genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*), as well as deletions in the *EPCAM* gene (which leads to constitutional repression of *MSH2* through promoter methylation) (Ruiz-Bañobre and Goel 2019). The proofreading activity of DNA polymerase-ε (*POLE*) or DNA polymerase-δ (*POLD*) is also crucial for the efficiency of the MMR. Collectively, MMR genes are involved in relevant stages of the DNA repair process, and their encoded products participate in a coordinated manner conforming heterodimers (e.g. MutSα, MutSβ, MutLα, MutLβ, MutLγ) (Baretti and Le 2018). These protein complexes specifically identify and remove mismatches and small insertion or deletion loops that arise during DNA replication, driving a correct resynthesis. When these MMR components are altered, drastic increments of the mutation rate are found across the genome, and the MSI phenotype is promoted (Ruiz-Bañobre and Goel 2019; Baretti and Le 2018).

C3.1.3 – Role of MSI in human cancer and immunotherapy

Over the last decades, the role of MSI in cancer development has been investigated, identifying several cancer types where MMR and microsatellite DNA alteration plays a relevant role.

The Lynch syndrome (LS) and the Constitutional Mismatch Repair Deficiency (CMMRD) are two hereditary cancer-associated syndromes driven by germline MMR alteration (Ruiz-Bañobre and Goel 2019). The first one (LS) is an hereditary cancer syndrome characterized by germline mutations in DNA mismatch repair genes, leading to an increased risk of colorectal, endometrial and other cancers, generally occurring at an earlier age (Martín-López and Fishel 2013). In the LS, somatic loss of function of the remaining wild-type allele of the germline altered MMR gene is required, and this can occur either through genetic or epigenetic alteration. The second (CMMRD), is an aggressive, highly penetrant childhood cancer predisposition syndrome, caused by homozygous germline mutations in any of the four canonical MMR genes (Giardiello et al. 2014). Both diseases are associated with severe disruption of the MMR machinery and are associated with the development of tumours with MSI.

MSI is also a molecular condition involved in the development of several types of sporadic human cancer (Baretti and Le 2018; Cortes-Ciriano et al. 2017; Kautto et al. 2016). Tumours with somatic alterations of the MMR system genes tend to acquire hypermutator phenotypes associated with MSI (Ruiz-Bañobre and Goel 2019). These somatic alterations can occur at genetic or epigenetic level. MMR genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) are often inactivated by somatic mutations on their coding sequences, and also by microRNA downregulation (Cortes-Ciriano et al. 2017). Remarkably, biallelic *MLH1* promoter methylation is primarily the key somatic event responsible for the loss of *MLH1* expression in ~75-80% of sporadic cancers with MSI (Ruiz-Bañobre and Goel 2019).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

MSI is frequently observed in sporadic solid tumours such as colorectal, stomach and endometrial cancers, and has been also reported in glioblastomas, lymphomas and ovarian cancer, among others (Cortes-Ciriano et al. 2017) (Table 12). Remarkably, MSI tumours are generally associated with a better prognosis, and are usually less prone to develop metastasis (Cortes-Ciriano et al. 2017).

Table 12: Frequency of MSI in tumour samples from The Cancer Genome Atlas.

TCGA cohort	Tumour type	N	MSI cases	MSI %
UCEC	Uterine corpus endometrial carcinoma	265	75	28.3 %
STAD	Stomach adenocarcinoma	292	64	21.9 %
COAD	Colon adenocarcinoma	271	45	16.6 %
READ	Rectal adenocarcinoma	76	3; 4*	9.2 %
ACC	Adrenal cortical carcinoma	92	5*	5.4 %
ESCA	Oesophageal carcinoma	183	3; 3*	3.3 %
OV	Ovarian cancer	436	14*	3.2 %
LIHC	Liver hepatocellular carcinoma	375	11*	2.9 %
CESC	Cervical squamous cell carcinoma	305	7*	2.3 %
BRCA	Breast cancer	922	16*	1.7 %
GBM	Glioblastoma multiforme	316	4*	1.3 %
HNSC	Head and neck squamous cell carcinoma	505	6*	1.2 %
LUSC	Lung squamous cell carcinoma	407	5*	1.2 %
KIRC	Kidney renal clear cell carcinoma	377	4*	1.1 %
PAC	Pancreatic cancer	171	2*	1.1 %
BLCA	Urothelial bladder cancer	368	2*	0.8 %
KIRP	Papillary kidney carcinoma	286	2*	0.7 %
LGG	Low grade glioma	514	3*	0.6 %
PRAD	Prostate adenocarcinoma	497	3*	0.6 %
LUAD	Lung adenocarcinoma	482	1*	0.2 %
SKCM	Cutaneous melanoma	109	0*	0 %
PHCA	Pheochromocytoma and paraganglioma	176	0*	0 %
THCA	Thyroid cancer	493	0*	0 %

Abbreviations: TCGA=The Cancer Genome Atlas; MSI=Microsatellite instability; Source: Cortes-Ciriano et al. 2017

It is considered that MSI tumours are generally less aggressive as a consequence of the higher incidence of somatic mutations on their genomes (Cortes-Ciriano et al. 2017). The accumulation of somatic mutations in specific genes can result in the generation of neoantigens and the production of novel immunogenic peptides, which have the potential to trigger a more effective immune response. This increased immunogenicity may contribute to the recognition and targeting of cancer cells by the immune system, leading to a limited tumour progression (Ruiz-Bañobre and Goel 2019).

Several lines of evidence support this cancer regulation model. For example, a classical hallmark of the MSI colorectal tumours is an important lymphocytic infiltration, which correlates with a higher neoantigen load, as well as with a higher expression of various immune checkpoint molecules (PD-L1, CTLA4, LAG3 and IDO) (Ruiz-Bañobre and Goel 2019). In melanoma and non-small cell lung cancer, high mutation rates have been associated with an increased rate of response to anti-CTLA-4 (Snyder et al. 2014) and anti-PD-1 therapies (Rizvi et al. 2015), respectively, probably as a result of a higher neoantigen burden leading to antitumour immune response by CD8+ T cells. Other reports have shown that patients with MMR-deficient colorectal cancer obtain a higher benefit from immune checkpoint blockers

such as nivolumab +/- ipilimumab (anti-PD-1 and anti-CTLA-4 antibodies, respectively) or pembrolizumab (anti-PD-1 antibody) (Diaz et al. 2022; André et al. 2022).

In light of the robust association between MSI and anti-PD-1 blockade benefit in many tumor types, The United States Food and Drug Administration (FDA) approved these immunotherapies for the treatment of MSI or MMR-defective solid tumours, regardless of their tissue of origin (Ruiz-Bañobre and Goel 2019). Therefore, developing efficient strategies for the determination of the MSI status has become a priority from a clinical and scientific point of view.

C3.1.4 - Predictors of the MSI phenotype

Over the last years, PCR-based profiling of informative microsatellite DNA markers and immunohistochemistry of MMR proteins have been the gold standard techniques to determine MSI in clinical and research contexts (Baretti and Le 2018). One of the most common strategies is the profiling of the Bethesda panel, which typically include two mononucleotide (BAT25 and BAT26) and three dinucleotide (D5S346, D2S123 and D17S250) informative microsatellites (Cortes-Ciriano et al. 2017). Assessing this panel was the strategy followed by international research consortia such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, Weinstein, et al. 2013), and it is still a common strategy in daily clinical practice. However, these molecular biology techniques require time and specialized knowledge and equipment to obtain reliable MSI predictions. In addition, their limited profiling of target regions imply that these tests might miss the detection of around 5 - 10% of MSI cases (Baretti and Le 2018).

With the advent of next generation sequencing technologies, whole-genome profiling of MSI cancer genomes has become affordable, allowing for a better understanding of the mutator phenotypes (Cortes-Ciriano et al. 2017). A set of bioinformatic methods have been developed to achieve MSI prediction from whole-genome or whole-exome profiles of microsatellite expansion patterns. These methods include mSINGS (Salipante et al. 2014), MSIsensor (Niu et al. 2014), MOSAIC (Hause et al. 2016), MANTIS (Kautto et al. 2016) and MSIsensor-pro (Jia et al. 2020). Some of these methods, such as MANTIS and MSIsensor, have been validated by health agencies such as the FDA. However, some of these methods often present important limitations, such as the requirement of important computational resources to store and analyze whole-genome data (e.g. BAM files), and the necessity of sequencing matched-normal samples (sometimes not accessible). In addition, these methods assess a large number of marker loci, which makes them inefficient for most standard clinical applications.

Over the last years, novel bioinformatic tools have been designed in order to overcome these limitations (Kather et al. 2019; T. Chen et al. 2019; Danaher et al. 2019; Fu et al. 2019; Saillard et al. 2023). The development of multi-omics and machine learning-based modelling is leading to the development of orthogonal MSI predictors which use additional layers of genomic, transcriptomic, epigenomic and even histologic information. Some methods, such as PreMSIm (L. Li, Feng, and Wang 2020), a K-NN-based classifier, use gene expression patterns to predict MSI from tumour transcriptomes, and others have mined lncRNA profiles for the same purpose (T. Chen et al. 2019). Interestingly, other approaches based on gene expression profiling has also been developed to determine the composition of the tumour microenvironment (e.g. CIBERSORTx) (Newman et al. 2019), providing complementary resources for cancer

immunotherapy response prediction. Additionally, considering the role of epigenetic changes in the origin of MMR alterations and the development of the MSI phenotype (Weisenberger et al. 2006), some methods have been devised. One of them is MIRMMR (Foltz et al. 2017), which integrates DNA methylation and mutation information of MMR pathway genes to predict MSI. Another one is MSI-XGNN (Cao et al. 2023), a graph neural network-based framework which integrates DNA methylation profiling with gene expression in order to generate predictions of the MSI status.

During this PhD thesis, we developed the first MSI predictor exclusively based on DNA methylation profiles (Infinium HumanMethylation 450K array). The development of this predictor not only represents a new resource for MSI prediction in colorectal cancer, but also contributes to the understanding of the epigenetic determinants of MMR tumours and their connection with immune response in cancer. Interestingly, further validation of this approach, might provide a cost-effective resource for the development of MSI testing in liquid biopsy samples.

C3.2 - RESULTS AND DISCUSSION

C3.2.1 - Training and optimization of a methylation-based MSI predictor

We developed a MSI predictor algorithm, assessing the epigenetic status of informative CpG methylation probes across the human genome. First, to find differentially methylated CpG probes between MSI and MSS tumours, we analysed Infinium Human Methylation 450K array data derived from 388 colorectal primary tumours from the pooled TCGA-COAD/READ cohort. We identified 780 probes with an absolute $\Delta\beta > 0.3$ and B-H false discovery rate (FDR)-adjusted p value <0.05, which included valuable CpG predictors for the determination of the MSI status (Supp. Figure 10). As expected, most of these probes (758/780; 97.2%) were hypermethylated, and a minor fraction of them (22/780; 2.8%) was hypomethylated.

We used the set of 780 differentially methylated CpG probes to train 7 alternative supervised machine learning models on the TCGA-COAD/READ dataset (Figure 41). The generalized linear model (GLM) implemented in GLMNET provided the best performance after five repeats of 10-fold cross validation (accuracy=0.98, kappa=0.92). This performance was reached when setting alpha = 0.1 and lambda = 0.0182 as GLMNET tuned parameters. Notably, the GLM performance was significantly better when compared against the second (Wilcoxon test Holm FWER-adjusted p value=0.041) and third (Wilcoxon test Holm FWER-adjusted p value=0.017) best alternative models, leading to the selection of the algorithm.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

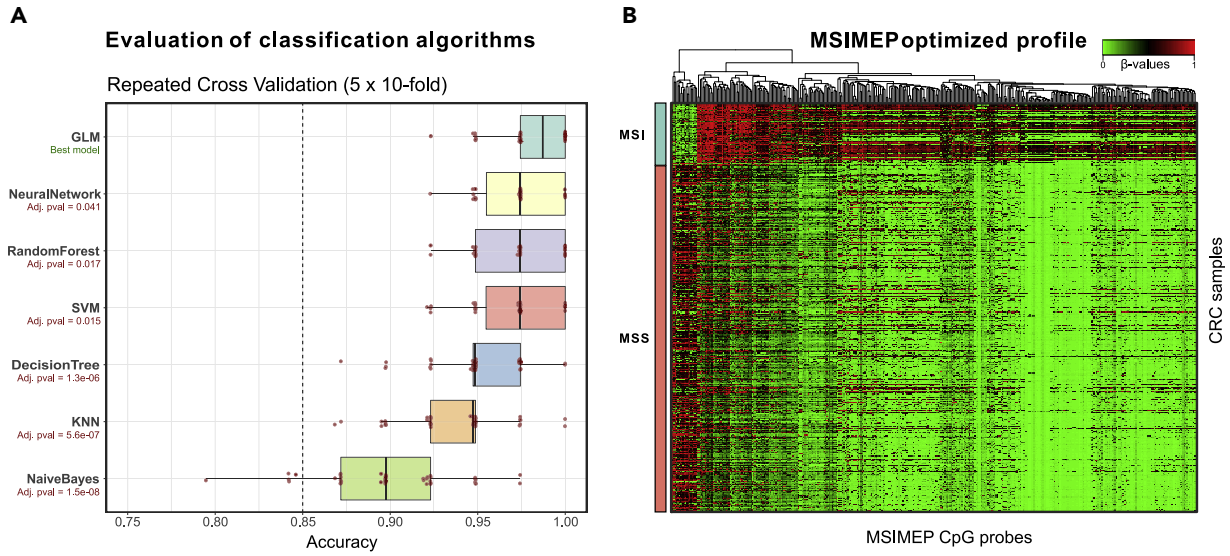


Figure 41: Development of the MSIMEP classifier.
Development of the MSIMEP classifier. A) Evaluation of alternative algorithms for MSI/MSS phenotype prediction from array methylation tumour profiles. Model performance was assessed through 5 rounds of 10-fold cross-validation on the TCGA-COAD/READ dataset, and an optimal algorithm (GLM) was selected based on significant improvement in classification accuracy. Alternative algorithms (Neural Network, Random Forest, SVM, Decision Tree, KNN and Naive Bayes) are displayed sorted by average accuracy, with Holm FWER-adjusted p values showing statistical differences against GLM. B) Heatmap for the methylation tumour profile (β-values) of TCGA COAD/READ patients for the CpG probes incorporated into the MSIMEP optimized model. CpG probes (x axis) are displayed following hierarchical clustering criteria, and TCGA COAD/READ patients (y axis) are grouped by MSI/MSS phenotype (green=MSI, red=MSS).

Once the best algorithm was selected, we performed a recursive feature elimination (RFE) to optimize our predictive model (Supp. Figure 11). The optimized model, hereafter MSIMEP, is composed of 400 CpG predictors (Figure 41, Table 13), which rendered the following classification performance: accuracy = 0.98, kappa = 0.92, area under the receiver operating characteristic curve (AUROC) = 0.98, sensitivity = 0.88, and specificity = 1.00. Notably, those cases incorrectly classified by our model presented a MANTIS score statistically closer to 0.4 (the average cutoff established to differentiate MSI from MSS tumours) (Kautto et al. 2016) than those correctly classified (MSS vs failed, Wilcoxon test p value < 0.001; failed vs MSI, Wilcoxon test p value = 0.009) (Supp. Figure 12) . Importantly, MSIMEP performance was consistent across different clinical and molecular subgroups (Supp. Figure 13, Supp. Figure 14).

Table 13: MSIMEP optimized model performance in different cohorts.

Cohort	Cancer type	Phase	N	MSI %	MSS %	ACC	KAP	AUC	SENS	SPEC	PREC
COAD/READ	CRC	Training	387	15	85	0.98	0.92	0.98	0.88	1.00	0.99
External I	CRC	Validation	79	14	86	0.97	0.90	0.99	1.00	0.97	0.85
External II	CRC	Validation	262	15	85	0.95	0.83	0.95	0.88	0.97	0.83
External III	CRC	Validation	81	11	89	0.86	0.52	0.86	0.89	0.86	0.44
External IV	CRC	Validation	95	100	NA	NA	NA	NA	0.86	NA	1.00
Ext. Pool	CRC	Validation	517	30	70	0.93	0.83	0.93	0.88	0.95	0.88
TCGA-STAD	STAD	Exploratory	394	18	82	0.84	0.56	0.94	0.90	0.82	0.52
TCGA-UCEC	UCEC	Exploratory	420	31	69	0.88	0.70	0.92	0.72	0.95	0.86

Performance metrics: ACC=accuracy; KAP=kappa; AUC=area under the ROC curve; SENS=sensitivity; SPEC=specificity; PREC=precision.

USC DE SANTIAGO DE COMPOSTELA

Additionally, considering a potential applicability to liquid biopsy samples and assuming a limited sacrifice of classification performance (1% decrease in accuracy), we developed a reduced version of MSIMEP composed of 25 CpG probes **(**Table 14**,** Supp. Figure 11).

MSIMEP was specifically developed to determine the MSI status in the context of colorectal cancer. However, we were also interested to test its performance in other tumour types with high prevalence of MSI, such as gastric and endometrial cancers (Cortes-Ciriano et al. 2017). For this purpose, we interrogated Infinium Human Methylation 450K array data from the TCGA-Stomach Adenocarcinoma (STAD) and TCGA-Uterine Corpus Endometrial Carcinoma (UCEC) cohorts (Supp. Figure 15). In STAD, MSIMEP yielded accuracy = 0.84, kappa = 0.56, AUROC = 0.94, sensitivity = 0.9, specificity = 0.82, and precision = 0.52 under the optimized model, and accuracy = 0.75, kappa = 0.43, AUROC = 0.93, sensitivity = 0.94, specificity = 0.71, and precision = 0.41 under the reduced model. In UCEC, MSIMEP yielded accuracy = 0.88, kappa = 0.70, AUROC = 0.92, sensitivity = 0.72, specificity = 0.95, and precision = 0.86 under the optimized model, and accuracy = 0.89, kappa = 0.73, AUROC = 0.93, sensitivity = 0.82, specificity = 0.91, and precision = 0.81 under the reduced model. Overall, these results indicated a reduction in MSI classification performance in STAD and UCEC compared with CRC.

C3.2.2 - Biological features of the MSIMEP classifier

The MSIMEP composition relies on a set of 400 CpG probes distributed across 22 human chromosomes (Figure 42). Chromosomes 3, 1 and 6 contain the largest number of MSIMEP probes, with counts of 41, 34 and 28 CpGs, respectively. The highest probe density is associated with chromosome 19, with 21 CpGs equivalent to 0.37 CpGs/Mb. In general, MSIMEP probes tend to be located in high GC content regions (88%). This likely reflects not only the composition of the Illumina 450K Infinium Array, but also the CpG island methylator phenotype (CIMP) typically displayed by CRC tumours with MSI (Weisenberger et al. 2006).

We further explored the genomic distribution of MSIMEP CpG probes. We observed that 63 MSIMEP probes (15.8%) are located at intergenic regions, whereas 337 probes (84.3%) are associated with coding genes. Notably, 34 MSIMEP probes (8.5%) are associated with COSMIC human cancer genes (Sondka et al. 2018), including 19 probes associated with genes involved in cancer syndromes such as *MLH1*, *EXT1*, *WRN* and *AXIN2*. Remarkably, *MLH1* is one of the key genes involved in the DNA mismatch repair process, and its epigenetic inactivation is a primary source event responsible for the majority of sporadic MSI CRCs (Ruiz-Bañobre and Goel 2021). Notably, the top five relevant genes for the MSIMEP model are *MLH1*, *TRIP10*, *C1orf95, EXT1* and *GPR160*, respectively (Figure 42).

Considering their connection with gene structure, 219 MSIMEP CpG probes (54.8%) were found associated with the proximal promoter region of human genes, including 12 probes associated with the *MLH1* promoter. One hundred and nine probes (26.3%) were found located within gene bodies, and 13 probes (3.3%) were found located within 3'UTR regions. Interestingly, MSI-related hypermethylated probes were predominantly located on promoter regions and in CpG islands (Fisher exact test p value < 0.001, hypermethylated vs hypomethylated), whereas MSI-related hypomethylated probes were preferentially located on gene bodies and in open sea regions, respectively (Fisher exact test p value <0.001,

hypermethylated vs hypomethylated) (Figure 42). In addition, when evaluating the composition of MSIMEP from a functional perspective, 250 CpG probes (62%) were found associated with regulatory elements. Compared with the Infinium Human Methylation 450K array, MSIMEP comprises a higher proportion of regulatory regions at both genic (χ^2 test p value = 0.003) and intergenic (χ^2 test p value = 0.012) regions, which likely evidence an overall overrepresentation of relevant CpG positions in the MSI predictor.

The composition of the MSIMEP reduced model contains 25 CpG predictors (Table 14**,** Supp. Figure 10). It includes 4 intergenic probes and 21 CpG probes associated with genes. Notably, 9 of these probes are attributed to relevant tumour suppressor genes (*MLH1*, *EXT1* and *WNK2*), including 6 probes associated with the MLH1 gene. Of note, the probe cg15048832 from the gene *TSC22D2* was also integrated into the model, a selection shared with another recently-developed MSI predictor (Cao et al. 2023)

Table 14: Description of the 25 CpG probes of the MSIMEP reduced model.

CpG probe	Chromosome	Position	Strand	Gene	Gene region	Methylation status (MSI vs MSS)
cg15592945	1	226736711	F	*C1orf95*	1st Exon	Hypermethylated
cg03745431	1	226736713	F	*C1orf95*	1st Exon	Hypermethylated
cg17621259	3	37035168	F	*MLH1*	TSS200	Hypermethylated
cg14671526	3	37035200	R	*MLH1*	TSS200	Hypermethylated
cg27331401	3	37035207	F	*MLH1*	TSS200	Hypermethylated
cg06590608	3	37035228	F	*MLH1*	TSS200	Hypermethylated
cg11224603	3	37035282	R	*MLH1*	1st Exon	Hypermethylated
cg14598950	3	37035355	F	*MLH1*	1st Exon	Hypermethylated
cg04563996	3	57093519	R	*ARHGEF3*	5'UTR	Hypomethylated
cg15048832	3	150130775	R	*TSC22D2*	Body	Hypermethylated
cg12350863	3	169758289	F	*GPR160*	5'UTR	Hypermethylated
cg20288341	6	31683131	F	*LY6G6E*	TSS1500	Hypermethylated
cg06226516	7	13072314	R	NA	IGR	Hypermethylated
cg02200207	7	29605237	R	*PRR15*	5'UTR	Hypermethylated
cg05313153	8	119122430	F	*EXT1*	1st Exon	Hypermethylated
cg21602557	8	119122878	R	*EXT1*	1st Exon	Hypermethylated
cg13563298	9	95948059	R	*WNK2*	Body	Hypermethylated
cg27240158	11	12698019	R	*TEAD1*	5'UTR	Hypermethylated
cg19524009	13	52735075	R	*NEK3*	TSS1500	Hypermethylated
cg06223834	16	4103161	R	*ADCY9*	Body	Hypermethylated
cg09294739	16	55218782	R	NA	IGR	Hypomethylated
cg11582717	17	63452460	R	NA	IGR	Hypermethylated
cg13860006	19	6741178	R	*TRIP10*	Body	Hypermethylated
cg06410591	19	6741181	R	*TRIP10*	Body	Hypermethylated
cg05232694	20	48809539	R	NA	IGR	Hypermethylated

Note: methylation status is described based on a comparison of MSI vs MSS tumours from TCGA COAD/READ.

Finally, to fully characterize the biological basis of MSIMEP, we performed a gene set enrichment analysis with a methylation-specific approach (Ren and Kuan 2019). As expected, when evaluated at promoter level, the analysis showed a significant enrichment of pathways involved in mismatch repair (B-H FDR-adjusted p value < 0.001). When evaluated either all CpG probes or those located at body gene level, mismatch repair (B-H FDR-adjusted p value < 0.001) and pathways in cancer (B-H FDR-adjusted p value < 0.001) appear among the most significant pathways, confirming again the biological connection of MSIMEP with MSI

USC DE COMPOSTELA

phenotype in cancer. Other significantly enriched pathways are described in Supp. Figure 16. Moreover, we used a computational immune cell deconvolution strategy (Newman et al. 2019) to confirm that MSIMEP-predicted MSI CRCs present the same immune cell context than those MSI cases originally classified through PCR-based profiling in the TCGA COAD/READ cohort (Supp. Figure 17).

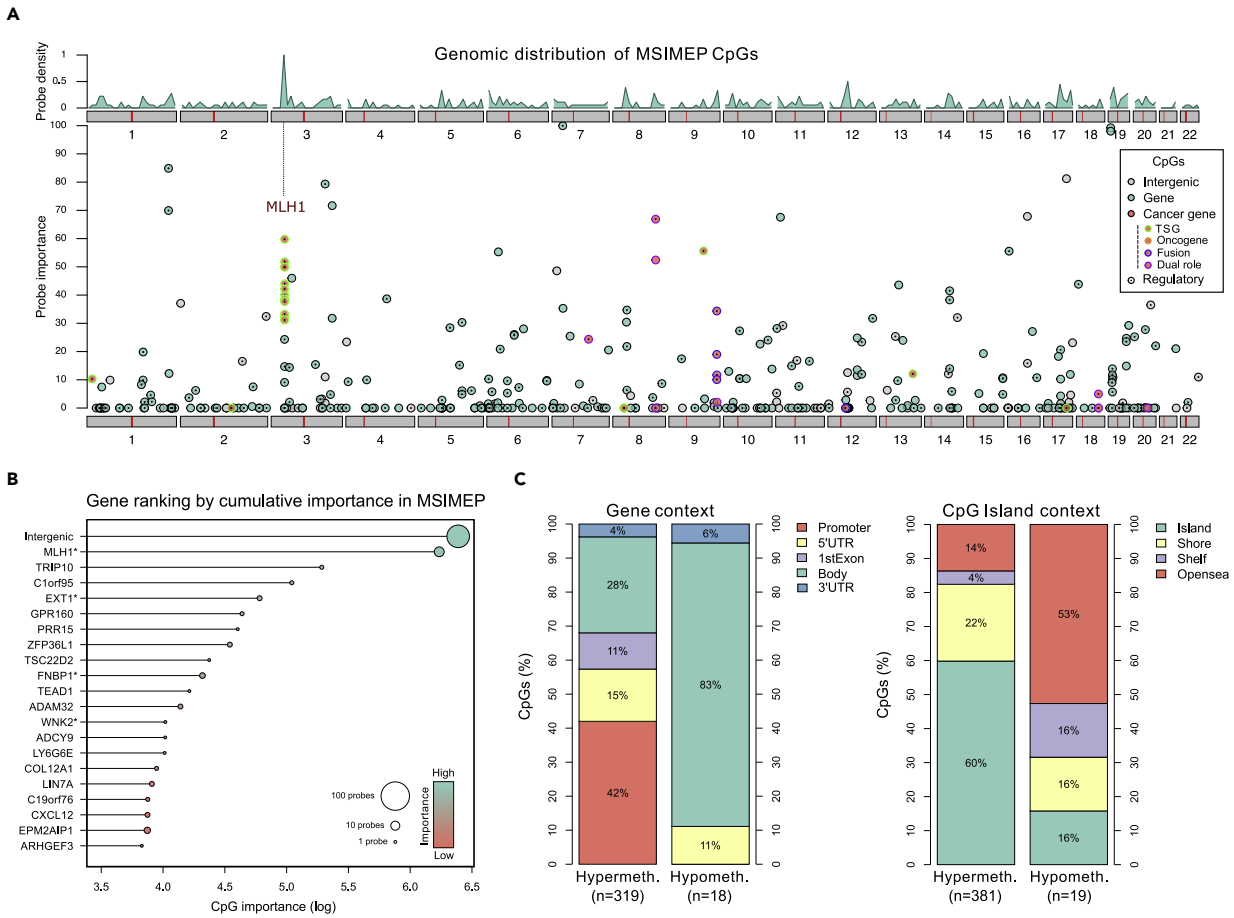


Figure 42: Features of the MSIMEP model.
A) Genomic distribution of MSIMEP CpG probes. The x axis shows the CpG coordinates across the human autosomes (chromosomes 1–22). The y axis shows relative density of MSIMEP CpG probes (top), and the relative importance of each CpG for the GLM classification model (bottom). Probes are classified as genic (green), intergenic (grey), or cancer-related (red). Cancer-related probes are further classified according to the role/s in cancer (tumour suppressor gene, oncogene, fusion gene, or dual role) of the associated gene. Probes associated with regulatory elements are highlighted with a central dot. Location of the MLH1 gene is shown. B) Top 20 genes profiled by MSIMEP, ranked by the cumulative importance for the GLM model of the associated CpG probes. Dot size reflects the number of probes associated with each gene. Cancer genes are highlighted with an asterisk. C) Comparison of the distribution of hypermethylated and hypomethylated MSIMEP CpG probes (MSI vs MSS tumours from TCGA COAD/READ cohort) according to gene and CpG island context regions.

C3.2.3 - External validation of MSIMEP classifier

To confirm the accuracy of MSIMEP models in determining MSI status from methylation profiles in CRC samples, we evaluated its performance in four independent CRC cohorts.

First, we evaluated MSIMEP in the external cohort I, which consisted of 79 CRCs, each with three alternative tumour regions per case. In this cohort, MSIMEP performance was remarkable

when tested in either invasive front (accuracy = 0.99, kappa = 0.95, AUROC = 1.00, sensitivity = 1.00, specificity = 0.99, and precision = 0.92), or in luminal (accuracy = 0.99, kappa = 0.95, AUROC = 0.99, sensitivity = 1.00, specificity = 0.99, and precision = 0.92) or centre (accuracy = 0.99, kappa = 0.95, AUROC = 0.99, sensitivity = 1.00, specificity = 0.99, and precision = 0.92) regions (Table 13, Figure 43). Notably, the concordance of predictions among the three different tumour regions was extremely high (Cochran's Q test p value = 0.607, no statistically significant differences found in predictions by tumour region) (Figure 43).

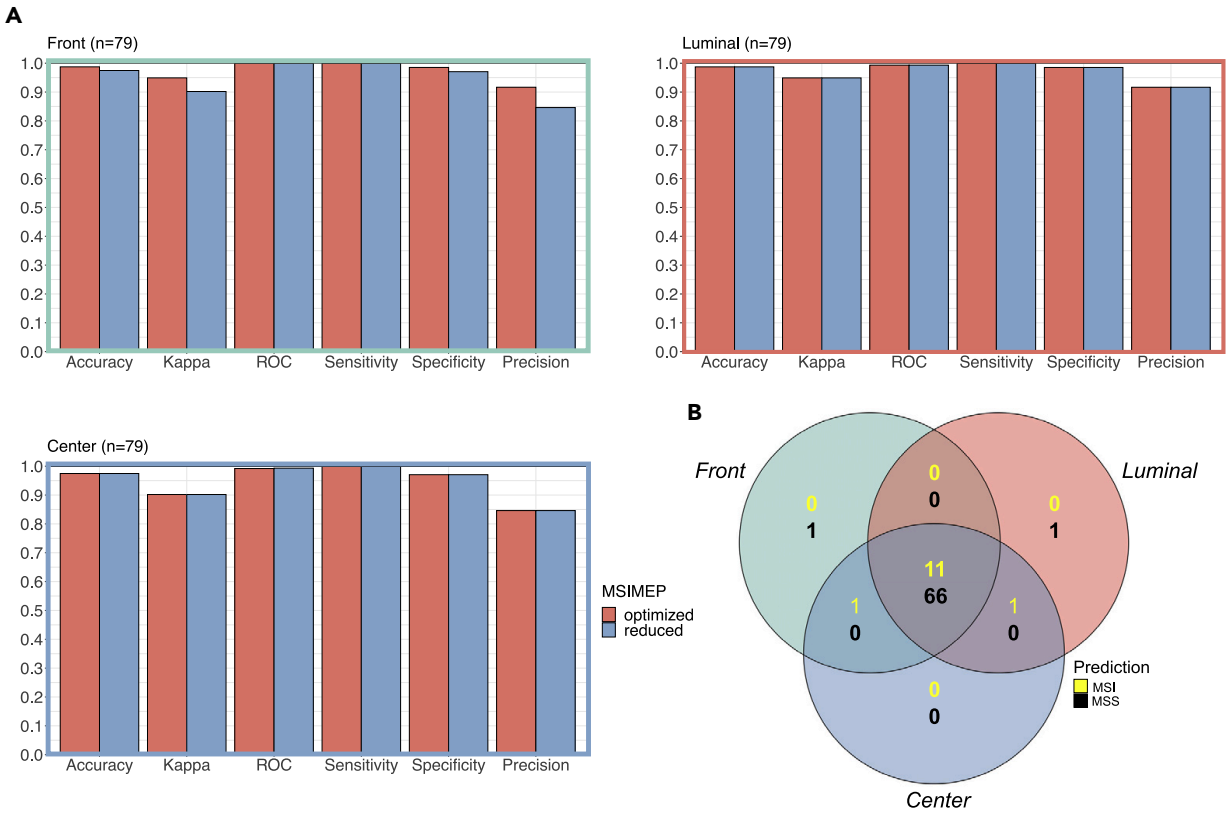


Figure 43: MSIMEP evaluation in CRC external cohort I.
A) Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity and precision) for the evaluation of MSIMEP classification capacity of MSI/MSS patients across different CRC regions (invasive front, luminal and centre). Models: MSIMEP optimized (red) and MSIMEP reduced (blue). B) Venn diagram showing the concordance in MSI/MSS prediction across different CRC regions from external cohort I patients. Cohort: external I (n=79, 14% MSI).

Next, we evaluated MSIMEP in cohorts II and III, composed of 262 and 81 CRC cases, respectively (Table 13, Figure 44). MSIMEP showed consistent results, especially in cohort II, where the performance of the classifier remained high (accuracy = 0.95, kappa = 0.83, AUROC = 0.95, sensitivity = 0.88, specificity = 0.97, and precision = 0.83). In cohort III, MSIMEP experienced a decrease in performance, but remained informative (accuracy = 0.86, kappa = 0.52, AUROC = 0.86, sensitivity = 0.89, specificity = 0.86, and precision = 0.44) This decrease in performance might be partially attributed to the higher proportion of serrated CRC cases present in this cohort (Conesa-Zamora et al. 2015).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Moreover, we assessed the predictive potential of MSIMEP in cohort IV, composed of 95 MSI-H CRC cases. Even when this might represent a challenging cohort due to the lack of MSS cases, MSIMEP still rendered a sensitivity of 0.86 (Table 13, Figure 44).

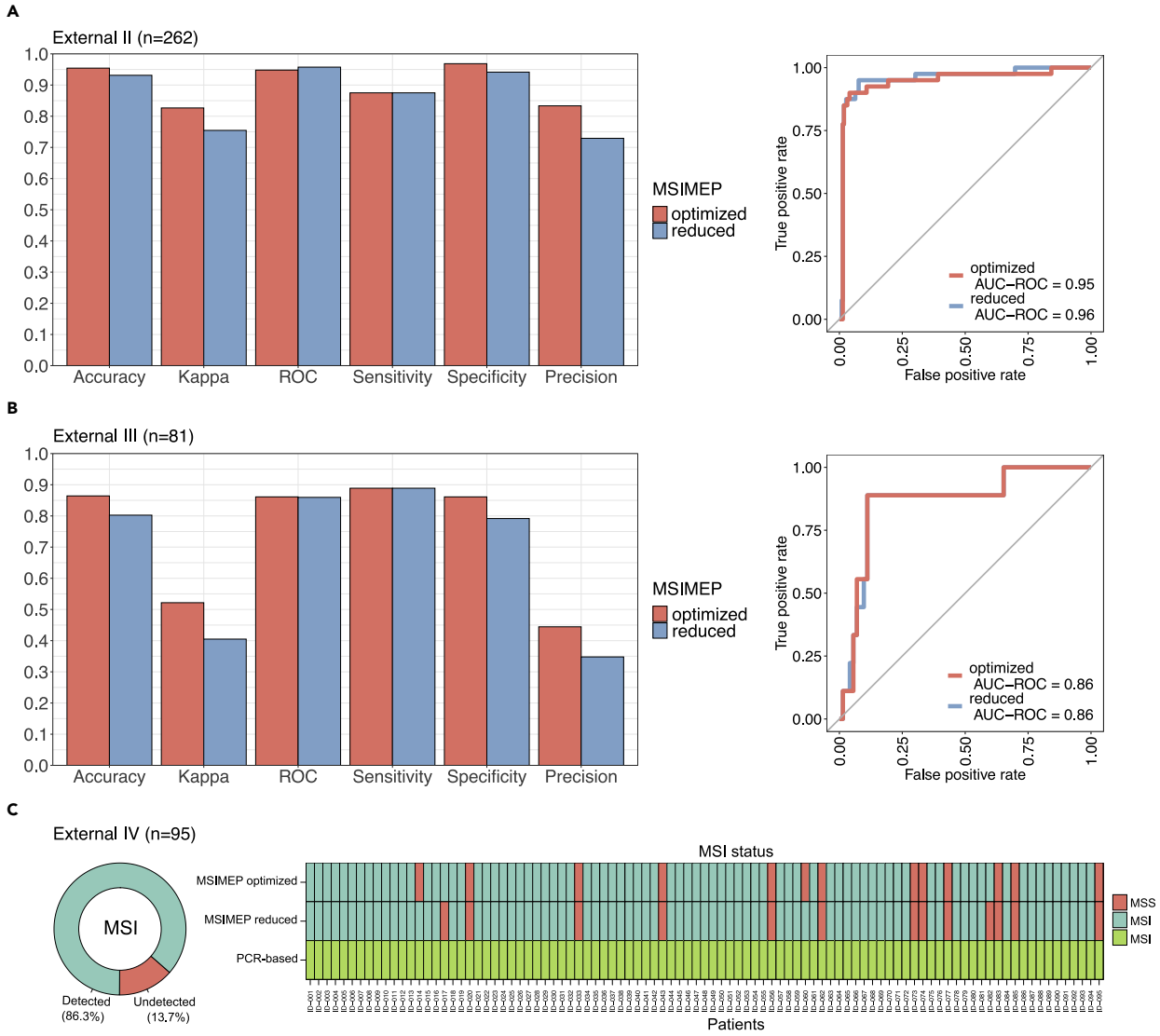


Figure 44: MSIMEP evaluation in CRC external cohorts II-IV.
Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity and precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients. ROC curves showing MSIMEP classification performance at alternative predictive thresholds. Models: MSIMEP optimized (red) and MSIMEP reduced (blue). Circle chart showing the fraction of detected cases in a cohort of MSI patients (ID-001 to ID-095). A) External cohort II (n=262, 15% MSI), B) External cohort III (n=81, 11% MSI) and C) External cohort IV (n=95, 100% MSI).

To obtain a general overview of MSIMEP performance on the external datasets, we evaluated a pooled CRC cohort (external cohorts I to IV, n=517 cases). In the pooled CRC cohort, MSIMEP optimized exhibited a high classification performance (accuracy = 0.93, kappa = 0.83, AUROC = 0.93, sensitivity = 0.88, specificity = 0.95, and precision = 0.88) (Table 13). MSIMEP reduced model version showed a slight decrease in performance metrics, but remained effective (accuracy = 0.91, kappa = 0.78, AUROC = 0.93, sensitivity = 0.88, specificity = 0.92, and precision = 0.82) (Table 15). Notably, MSIMEP performance was balanced across different clinical and molecular subgroups (Supp. Figure 13). Altogether, these results suggest that MSIMEP models can be applied to predict MSI status from microarray

USC DE COMPOSTELA

DNA methylation profiles from both fresh-frozen FF and formalin-fixed paraffin-embedded (FFPE) tumour CRC samples.

Table 15: MSIMEP reduced model performance in different cohorts.

Cohort	Cancer type	Phase	N	MSI (%)	MSS (%)	ACC	KAP	AUC	SENS	SPEC	PREC
COAD/READ	CRC	Training	387	15	85	0.98	0.92	0.99	0.86	1.00	0.95
External I	CRC	Validation	79	14	86	0.97	0.90	0.99	1.00	0.97	0.85
External II	CRC	Validation	262	15	85	0.93	0.75	0.96	0.88	0.94	0.73
External III	CRC	Validation	81	11	89	0.80	0.40	0.86	0.89	0.79	0.35
External IV	CRC	Validation	95	100	NA	0.86	NA	NA	0.86	NA	1.00
Ext. Pool	CRC	Validation	517	30	70	0.91	0.78	0.93	0.88	0.92	0.82
TCGA-STAD	STAD	Exploratory	394	18	82	0.75	0.43	0.93	0.94	0.71	0.41
TCGA-UCEC	UCEC	Exploratory	420	31	69	0.89	0.73	0.93	0.82	0.91	0.81

Performance metrics: ACC=accuracy; KAP=kappa; AUC=area under the ROC curve; SENS=sensitivity; SPEC=specificity; PREC=precision.

C3.2.4 - Comparison of MSIMEP with a MLH1 promoter methylation-based model

Considering the role of biallelic *MLH1* promoter methylation as source of microsatellite instability in most of the sporadic CRCs with this phenotype (Ruiz-Bañobre and Goel 2021), we compared the performance of our MSIMEP models with a MLH1 promoter methylation-based model in a pooled CRC cohort (External cohorts I to IV). As expected, MSIMEP optimized and reduced models provided a better sensitivity (0.88 for booth MSIMEP models vs 0.71 for *MLH1* promoter methylation-based model) at the expense of a slightly lower specificity than the *MLH1* methylation-based model (0.95 and 0.92 for MSIMEP optimized and reduced models, respectively, vs 0.96 for *MLH1* promoter methylation-based model) (Figure 45). As expected, stratified analysis by *MLH1* promoter methylation status confirmed the increased sensitivity of MSIMEP models to detect those non-*MLH1* dependent cases (Supp. Figure 20).

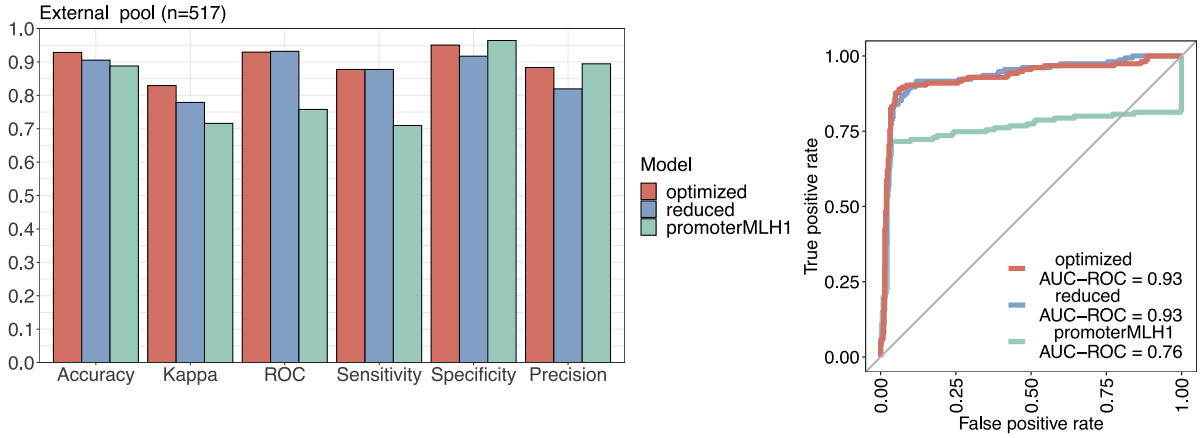


Figure 45: Comparison of the performance of three GLM models in a pooled CRC cohort. (External cohorts I to IV). Barplots showing complementary metrics (accuracy, kappa, AUROC, sensitivity, specificity and precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients. ROC curves showing MSIMEP classification performance at alternative predictive thresholds. Models: MSIMEP optimized (red), MSIMEP reduced (blue), and promoter MLH1 (green). Cohort: external pool (n=517, 30% MSI).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

C3.3 – CONCLUSIONS OF CHAPTER III

I. We have developed MSIMEP, an accurate computational tool for predicting the MSI status from microarray DNA methylation profiles of CRC samples. The GLM-based classifier models that MSIMEP implements were extensively validated in different CRC cohorts, showing potential to be computationally applied to a wide range of CRC cases without a significant decrease in performance.

II. MSIMEP models incorporate relevant CpG predictors from both genic and intergenic genomic locations. A significant proportion of MSIMEP predictors is accounted of *MLH1* probes, reflecting the role of the methylation of its promoter as the primary somatic event responsible for the majority of sporadic MSI-H CRCs. However, MSIMEP incorporates additional genic and intergenic CpG sites that are relevant to trace those non-*MLH1*-dependent MSI cases.

III. MSIMEP still presents important limitations: (1) it should be formally adapted before being applied to other platforms aside from Infinium Human Methylation 450K arrays; (2) Its performance in non-CRC tumour types requires further investigation; (3) a centralized re-evaluation of the PCR-based MSI status across CRC datasets might be convenient for an improved MSIMEP evaluation; (4) The potential clinical value of MSIMEP reduced model needs to be tested in circulating tumour DNA samples.

C3.4 - PUBLICATIONS

The contents of this chapter have been published as an article in the journal iScience [1]. This publication represents a collaborative effort to advance our understanding of the sources of microsatellite instability in colorectal cancer. My contributions to this work were focused on the development and validation of a computational predictor of microsatellite instability (MSIMEP). I performed differential methylation analyses using microarray TCGA tumour datasets, followed by variable selection to optimize MSI prediction in colorectal cancer. I tuned and evaluated alternative algorithms for MSI prediction in colorectal cancer. I investigated the biological features of MSIMEP, conducting pathway enrichment analyses and determining the genomic context of critical CpG sites within the classification model. I assessed the performance of MSIMEP across multiple colorectal cancer cohorts and tested its applicability to other tumour types. I compared MSIMEP with a *MLH1* methylation-based model and characterized the immune cell context of MSIMEP predictions. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work. The contents of the publication have been adapted to align with the format of this thesis. Further detail regarding author's contributions is provided in the appendix, alongside the full list of authors and their affiliations.

References:

1. Santamarina-García M, Brea-Iglesias J, Bramsen JB, et al. MSIMEP: Predicting microsatellite instability from microarray DNA methylation tumor profiles. *iScience*. 2023;26(3):106127. Published 2023 Feb 3. doi:10.1016/j.isci.2023.106127

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CONCLUSIONS

Concluding remarks

This work has presented key findings from my research within the Molecular Medicine PhD program, focusing on genomic instability in cancer. Collectively, these results contribute modestly to a deeper understanding of the molecular mechanisms driving genomic instability in cancer. Our findings demonstrate the potential of emerging high-throughput technologies, alongside advanced bioinformatic approaches, to identify and characterize the genetic and epigenetic elements involved in genomic structural variation in cancer. The analysis of genetic and epigenetic variants, as well as alterations distributed across large-scale chromosome territories, has been valuable in providing insights into the patterns of somatic evolution in tumours. Although these results fall within the framework of basic research in cancer genomics, further development of these research lines in primary, metastatic, and transmissible cancers may contribute to novel therapeutic advances in the fields of personalized medicine and molecular oncology.

Conclusion from Chapter I

I. Long-read sequencing technologies implemented by PacBio and ONT enable a comprehensive characterization of L1 source elements and their derived insertions. This advanced characterization allows to determine some relevant aspects of L1 biology, such as the integrity of their ORFs, their allelic variation and their diversification patterns, at unprecedented level of resolution.

II. Source Inference is a reliable bioinformatic method to trace L1 somatic retrotranspositions to their element-of-origin based on internal diagnostic nucleotides. While insertion length limits the sensitivity of the approach, our method successfully traces a significant fraction of somatic solo-L1 insertions back to their original source elements in tumours.

III. L1 activity in human cancer is driven by a diverse set of source elements belonging to multiple L1Hs sublineages. Notably, while some of these source elements predominantly generate solo-L1 derived insertions, others preferentially generate transductions, likely influenced by local sequence variation in their termination signals.

Conclusion from Chapter II

I. The genome of the bivalve transmissible neoplasia of *Cerastoderma edule* has a notable burden of somatic structural variation. Deletions are the most common type of somatic structural alterations, followed by duplications and potential interchromosomal rearrangements. These variants are widely distributed across all the transmissible CedBTN chromosomes.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

I. Two divergent lineages of the bivalve transmissible neoplasia coexist in the marine environment: CedBTN1 and CedBTN2, as supported by complementary genetic variants (SNVs and SVs). Their large fraction of private candidate somatic variants, and their divergence against extant cockle populations, suggest that both lineages might be extremely old.

II. Oncogene amplification (*MDM2*, *CCND3*, and *MYC*) and tumour suppressor loss (*MGMT*) might be among the driver genetic alterations driving CedBTN1 and CedBTN2 early evolution. The complex copy number alterations profiles found in CedBTN tumours, including whole genome doubling, avoid conclusive findings about other specific driver alterations.

Conclusion from Chapter III

I. We have developed MSIMEP, an accurate computational tool for predicting the MSI status from microarray DNA methylation profiles of CRC samples. The GLM-based classifier models that MSIMEP implements were extensively validated in different CRC cohorts, showing potential to be computationally applied to a wide range of CRC cases without a significant decrease in performance.

II. MSIMEP models incorporate relevant CpG predictors from both genic and intergenic genomic locations. A significant proportion of MSIMEP predictors is accounted of *MLH1* probes, reflecting the role of the methylation of its promoter as the primary somatic event responsible for the majority of sporadic MSI-H CRCs. However, MSIMEP incorporates additional genic and intergenic CpG sites that are relevant to trace those non-*MLH1*-dependent MSI cases.

III. MSIMEP still presents important limitations: (1) it should be formally adapted before being applied to other platforms aside from Infinium Human Methylation 450K arrays. (2) Its performance in non-CRC tumour types requires further investigation; (3) a centralized re-evaluation of the PCR-based MSI status across CRC datasets might be convenient for an improved MSIMEP evaluation; (4) The potential clinical value of MSIMEP reduced model needs to be tested in circulating tumour DNA samples.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

METHODOLOGY

METHODOLOGY OF CHAPTER I

HGSVC2 cohort

We initially selected 34 unrelated individual genomes for de novo sequencing, aiming to have at least one representative from each of the 26 1000GP populations, of which 30 samples passed the initial quality control (QC). Additionally, we sequenced three previously studied child samples, completing three parent-child trios, and included publicly available sequencing data for two samples generated as part of the Genome in a Bottle effort. The final set of 35 genomes consists of 19 females and 16 males from various ancestries, including African (AFR, n=11), Admixed American (AMR, n=5), East Asian (EAS, n=7), European (EUR, n=7), and South Asian (SAS, n=5) descent. All genomes were sequenced using continuous long-read (CLR) sequencing (n=30) with more than 40-fold coverage, or high-fidelity (HiFi) sequencing (n=12) with more than 20-fold coverage. Strand-seq data was generated for each of the samples. This data was used to produce 64 unrelated phased and assembled human haplotypes using a reference-free assembly approach in the absence of parent-child trio information.

Note: this work was conducted by researchers and technicians from the HGSCV2 consortium.

Mobile element insertion calling and MEIGA-PAV annotation

Based on the 64 unrelated phased genome assemblies, a collection of sequence-resolved non-reference mobile element insertions (MEIs) was identified (n=9453), including 7738 Alu elements, 1175 L1Hs, and 540 SVAs. MEIs were primarily discovered using PAV, followed by MEI annotation using MEIGA-PAV. Phased assembly structural variant (SV) calls, in the form of a multisample VCF, were processed using a custom pipeline derived from the MEIGA algorithm to detect non-reference L1 insertions. First, candidate retrotransposition events were identified by searching for poly-A/T tails at the 5’ and 3’ ends of assembled inserted sequences. These tails were required to be at least 10 bp in length, have a minimum purity of 80%, and be within 30 bp of the insert end. Then, the inserted sequences for candidates were realigned using minimap2 (H. Li 2018) v2.10 against a database that included L1 consensus sequences. Sequence alignments were chained based on complementarity to identify the minimum set of non-overlapping alignments that maximized the resolution of the inserted sequence. Based on the alignment chain, MEIs were called, and multiple insertion features were inferred, including orientation, insertion length, 5' inversion, and truncation lengths. Unresolved sequence 3' ends were interrogated for poly-A/T tracts using the same criteria described above. MEI calls with a single tract were classified as "solo" insertions, while events with multiple tracts were considered potential 3' transduction events, with the transduced sequences located between the tracts.

Note: this work was conducted by Bernardo Rodriguez-Martín (EMBL) and other researchers from the HGSCV2 consortium.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Phylogenetic analysis and age estimation for active FL-L1s

DNA sequences for 142 human-specific (L1Hs) and L1Pt [GenBank: KF661301.1] full-length L1 elements were aligned using MUSCLE (Edgar 2004) v3.8 with default number of iterations. Manual inspection of the multiple alignment was performed with Jalview (Waterhouse et al. 2009) v2.11 in order to remove short upstream and downstream spurious sequences. L1 phylogenies were built using the following R packages: Ape (Paradis, Claude, and Strimmer 2004) v5.4 and Phangorn (Schliep 2011) v2.5. Nucleotide substitution model selection was performed with phangorn::modelTest. GTR+G+I was set as the best model for tree inference according to both AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) scores. Maximum likelihood tree was estimated with phangorn::optim.pml function, starting from an initial distance-based tree. Tree topology was adjusted via nearest neighbor interchange (Nguyen et al. 2015). After maximum likelihood optimization, a midpoint rooting strategy was followed, rendering L1Pt as the expected tree output. Non-parametric bootstrap analysis with 1000 replicates was done to determine the node consistency along the inferred tree. For display purposes, only values informative of high support (≥80%) were labeled to the tree nodes.

The average age of L1 elements could be estimated by measuring the level of sequence divergence from each subfamily consensus sequence (Marchani et al. 2009; Salem et al. 2003). We used DECIPHER v2.14 to generate a consensus sequence for each L1Hs subfamily (*pre-Ta*, *Ta-0* and *Ta-1*) from our set of active elements. For each element, we identified the number of nucleotide substitutions acquired from the ancestral state represented by this consensus. Using a L1-specific mutation rate of 0.25% per million years (Boissinot, Chevret, and Furano 2000), and assuming constant and neutral rate of evolution, we obtained rough estimates about the age of individual L1 elements.

Note: this research work was performed by Martín Santamarina (Universidade de Santiago de Compostela)

Subfamily assignation and ORF status determination for active FL-L1s

For L1 events, subfamily assignment was performed through the identification of subfamily diagnostic nucleotide positions on their 3' end (Boissinot, Chevret, and Furano 2000). L1 integrations bearing the diagnostic "ACG" or "ACA" triplet at 5929-5931 position were classified as "*pre-Ta*" and "Ta", respectively. Ta elements were subclassified into "*Ta-0*" or "*Ta-1*" according to diagnostic bases at 5535 and 5538 positions (*Ta-0*: G and C; *Ta-1*: T and G). Elements that did not display any of these diagnostic profiles could not be assigned to a particular category, and their subfamily status remained undetermined. The coding potential of each L1 sequence was assessed in every of the six potential reading frames. Codon equivalencies followed the Standard Genetic Code currently adopted by NCBI, considering AUG, UUG and CUG as starting codons, and UAA, UAG and UGA as terminators. Every candidate open reading frame (ORF) identified through this strategy was compared with reference L1 ORF1 and ORF2 in terms of 1) length and 2) sequence identity, before being cataloged as one of the two main components of the L1 coding system. Reference amino acid sequences encoded by L1Hs ORF1 and ORF2 proteins were obtained from Uniprot (Apweiler et al. 2004). Finally, a multi-sample VCF containing MEI calls together with all collected pieces of annotation was generated as output.

USC
UNIVERSIDADE DE COMPOSTELA

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

MEIGA cohort

Primary tumours were obtained from donors with three cancer types (HNSC, LUSC, and COAD), which are known to display some of the highest rates of somatic retrotransposition in human cancer (Rodriguez-Martin et al. 2020). An initial screening was conducted on 150 primary tumours (50 HNSC, 50 LUSC, and 50 COAD), from which samples with high retrotransposition rates (≥100 candidate somatic L1 events) were selected. This screening was performed using shallow-Illumina paired-end sequencing (average coverage 11.7X), with xTea (Chu et al. 2021) used to identify retrotransposition events. This analysis led to the identification of 10 tumours (PD0266a, PD0270a, PD0274a, PD0277a, PD0287a, PD0307a, PD0312a, PD0319a, PD0331a, and PD0351a) which were selected for high-coverage sequencing analysis with long-reads. All the samples for the MEIGA project were provided by the Basque Biobank, part of the Spanish Biobank Network. Approval for the study was granted by a Clinical Research Ethics Committee dependent on the Galician Ministry of Health.

Note: This work was conducted by researchers and technicians from the MEIGA project.

DNA isolation and library preparation

Genomic DNA extraction was performed on fresh-frozen tissues from primary tumors and their matched-normal samples using the AllPrep DNA/RNA Mini Kit (Qiagen, California, USA). During long-read library preparation, short DNA fragments (less than 5 Kb) were removed with the Short Read Eliminator XS buffer (Circulomics, Maryland, USA), and multiple purification steps were carried out using Agencourt AMPure XP magnetic beads (Beckman Coulter, California, USA). Long-read whole genome libraries were constructed using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies Ltd). DNA amount, integrity, and purity were assessed with a Qubit fluorometer (Thermo Fisher Scientific, Massachusetts, USA) and TapeStation electrophoresis (Agilent Technologies, California, USA).

Note: this work was conducted by Ana Pequeño-Valtierra and Jorge Rodríguez-Castro (Universidade de Santiago de Compostela).

Long-read whole-genome sequencing

ONT long-read sequencing of tumours and matched normal samples was performed with MinION R9.4 flow cells (FLO-MIN106, Oxford Nanopore Technologies). MinION devices were controlled with the MinKNOW software v18.12.09. High accuracy basecalling was performed in real-time using GPU-based Guppy software v2.3.1. Long reads were aligned to the human reference genome (GRCh38) with minimap2 v2.24 (H. Li 2018), and the aligned bam files were refined with Samtools v1.12 (H. Li et al. 2009). Aligned ONT tumour genomes presented a median depth of coverage of 33.7X. Haplotype-phasing of tumour and matched-normal ONT genomes was performed with WhatsHap (Martin et al. 2016).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Note: this work was conducted by Jorge Rodríguez-Castro, Javier Temes and other research technicians from Mobile Genomes (Universidade de Santiago de Compostela).

Reconstruction of source L1 elements

Taking advantage of our phased long-read sequencing data, we reconstructed the pool of potentially active L1 elements for each donor. These elements were defined as full-length L1s (FL-L1s) (>6Kb) detected in the matched-normal samples of each cancer patient. To reconstruct their sequences, we applied alternative approaches depending on whether the source elements were present or absent in the reference genome. For FL-L1 elements present in the reference genome, the reconstruction of their full sequence was achieved using the 'refSrc_seqs.py' workflow (included in the MEIGA suite). This workflow begins by collecting the aligned reads that span the L1 element and its surrounding regions. These reads are then discriminated based on their haplotype tags and mapped to an L1Hs consensus sequence. For each haplotype, the resulting alignments are combined and assembled using Wtdbg2. (Ruan and Li 2020). Next, the assembled FL-L1 sequences undergone a polishing step with Racon (Vaser et al. 2017). Finally, the FL-L1 sequences are realigned to the L1Hs consensus to define their precise breakpoints. For FL-L1s elements absent in the reference genome, we used our MEIGA calling to identify FL-L1 solo insertions in the matched-normal tissue. For the homozygous non-reference FL-L1s, we re-applied MEIGA on the target regions using the '--hpTag' parameter for the reconstruction of allele-specific sequences. For those heterozygous non-reference FL-L1s, we relied on the insertion sequences initially reconstructed by MEIGA under its default execution parameters.

Note: this work was conducted by Sonia Zumalave (Universidade de Santiago de Compostela).

Somatic retrotransposition calling on long-reads

We used MEIGA to detect and characterize L1 somatic retrotransposition events in Oxford Nanopore Technologies (ONT) long-read sequencing data. Somatic detection was performed through paired analysis using BAM files from tumour and matched-normal tissues, identifying cancer-specific events. This approach enabled the detection of different types of L1 somatic retrotransposition events, including solo-L1s (n=3611), partnered-L1 transductions (n=1535), and orphan-L1 transductions (n=705). For the L1 transduction analysis, we customized the set of potential source L1 elements for each donor to improve precision when tracing activity back to specific elements of origin. The process for obtaining these donor-specific candidate source L1 elements is described in the previous section. Each pool of potential L1 source elements was provided to MEIGA using the '--transduction-search' parameter, enabling the detection of 3' transduction events that mobilize FL-L1 downstream sequences.

Note: this work was conducted by Sonia Zumalave and Nuria Pereira-Espasandín (Universidade de Santiago de Compostela).

Source inference of solo-L1 retrotransposition

USC UNIVERSIDADE DE COMPOSTELA

We applied our Source Inference strategy to determine the element-of-origin for somatic retrotransposition events retaining any portion of the L1 sequence, including solo-L1s and

partnered transductions. To ensure precision, the program was executed over a specific set of candidate source L1 elements for each donor. In heterozygous candidate sources, the sequence of the reconstructed FL-L1 allele was considered, while for homozygous candidate sources with alternative FL-L1 haplotypes, one of the two alleles was selected arbitrarily. Somatic FL-L1 copies detected by MEIGA in the tumour genome were also included in the set of candidate sources.

To mitigate the impact of ONT sequencing errors, the source inference analysis focused exclusively on base substitutions as diagnostic variants. Diagnostic L1 nucleotides were identified based on minimap2 (H. Li 2018) alignments of all candidate sources against the L1Hs consensus, excluding divergent (>1%) FL-L1s to refine the catalogue of diagnostic variants and improve source inference sensitivity.

Informative positions were detected for both somatic L1 insertions and candidate sources based on pairwise alignments against the same L1Hs consensus. Source inference was performed under default parameters, requiring: 1) a 100% ratio of diagnostic variants shared between the insertion and the element-of-origin, 2) less than 100% ratio of diagnostic variants shared between the insertion and all alternative sources, and 3) a ≥75% ratio of total variants shared between the insertion and the element-of-origin. The source inference program was executed independently on each of the 10 golden set tumour genomes with high retrotransposition rates, leading to the identification of the element-of-origin for 899 solo-L1 somatic insertions.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Allelic tracing of somatic L1 retrotransposition

To explore allelic resolution in somatic L1 tracing, we adopted a strategy based on the identification of informative SNVs that discriminate between the two alleles of a given phased L1 locus. These discriminative positions were defined as nucleotide substitutions present only in one allele (hp1/hp2) and absent in the other (hp1/hp2), while also being undetected in the majority (≥99%) of the pool of candidate FL-L1 source alleles identified in the cancer patient. Discriminative positions were identified for each L1 source by aligning the two alternative allele sequences (hp1 and hp2) with minimap2 (H. Li 2018) against a L1Hs consensus, followed by Rsamtools (Danecek et al. 2021) pileup screening of the alignments to determine the informative status of each nucleotide variant. Once discriminative variants were identified for each homozygous L1 locus, we applied our source inference commands (steps 2 and 3) to genotype those variants in the set of somatic insertions previously ascribed to that locus. This allowed us to identify somatic L1 insertions sharing 100% of detected discriminative variants, which could be unambiguously traced to a particular L1 allele. To maximize the specificity of the analysis, allelic resolution was limited to homozygous source L1 loci with evidence of somatic activity derived from both L1 diagnostic nucleotide tracing and 3' transduction analysis.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Identification of MEIGA hot L1 elements and L1 phylogenetic analysis

Hot L1 elements are defined as source elements with greater somatic activity in our set of MEIGA tumors. They were identified based on one or more of the following conditions: 1) L1 activity rate (the average number of mobilizations mediated by a source element when found activated) equal to or greater than 10; 2) relative contribution to the total number of traced mobilizations in the cohort equal to or greater than 1.5%. These cutoffs were determined using density-based clustering with DBSCAN (Hahsler, Piekenbrock, and Doran 2019) v1.1-11.

Once identified, MEIGA hot source element sequences were used to build an L1 phylogeny. A multiple sequence alignment of FL-L1 sequences was performed using MAFFT v7.490 with default parameters, and upstream and downstream unspecific sequences were trimmed after visual inspection with Jalview v2.11. Consensus sequences from the golden set MEIGA samples were retrieved for each L1 source element using DECIPHER.

Consensus sequences from 27 L1 hot elements and the L1Pt sequence [GenBank: KF661301.1] were aligned with MUSCLE v3.36.0. The L1 phylogenetic tree was constructed using the following R The nucleotide substitution model was selected using phangorn::modelTest, with the GRT + G + I model chosen based on the Akaike Information Criterion (AIC) score. A maximum likelihood tree was estimated with phangorn::optim.pml, starting from a distance-based tree. Tree topology was adjusted via nearest neighbor interchange (Nguyen et al. 2015). After maximum likelihood optimization, midpoint rooting was applied, rendering L1Pt as the expected outgroup. A non-parametric bootstrap analysis with 1000 replicates was conducted to assess node consistency in the inferred tree, and for clarity, only nodes with high support (≥80%) were labeled.

L1Hs subfamily assignment was performed by identifying subfamily diagnostic nucleotide positions on the 3' end (Boissinot, Chevret, and Furano 2000). FL-L1 source elements bearing the diagnostic "ACG" or "ACA" triplet at positions 5929-5931 were classified as *pre-Ta* and Ta, respectively. Ta elements were further subclassified into *Ta-0* or *Ta-1*, based on diagnostic nucleotides at positions 5535 and 5538 (*Ta-0*: G and C; *Ta-1*: T and G). Source elements lacking any of these diagnostic profiles could not be assigned to a particular category, and their subfamily status remained undetermined.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Profiling of polyadenylation signals

We used APARENT (Bogard et al. 2019) to evaluate the polyadenylation sites potentially used by L1 source elements as termination signals to end their transcription. We applied the algorithm to profile the canonical termination site in the 3' end of the FL-L1 sequence, and alternative termination signals present in unique downstream regions, estimating their relative strengths. The strength of the detected polyadenylation sites was assessed against the propensity of FL-L1 sources to mediate solo-L1 insertions and 3' transductions events.

Note: this work was conducted by Sonia Zumalave (Universidade de Santiago de Compostela).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Simulation of retrotransposition events

To establish an in-silico benchmarking framework for the MEIGA project, we developed MEIsimulator, a method to generate sequencing datasets that mimic the characteristics of retrotransposition activity across whole genomes. MEIsimulator is an algorithm included in the MEIGA suite. It functions by placing various retrotransposon sequences at predetermined random positions within a mock reference genome and simulating sequencing reads based on the modified genome. To generate Illumina short-read sequencing datasets, ART (Huang et al. 2012) is utilized internally, while PBSIM (Ono, Asai, and Hamada 2013) is used to obtain ONT long-read simulated data. Notably, MEIsimulator enables the simulation of retrotransposition events according to different variant allele frequency (VAF) profiles. Currently, MEIsimulator allows for the generation of the following retrotransposition events: solo-L1, solo-Alu, solo-SVA, L1-transductions (partnered or orphan), and L1-mediated structural variants (deletions, duplications, inversions, and balanced translocations).

To evaluate our Source Inference strategy, we used MEIsimulator to model the mobilization of all candidate source elements (n=266) from the patient in the cohort with the highest L1 activity rate (PD0270), assuming homogeneous activity (20 mobilizations per source), with derived copies showing the same distribution of L1 insertion sizes as in 2,954 cancer genomes from a previous somatic retrotransposition study (Rodriguez-Martin et al. 2020). This led to the simulation of 5320 solo-L1 insertion events of variable size, whose sequences were retrieved for a Source Inference analysis in the same PD0270 donor. Source Inference was executed with default parameters, considering the same set of candidate L1 source elements from PD0270. After performing the Source Inference analysis, solo-L1 events that were traced to the correct element-of-origin were considered true positives, and sensitivity was then assessed considering two relevant variables: L1Hs subfamily (*pre-Ta*, *Ta-0*, and *Ta-1*) and derived L1 insertion size.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

METHODOLOGY OF CHAPTER II

Sample collection, processing and diagnosis

6854 *Cerastoderma edule* adult specimens were sampled from 36 locations in 11 countries including Morocco, Spain, Portugal, France, Ireland, United Kingdom, Germany, Netherlands, Denmark, Norway and Russia. Sampling covered the species' geographical range, spanning from the northern Barents Sea to the Atlantic coast of Morocco. Individuals were collected between 2016 and 2021 from natural intertidal sand beds in collaboration with shellfishermen's guild and researchers from the International Common Cockle Genome Consortium. Bivalve specimens were collected and processed with the approval of the Standing Committee on Conflict of Interest, Scientific Misconduct and Ethical Issues (CoIME) of the European Research Council, and under regional licenses for mollusc extractions and trading authorizations. Specimens were excluded for the study if were found to belong to a species other than *Cerastoderma edule*, as determined through morphological and molecular markers.

USC DE SANTIAGO

After sampling and shipment, cockles were maintained in closed-circuit seawater tanks for 48h before being dissected for obtaining histological, cytological and DNA samples for diagnosis and sequencing.

Disseminated neoplasia (DN) diagnosis was initially performed by microscopy examination of haemolymph cell monolayers, and further confirmed with histological analysis. Tumour purity, defined as neoplastic cell percentage, was estimated by inspecting 500 cells per sample on a light microscope. Tumour samples were ranked according to a scale of disease severity, comprising N0 (non-affected), N1 (early-stage; purity <15%), N2 (intermediate stage; purity 15-75%) and N3 (severe stage; purity >75%) levels. Histological analysis on 5-µm cockle sections enabled to determine disseminated neoplasia types (A/B) based on cell morphology and cell interaction features. Type-A neoplasia was associated to larger tumour cells, show pleomorphic nucleus and appear loosely packed, while type-B cells were smaller, showed rounded nucleus and present tighter clustering (M. J. Carballal et al. 2001; Metzger et al. 2016).

Note: this work was commanded by Alicia L. Bruzos and Seila Díaz, in association with shellfishermen's guild and collaborators from the International Common Cockle Genome Consortium.

DNA isolation and sequencing

Cockle DNA was isolated using QIAamp DNA Mini Kit (Qiagen), with an additional precipitation step with SDS/CH_3COOH (70°C, 10 min). DNA purity was evaluated using a Nanodrop One spectrophotometer (Thermo Fisher Scientific) and DNA integrity was evaluated with a 4200 TapeStation System (Agilent). Samples with poor DNA yields were subjected to whole-genome-amplification using REPLI-g Mini Kit (Qiagen). DNA libraries were prepared using Illumina whole-genome protocols, multiplexed and sequenced on an Illumina NovaSeq 6000 platform to generate 150bp paired-end reads. DNA was sequenced for 61 tumour samples, 40 normal samples from host cockles, and 462 normal samples from non-neoplastic cockles. Sequencing depth ranged between ~20X and 150X, depending on the type and purpose of each sample.

Note: this work was conducted by Alicia L. Bruzos, Ana Pequeño-Valtierra, Jorge Rodriguez-Castro, and other collaborators from the International Common Cockle Genome Consortium.

Computational sequence read alignment

Sequenced paired-end DNA reads were aligned to the *Cerastoderma edule* reference genome sequence using the BWA-mem algorithm as implemented in BWA v0.7.17 (with default settings). Aligned reads were sorted and indexed with samtools v1.9, and duplicated reads were marked with the bammarkduplicates tool from biobambam2 v2.0.87. Quality control and coverage metrics were obtained with samtools v1.9 and

Note: this work was conducted by Jorge Zamora, Daniel García-Souto, Alicia L. Bruzos and Martín Santamarina (Universidade de Santiago de Compostela).

Selection of a high-purity tumour set

A set of ten high-quality tumour samples was selected to constitute a golden set, required for analyses sensitive to host DNA contamination and coverage or allelic biases. The golden set

USC DE COMPOSTELA

was represented by samples that had sequencing depth ≥ 90GB, had purity estimates > 97%, had not undergone whole-genome amplification, and showed no evidence of co-infection by alternative tumour mtDNA lineages.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Calling, filtering and annotation of nuclear SNVs and indels

Calling of SNVs and indels in the 'golden set' of 10 tumour genomes (see section 'Selection of a high-purity tumour set') was performed using GATK MuTect2 (v4.1.6.0) in 'tumour-only mode' with default settings. Variant calling in samples from healthy (non-neoplastic) cockles was performed using Platypus (v0.8.1) with default settings. Our reference genome assembly was used as the reference sequence for all samples. MuTect2 calls were first filtered by assigning filter tags using the FilterMutectCalls tool in GATK (v4.1.6.0), and then selecting calls showing only filter tags 'PASS' or 'clustered_events'. To isolate potentially somatic variants and filter contaminating germline variation from the hosts, we identified likely germline variants from the sets of tumours variants by comparing them against the combined set of 'PASS'-tagged Platypus calls from the 462 healthy samples included in our 'panel of normals' (PON). This approach was required for two reasons: first, matched-normal tissue samples were found to harbour a considerable fraction of cancer cells, and were therefore unsuitable for filtering host contamination in tumour samples; second, because bivalve transmissible neoplasia (BTN) is transmitted as and allogeneic transplant, tumour cells are genetically unrelated to host cells, and thus the germline variation from hosts alone is insufficient for effective filtering of the germline variation inherited from the 'founder' animal which originally gave rise to each BTN clone.

The tumour variants that were absent from the PON were subsequently re-genotyped across all tumour samples using Platypus (options '--source=<VCF> --getVariantsFromBams=0 --minPosterior=0') and discarded if they were assigned any of the following Platypus filter tags: 'badReads', 'MQ', 'strandBias', 'SC', 'QD'. To maximize filtering of host-derived germline contamination, we then discarded variants fulfilling any of the following three criteria: (i) coverage < 20 reads in any tumour sample, (ii) VAF ≤ 0.2 in all tumour samples, or (iii) 0 < VAF < 0.05 in any tumour sample. This approach yielded a final filtered set of 5341963 variants (4340713 SNVs, 311113 MNVs, and 690137 indels), which are hereafter referred to as 'BTN-specific' variants. This set includes both somatic variants and ancestral germline variants that escaped filtering by the PON. For all subsequent analyses, a BTN-specific variant was considered to be present in a tumour sample if it had a VAF>0.

Annotation of BTN-specific variants was performed using the Ensembl Variant Effect Predictor (v104.3) with standard settings. The gene annotation for our reference genome assembly was used as a reference. Predicted variant consequences, CDS and protein positions, SIFT scores and other information were obtained for each variant. For subsequent analyses, we considered seven classes of nonsynonymous SNVs and indels: frameshift, stop gained, stop lost, start lost, missense, splice regions and coding sequence.

Note: This work was conducted by Adrián Baez-Ortega (University of Cambridge) and Iago Otero (University of Santiago de Compostela).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Calling, filtering and annotation of structural variants

SVs were detected and characterized with a combination of three state-of-the-art algorithms: DELLY (Rausch et al. 2012) v0.7.9, LUMPY ('LUMPY: A Probabilistic Framework for Structural Variant Discovery - PubMed', n.d.) v0.2.13, and Manta (Xiaoyu Chen et al. 2016) v1.6.0. Delly and Lumpy call structural variants using both discordant paired-end and split-read alignments while Manta in addition incorporates a local-assembly breakpoint validation to improve the detection of the events. To limit false positives, we considered only candidate events with precise (base-level) breakpoint resolution, and belonging to the following SV categories: deletions, duplications, inversions and breakends (or BNDs, including translocations). We integrated the SV events into a single callset using SVIMMER v0.1, requiring the events to have been detected by at least two SV callers (Manta and another method) to be considered as candidate variants. We genotyped all candidate SVs using GraphTyper(Eggertsson et al. 2019) v2.0 with default settings. Structural variants were genotyped both in the set of high-purity tumours and in a diverse set of 462 non-neoplastic cockles (PON) and annotated with the Ensembl Variant Effect Predictor (v104.3).

Sets of potentially somatic SVs were identified through additional filtering steps. First, SV genotypes were encoded in a binary format, indicating presence (1) or absence (0). Potentially somatic SVs were defined as those identified in at least one tumour genome and undetected in the PON. In order to explore the general patterns of structural variation within each CedBTN, this main set of candidate somatic variation was divided into 3 additional subsets: (i) 'ancestral' SVs, including variants detected in all the tumours from one cancer clone while being absent in any other lineage (likely representing early or pre-divergence somatic variation); (ii) 'internal' SVs, including variants detected in multiple tumours from one cancer clone but not all of them, while being absent in any other lineage (likely representing post-divergence somatic variation); and (iii) 'private' SVs, including variants detected in only one tumour sample (likely representing recent somatic variation).

Circos plots were produced using circlize (v0.4.14) to present the genomic distribution of candidate somatic SVs. For greater interpretability, circos plots did not include deletion and duplication events shorter than 10 kb.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Copy number inference

Copy number (CN) calling was performed with DELLY (v1.0.3), using the *Cerastoderma edule* reference genome assembly to correct for mappability and GC-content. The minimum CN alteration size was set to 10 kb. Read counts were obtained for variable size bins with 10 kb uniquely mappable (mapping quality ≥ 10) sequence. These bins were constructed by first simulating 2x150-bp PE reads from the reference genome sequence using dicey (v0.1.8) with otherwise default parameters. Simulated reads were aligned back to the indexed reference genome using BWA-MEM (v0.7.17), sorted, converted into BAM format, and indexed using samtools (v1.9). The final mappability map was generated using the 'dicey mappability2' tool with default parameters. CN segments were called for a range of ploidy values between 2n and 6n, and the most likely ploidy was then selected for each sample as the value providing the best fit between expected CN states and observed CN modes. The most likely ploidy was found to

be 4n for all samples except UGCE17/2401H (which was best fit by 5n), PACE17/421H1 and PACE17/433H (for both of which a ploidy of 4n was assumed, as their CN distributions were monomodal and thus uninformative).

Note: this work was conducted by Jonas Demeulemeester (The Francis Crick Institute).

Phylogenetic inference from nuclear variants

Tumour phylogenetic trees were estimated from nuclear sets of BTN-specific SNVs and SVs detected in the golden set. For SNVs, variant alleles were concatenated into an alignment containing 4340713 sites and 3724 different site patterns. ML trees were estimated with RAxML v8.2.12 using a single partition for the whole nuclear genome. A GTRGAMMA substitution model was assumed, given the low number of sequences and the high number of sites. Stamatakis ascertainment correction was applied, incorporating the exact nucleotide frequencies of invariable sites along the partition. A hundred trees were generated by optimizing alternative parsimony starting trees, and the tree with the best Gamma-based likelihood was selected. Tree consistency was evaluated using nonparametric bootstrap analysis with 1000 replicates. This tree was rooted using the reference sequence as an outgroup.

For SVs, binary genotypes derived from GraphTyper v0.2 were concatenated into an alignment using functions from the phangorn v2.8.1 package. Heuristic parsimony tree searches were performed with the implementation of the parsimony ratchet in phangorn. To evaluate the level of homoplasy, tree consistency indexes (CI) were calculated for the alternative phylogenies estimated from different types of SVs: deletions (CI=0.81), duplications (CI=0.91), inversions (CI=0.79) and breakends (CI=0.77). A maximum parsimony tree search was performed with PAUP v4.0a168. An alignment of 18272 SV binary genotypes was analysed, encoding the variants as unordered reversible characters with equal weights, and an exhaustive parsimony tree search was performed. Rooting was done using a user-specified outgroup corresponding to the reference genome sequence. Consistency of the tree was evaluated using nonparametric bootstrap analysis with 1000 replicates.

Principal component analysis was performed on a random subset of 100.000 germline polymorphisms (SNVs) found in the genomes of 462 non-neoplastic cockles (PON). Variants were genotyped across 10 tumour samples and 100 diverse non-neoplastic cockle samples covering all sampled locations from the species geographical distribution. Genotyping was performed with Platypus (v0.8.1). Dimensionality reduction of the binary genotype data was performed using the 'logisticSVD' function in the logisticPCA (v0.2) R package on default parameters. Samples were represented based on the scores of the first two principal components.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Identification of candidate driver mutations

To identify candidate early cancer-driver mutations in CedBTN, a set of cancer gene orthologs was first defined. The COSMIC Cancer Gene Census database of genes causally involved in human cancer (COSMIC v95) was retrieved. EggNOG gene identifiers were used to find *Cerastoderma edule* orthologs, rendering 226 putative cancer genes (located in 258 loci) across the cockle reference genome. A screen for potential early driver mutations in CedBTN1 and

USC DE SANTIAGO DE COMPOSTELA

CedBTN2 was conducted by searching for tumour-specific SNVs, indels and SVs satisfying the following criteria: I) the variant occurred pre-divergence; II) the variant affects the coding sequence of a cancer gene ortholog; III) the variant is nonsynonymous; and IV) the type of mutation matches one of the mutation types listed for the overlapping gene in the COSMIC Cancer Gene Census. Variant annotation was obtained with the Ensembl Variant Effect Predictor (v104.3)

This search yielded a list of early mutations in cancer genes. However, the combination of mutation consequence and gene copy number did not provide enough evidence that these events had affected the genes in a manner consistent with biological knowledge in human cancer, and therefore none was considered a high-confidence candidate early driver mutation in CedBTN.

Identification of candidate driver genes in CedBTN1 and CedBTN2 was also performed through the detection of CN changes associated with ancestral amplification of an oncogene, or with ancestral deletion of a TSG. CN estimation for each cancer gene was obtained through the intersection of unrounded CN segments with gene start and gene end coordinates. Intersections were obtained with the 'findOverlaps' function from the GenomicRanges v.1.46.1 R package. Gene-wise CN was defined as the average for intersected segments, normalized by their size. Gene-wise CN estimates were used to search for driver candidates in CedBTN1 and CedBTN2. Potential candidate driver oncogenes were defined as those with ancestral amplifications with CN > 6. Potential candidate driver TSGs were defined as those with ancestral losses with CN < 2. Potential candidate driver TSGs with multiple genomic copies, or with average CN status decrease indicative of hemizygous deletion, were further inspected in search of additional deleterious variants disrupting remaining alleles (homozygous inactivation). Candidate driver genes with greatest support were assessed for changes in gene expression relative to normal cockle tissues (differential gene expression analysis with RNAseq data).

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela) and Adrián Baez-Ortega (University of Cambridge).

METHODOLOGY OF CHAPTER III

TCGA cohorts

Methylation datasets were obtained from the GDC Data Portal (Project IDs TCGA-COAD, TCGA-READ, TCGA-STAD, TCGA-UCEC). Selected experimental platform was Illumina Infinium Human Methylation 450K. Downloaded data corresponded to 353 TCGA-COAD samples, 106 TCGA -READ samples, TCGA-STAD 397 samples and TCGA-UCEC 485 samples. Only 450K methylation profiles from fresh-frozen (FF) primary solid tumour samples were included.

Clinical information was downloaded from the GDC Data Portal [Project IDs: TCGA-COAD (458 cases), TCGA-READ (172 cases), TCGA-STAD (443 cases), TCGA-UCEC (560 cases)]. Gene expression and SNV information was downloaded from GDC Data Portal (TCGA-COAD). Additionally, CIMP status information of samples from TCGA COAD/READ cohort was downloaded from the UCSC Xena Portal (dataset ID: TCGA.COADREAD.sampleMap/COADREAD_clinicalMatrix).

MSI status was evaluated through PCR-based assays as previously described (Cancer Genome Atlas Research Network, Kandoth, et al. 2013; Cancer Genome Atlas Network 2012; Cancer Genome Atlas Research Network 2014). For each patient MSI status was encoded as a binary phenotype after considering MSS and MSI-Low profiles as MSS, and MSI-High profiles as MSI, in order to maximize the model predictive capacity toward relevant cases, as recommended by previous studies (Hause et al. 2016).

Among downloaded cases, only those from primary solid tumours with available DNA methylation beta (β)- values and MSI status information were used in this study. From the pooled TCGA-COAD/READ cohort, 387 out of 459 cases fulfilled these criteria. From TCGA-STAD and TCGA-UCEC cohorts, 394 out 397 cases and 420 out of 485 cases fulfilled these criteria, respectively.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

External cohorts

External cohort I: Methylation dataset was downloaded from the Gene Expression Omnibus (GEO) data repository (GEO accession number: GSE69550). Experimental platform was Illumina Infinium Human Methylation 450K. Downloaded data corresponded to 79 formalin-fixed paraffin-embedded (FFPE) colorectal primary tumours. Three regions per primary tumour were selected: the region nearest the digestive tract surface (luminal), the central bulk (centre) and the invasive front (front). MSI status was provided by the corresponding author upon request (Martínez-Cardús et al. 2016).

External cohort II: Methylation dataset was downloaded from the European Genome-phenome Archive (EGA accession number: EGAS00001004293). Experimental platform was Illumina Infinium Human Methylation 450K. Downloaded data corresponded to 262 FF colorectal primary tumours. MSI status and clinical characteristics were provided by the corresponding author upon request (Mattesen et al. 2020).

External cohort III: Methylation dataset and MSI status information were downloaded from the ArrayExpress Archive of Functional Genomics Data repository (accession number: E-GEOD-68060). Experimental platform was Illumina Infinium Human Methylation 450K. Downloaded data corresponded to 81 FF colorectal primary tumours. Clinical characteristics were provided by the corresponding author upon request (Conesa-Zamora et al. 2015).

External cohort IV: Methylation dataset was downloaded from the Japanese Genotype-phenotype Archive (JGA) (JGA accession number: JGAS00000000113). Experimental platform was Illumina Infinium MethylationEPIC array. Downloaded data correspond to 95 FF MSI-H colorectal primary tumours. MSI status and clinical characteristics were provided by the corresponding author upon request (Namba et al. 2019).

Note: this work was conducted by Martín Santamarina and Juan Ruiz-Bañobre (Universidade de Santiago de Compostela) in association with collaborators of the project.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

TCGA datasets annotation

Annotation for methylation datasets was obtained with the ChAMP v2.24.0 R package (Morris et al. 2014) retrieving the following information: probe type, strand, genomic coordinates, CpG island context (island ID and associated features) and gene context (gene symbol and associated features). Gene promoter was defined considering CpG probes located up to 1.5 kb upstream from the transcription start site (TSS1500 and TSS200 features). Annotation for cancer gene information was obtained from the Cancer Gene Census database (COSMIC v96 release) (Sondka et al. 2018). Genome-wide annotation for regulatory elements (curated regulatory regions, TFBS and regulatory polymorphisms) was obtained from the Open Regulatory Annotation track available at the UCSC Genome Browser database (Kent et al. 2002).

Note: this work was conducted by Martín Santamarina and Jenifer Brea-Iglesias (Universidade de Santiago de Compostela)

Analysis of DNA methylation arrays

Analysis of DNA methylation arrays was performed with the ChAMP v2.24.0 R package (Morris et al. 2014). Methylation probe filtering was done with *champ.filter*, setting the exclusion of (I) non-CpG probes, (II) Multi-Hit probes, (III) probes matching SNPs, and (IV) probes located in sex chromosomes. Methylation data imputation was done with *champ.impute* under combined method (partial removal followed by KNN imputation) setting ProbeCutoff = 0.2, SampleCutoff = 0.2 and k=5 parameters. Quality control was performed with *champ.QC*. Normalization of Type-I and Type-II probes was performed with *champ.norm* under peak-based correction (PBC) method. Identification of differential methylation positions was performed with *champ.DMP* by comparing MSI vs MSS samples, applying Benjamini-Hochberg (B-H) False Discovery Rate (FDR) p value adjustment after *limma* analysis. MLH1 promoter methylation status was defined as previously described (Benhamida et al. 2020). Briefly, if β-value for all four MLH1 CpG sites cg23658326, cg11600697, cg21490561, and cg00893636 were greater than or equal to 0.18, 0.27, 0.11 and 0.10, respectively, the sample was considered hypermethylated for MLH1 promoter. Otherwise, the sample was considered non-hypermethylated.

Note: this work was conducted by Martín Santamarina and Jenifer Brea-Iglesias (Universidade de Santiago de Compostela)

Machine-learning predictive models

The selection of predictors for MSIMEP was done by retaining those probes from the pooled TCGA COAD/READ dataset with an absolute delta β-value (Δβ) > 0.3 between MSI and MSS tumours, and B-H FDR adjusted p value <0.05. Both hypermethylated and hypomethylated probes were included as predictors, and patient binary MSI/MSS status was considered as response variable. Model development was performed with caret v6.0-91 R package (Kuhn 2008), training in the TCGA COAD/READ dataset the following machine-learning classifiers: GLM (*glmnet*), KNN (*knn*), SVM (*svmLinear*), NaiveBayes (*naive_bayes*), NeuralNetwork (*nnet*), DecissionTree (*rpart*), and RandomForest (*cforest*). Classifier performance was assessed through 5 repeats of 10-fold cross-validation. During classifier development, the hyperparameter tuning was performed through grid search. Average model accuracies were compared through Friedman Rank-Sum Test and Pairwise Wilcoxon Rank-Sum Test (paired, p.adjust.method = "holm"), and finally, the less complex model among those with significantly

USC DE COMPOSTELA

better performance was selected. Recursive feature elimination was used to decrease model complexity, selecting an optimized MSIMEP model and a reduced version with suitable performance (less than 1% accuracy loss). During RFE, an outer resampling method with 3 repeats of 2-fold cross-validation was applied. Predictors retained during each of the backward selection steps corresponded to 700, 600, 500, 400, 300, 200, 100, 50, 25, 10, 5 and 2 CpG probes.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Model description

Relative CpG probe importance for MSIMEP was measured with *caret::varImp* (Kuhn 2008), considering the absolute value of the coefficients corresponding to the optimized GLM model. Relative gene importance was estimated by adding the absolute importance of individual CpG probes associated with each gene, normalized by the total number of genes represented in the classification model. Gene set enrichment analysis was performed with the methylGSA v1.12.0 R package (Ren and Kuan 2019), which implements logistic regression adjusting for the number of methylation probes. Enrichment in KEGG categories was evaluated with the *methylglm* function. CpG sites located in different gene context regions of the Infinium Human Methylation 450K array were considered ("group = promoter", "group = body" and "group = all"). Gene sets composed by 20 - 500 members were displayed and ranked according to their B-H FDR-adjusted p value.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

MSIMEP exploratory analysis in alternative TCGA cohorts

MSIMEP prediction of MSI/MSS phenotype in TCGA-UCEC and TCGA-STAD cohorts was performed using the caret *predict* function, by assessing the methylation profiles for the same 400 CpG predictors (MSIMEP optimized model), or 25 CpG predictors (MSIMEP reduced model), and considering patient binary MSI/MSS status as response variable. Methylation β-values were processed under a common framework (see "*Analysis of DNA methylation arrays*"). Accuracy, Kappa, ROC, Sensitivity, Specificity and Precision metrics were calculated from the predicted class probabilities (caret *predict* "type = prob") and associated confusion matrix comparing predicted and expected MSI/MSS status. ROC curves were generated using the MLeval v0.3 R package (John 2020).

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

MSIMEP validation in external colorectal cancer cohorts

MSIMEP prediction of MSI/MSS phenotype in External cohort I, External cohort II, External cohort III and External cohort IV was performed using the caret *predict* function, by assessing the methylation profiles for the same 400 CpG predictors (MSIMEP optimized model), or 25 CpG predictors (MSIMEP reduced model), and considering patient binary MSI/MSS status as response variable. Methylation β-values were processed under a common framework (see "*Analysis of DNA methylation arrays*"). Accuracy, Kappa, ROC, Sensitivity, Specificity and

USC DE COMPOSTELA

Precision metrics were calculated from the predicted class probabilities (caret *predict* "type = prob") and associated confusion matrix comparing predicted and expected MSI/MSS status. ROC curves were generated using the MLeval v0.3 R package. Venn diagram was generated using the VennDiagram v1.7.1 R package (H. Chen and Boutros 2011).

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

Deconvolution of immune cell populations

The CIBERSORTx deconvolution algorithm (Newman et al. 2019) was used to determine immune cell infiltration from TCGA-COAD and TCGA-READ bulk RNA-seq gene expression data. CIBERSORTx job type: "Impute cell fractions" was executed in absolute mode on the LM22 signature, under recommended parameters for bulk RNA-seq mixtures. The CIBERSORTx analysis was executed with 500 permutations. Low-quality samples with p value for deconvolution > 0.05 were discarded. Relative cell proportions were determined by normalizing the CIBERSORTx output to the sample-level sum of cell scores rendering percentages of immune infiltration. Relative cell proportions were compared between MSIMEP-predicted MSI CRCs and MSI cases diagnosed through standard PCR-based assays. The immune cell types included in the analysis corresponded to B cells (naive and memory), plasmatic cells, T cells (CD8, CD4 naive, CD4 memory resting, CD4 memory activated, follicular helper, regulatory and gamma-delta), NK cells (resting and activated), monocytes, macrophages (M0, M1 and M2), dendritic cells (resting and activated), mast cells (resting and activated), eosinophils and neutrophils.

Note: this work was conducted by Martín Santamarina (Universidade de Santiago de Compostela)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

BIBLIOGRAPHY

1. Aktipis, C. Athena, Amy M. Boddy, Gunther Jansen, Urszula Hibner, Michael E. Hochberg, Carlo C. Maley, and Gerald S. Wilkinson. 2015. 'Cancer across the Tree of Life: Cooperation and Cheating in Multicellularity'. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1673): 20140219. https://doi.org/10.1098/rstb.2014.0219.
2. Allam, Bassem, and David Raftos. 2015. 'Immune Responses to Infectious Diseases in Bivalves'. *Journal of Invertebrate Pathology* 131 (October):121–36. https://doi.org/10.1016/j.jip.2015.05.005.
3. Álvarez, Eva G., Jonas Demeulemeester, Paula Otero, Clemency Jolly, Daniel García-Souto, Ana Pequeño-Valtierra, Jorge Zamora, et al. 2021. 'Aberrant Integration of Hepatitis B Virus DNA Promotes Major Restructuring of Human Hepatocellular Carcinoma Genome Architecture'. *Nature Communications* 12 (1): 6910. https://doi.org/10.1038/s41467-021-26805-8.
4. Apweiler, Rolf, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, et al. 2004. 'UniProt: The Universal Protein Knowledgebase'. *Nucleic Acids Research* 32 (Database issue): D115-119. https://doi.org/10.1093/nar/gkh131.
5. Baca, Sylvan C., Davide Prandi, Michael S. Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, et al. 2013. 'Punctuated Evolution of Prostate Cancer Genomes'. *Cell* 153 (3): 666–77. https://doi.org/10.1016/j.cell.2013.03.021.
6. Baez-Ortega, Adrian, Kevin Gori, Andrea Strakova, Janice L. Allen, Karen M. Allum, Leontine Bansse-Issa, Thinlay N. Bhutia, et al. 2019. 'Somatic Evolution and Global Expansion of an Ancient Transmissible Cancer Lineage'. *Science (New York, N.Y.)* 365 (6452): eaau9923. https://doi.org/10.1126/science.aau9923.
7. Baez-Ortega, Adrian, and Elizabeth P. Murchison. 2021. 'Searching for Transmissible Cancers among the Mussels of Europe'. *Molecular Ecology*, December. https://doi.org/10.1111/mec.16330.
8. Bagshaw, Andrew T.M. 2017. 'Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes'. *Genome Biology and Evolution* 9 (9): 2428–43. https://doi.org/10.1093/gbe/evx164.
9. Baretti, Marina, and Dung T. Le. 2018. 'DNA Mismatch Repair in Cancer'. *Pharmacology & Therapeutics* 189 (September):45–62. https://doi.org/10.1016/j.pharmthera.2018.04.004.
10. Beck, Christine R., Pamela Collier, Catriona Macfarlane, Maika Malig, Jeffrey M. Kidd, Evan E. Eichler, Richard M. Badge, and 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.
11. Beck, Christine R., José Luis Garcia-Perez, Richard M. Badge, and 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.
12. Belancio, V. P., A. M. Roy-Engel, and P. Deininger. 2008. 'The Impact of Multiple Splice Sites in Human L1 Elements'. *Gene* 411 (1–2): 38–45. https://doi.org/10.1016/j.gene.2007.12.022.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

13. Belancio, Victoria P., Megan Whelton, and Prescott Deininger. 2007. 'Requirements for Polyadenylation at the 3' End of LINE-1 Elements'. *Gene* 390 (1–2): 98–107. https://doi.org/10.1016/j.gene.2006.07.029.
14. Benhamida, Jamal K., Jaclyn F. Hechtman, Khedoudja Nafa, Liliana Villafania, Justyna Sadowska, Jiajing Wang, Donna Wong, et al. 2020. 'Reliable Clinical MLH1 Promoter Hypermethylation Assessment Using a High-Throughput Genome-Wide Methylation Array Platform'. *The Journal of Molecular Diagnostics: JMD* 22 (3): 368–75. https://doi.org/10.1016/j.jmoldx.2019.11.005.
15. Bogard, Nicholas, Johannes Linder, Alexander B. Rosenberg, and Georg Seelig. 2019. 'A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation'. *Cell* 178 (1): 91-106.e23. https://doi.org/10.1016/j.cell.2019.04.046.
16. Boissinot, S., P. Chevret, and A. V. Furano. 2000. 'L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History'. *Molecular Biology and Evolution* 17 (6): 915–28. https://doi.org/10.1093/oxfordjournals.molbev.a026372.
17. Bramwell, Georgina, James DeGregori, Frédéric Thomas, and Beata Ujvari. 2024. 'Transmissible Cancers, the Genomes That Don't Melt Down'. *Evolution; International Journal of Organic Evolution*, April, qpae063. https://doi.org/10.1093/evolut/qpae063.
18. Brouha, Brook, Christof Meischl, Eric Ostertag, Martin de Boer, Yue Zhang, Herman Neijens, Dirk Roos, and Haig H. Kazazian. 2002. 'Evidence Consistent with Human L1 Retrotransposition in Maternal Meiosis I'. *American Journal of Human Genetics* 71 (2): 327–36. https://doi.org/10.1086/341722.
19. Brouha, Brook, Joshua Schustak, Richard M. Badge, Sheila Lutz-Prigge, Alexander H. Farley, John V. Moran, and Haig H. Kazazian. 2003. 'Hot L1s Account for the Bulk of Retrotransposition in the 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.
20. Brown, Joel S., Sarah R. Amend, Robert H. Austin, Robert A. Gatenby, Emma U. Hammarlund, and Kenneth J. Pienta. 2023. 'Updating the Definition of Cancer'. *Molecular Cancer Research* 21 (11): 1142–47. https://doi.org/10.1158/1541-7786.MCR-23-0411.
21. Brügger, Kim, Peter Redder, Qunxin She, Fabrice Confalonieri, Yvan Zivanovic, and Roger A. Garrett. 2002. 'Mobile Elements in Archaeal Genomes'. *FEMS Microbiology Letters* 206 (2): 131–41. https://doi.org/10.1111/j.1574-6968.2002.tb10999.x.
22. Bruzos, Alicia L., Martín Santamarina, Daniel García-Souto, Seila Díaz, Sara Rocha, Jorge Zamora, Yunah Lee, et al. 2023. 'Somatic Evolution of Marine Transmissible Leukemias in the Common Cockle, Cerastoderma Edule'. *Nature Cancer*, October. https://doi.org/10.1038/s43018-023-00641-9.
23. Burioli, E. a. V., M. Hammel, N. Bierne, F. Thomas, M. Houssin, D. Destoumieux-Garzón, and G. M. Charrière. 2021. 'Traits of a Mussel Transmissible Cancer Are Reminiscent of a Parasitic Life Style'. *Scientific Reports* 11 (1): 24110. https://doi.org/10.1038/s41598-021-03598-w.
24. Burioli, E. a. V., M. Hammel, E. Vignal, J. Vidal-Dupiol, G. Mitta, F. Thomas, N. Bierne, D. Destoumieux-Garzón, and G. M. Charrière. 2023. 'Transcriptomics of Mussel Transmissible Cancer MtrBTN2 Suggests Accumulation of Multiple Cancer Traits and Oncogenic Pathways Shared among Bilaterians'. *Open Biology* 13 (10): 230259. https://doi.org/10.1098/rsob.230259.
25. Cancer Genome Atlas Network. 2012. 'Comprehensive Molecular Characterization of Human Colon and Rectal Cancer'. *Nature* 487 (7407): 330–37.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

https://doi.org/10.1038/nature11252.
26. Cancer Genome Atlas Research Network. 2014. 'Comprehensive Molecular Characterization of Gastric Adenocarcinoma'. *Nature* 513 (7517): 202–9. https://doi.org/10.1038/nature13480.
27. Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D. Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, et al. 2013. 'Integrated Genomic Characterization of Endometrial Carcinoma'. *Nature* 497 (7447): 67–73. https://doi.org/10.1038/nature12113.
28. Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. 'The Cancer Genome Atlas Pan-Cancer Analysis Project'. *Nature Genetics* 45 (10): 1113–20. https://doi.org/10.1038/ng.2764.
29. Cao, Yang, Dan Wang, Jin Wu, Zhanxin Yao, Si Shen, Chao Niu, Ying Liu, et al. 2023. 'MSI-XGNN: An Explainable GNN Computational Framework Integrating Transcription- and Methylation-Level Biomarkers for Microsatellite Instability Detection'. *Briefings in Bioinformatics* 24 (6): bbad362. https://doi.org/10.1093/bib/bbad362.
30. Carballal, M. J., D. Iglesias, J. Santamarina, B. Ferro-Soto, and A. Villalba. 2001. 'Parasites and Pathologic Conditions of the Cockle Cerastoderma Edule Populations of the Coast of Galicia (NW Spain)'. *Journal of Invertebrate Pathology* 78 (2): 87–97. https://doi.org/10.1006/jipa.2001.5049.
31. Carballal, María J., Bruce J. Barber, David Iglesias, and Antonio Villalba. 2015. 'Neoplastic Diseases of Marine Bivalves'. *Journal of Invertebrate Pathology* 131 (October):83–106. https://doi.org/10.1016/j.jip.2015.06.004.
32. Chen, Hanbo, and Paul C. Boutros. 2011. 'VennDiagram: A Package for the Generation of Highly-Customizable Venn and Euler Diagrams in R'. *BMC Bioinformatics* 12 (January):35. https://doi.org/10.1186/1471-2105-12-35.
33. Chen, Tao, Cangui Zhang, Yingqiao Liu, Yuyun Zhao, Dingyi Lin, Yanfeng Hu, Jiang Yu, and Guoxin Li. 2019. 'A Gastric Cancer LncRNAs Model for MSI and Survival Prediction Based on Support Vector Machine'. *BMC Genomics* 20 (1): 846. https://doi.org/10.1186/s12864-019-6135-x.
34. Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. 'Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications'. *Bioinformatics (Oxford, England)* 32 (8): 1220–22. https://doi.org/10.1093/bioinformatics/btv710.
35. Chen, Xuelan, Albert S. Agustinus, Jun Li, Melody DiBona, and Samuel F. Bakhoum. 2024. 'Chromosomal Instability as a Driver of Cancer Progression'. *Nature Reviews. Genetics*, July. https://doi.org/10.1038/s41576-024-00761-7.
36. Choudhary, Mayank N. K., Kara Quaid, Xiaoyun Xing, Heather Schmidt, and Ting Wang. 2023. 'Widespread Contribution of Transposable Elements to the Rewiring of Mammalian 3D Genomes'. *Nature Communications* 14 (1): 634. https://doi.org/10.1038/s41467-023-36364-9.
37. Chu, Chong, Rebeca Borges-Monroy, Vinayak V. Viswanadham, Soohyun Lee, Heng Li, Eunjung Alice Lee, and 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.
38. Chuang, Nelson T., Eugene J. Gardner, Diane M. Terry, Jonathan Crabtree, Anup A.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Mahurkar, Guillermo L. Rivell, Charles C. Hong, James A. Perry, and Scott E. Devine. 2021. 'Mutagenesis of Human Genomes by Endogenous Mobile Elements on a Population Scale'. *Genome Research* 31 (12): 2225–35. https://doi.org/10.1101/gr.275323.121.

39. Conesa-Zamora, Pablo, José García-Solano, María del Carmen Turpin, Patricia Sebastián-León, Daniel Torres-Moreno, Eduardo Estrada, Anne Tuomisto, et al. 2015. 'Methylome Profiling Reveals Functions and Genes Which Are Differentially Methylated in Serrated Compared to Conventional Colorectal Carcinoma'. *Clinical Epigenetics* 7 (1): 101. https://doi.org/10.1186/s13148-015-0128-7.
40. Coorens, Tim H. H., Thomas R. W. Oliver, Rashesh Sanghvi, Ulla Sovio, Emma Cook, Roser Vento-Tormo, Muzlifah Haniffa, et al. 2021. 'Inherent Mosaicism and Extensive Mutation of Human Placentas'. *Nature* 592 (7852): 80–85. https://doi.org/10.1038/s41586-021-03345-1.
41. Cordaux, Richard, and 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.
42. Cortes-Ciriano, Isidro, Sejoon Lee, Woong-Yang Park, Tae-Min Kim, and Peter J. Park. 2017. 'A Molecular Portrait of Microsatellite Instability across Multiple Cancers'. *Nature Communications* 8 (June):15180. https://doi.org/10.1038/ncomms15180.
43. Cosenza, Marco Raffaele, Bernardo Rodriguez-Martin, and Jan O. Korbel. 2022. 'Structural Variation in Cancer: Role, Prevalence, and Mechanisms'. *Annual Review of Genomics and Human Genetics* 23 (1): annurev-genom-120121-101149. https://doi.org/10.1146/annurev-genom-120121-101149.
44. Cosenza, Marco Raffaele, Bernardo Rodriguez-Martin, and Jan O. Korbel 'Structural Variation in Cancer: Role, Prevalence, and Mechanisms'. *Annual Review of Genomics and Human Genetics* 23 (August):123–52. https://doi.org/10.1146/annurev-genom-120121-101149.
45. Danaher, Patrick, Sarah Warren, SuFey Ong, Nathan Elliott, Alessandra Cesano, and Sean Ferree. 2019. 'A Gene Expression Assay for Simultaneous Measurement of Microsatellite Instability and Anti-Tumor Immune Activity'. *Journal for Immunotherapy of Cancer* 7 (1): 15. https://doi.org/10.1186/s40425-018-0472-1.
46. Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. 'Twelve Years of SAMtools and BCFtools'. *GigaScience* 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.
47. Deininger, P. L., M. A. Batzer, C. A. Hutchison, and M. H. Edgell. 1992. 'Master Genes in Mammalian Repetitive DNA Amplification'. *Trends in Genetics: TIG* 8 (9): 307–11. https://doi.org/10.1016/0168-9525(92)90262-3.
48. Deininger, Prescott. 2011. 'Alu Elements: Know the SINEs'. *Genome Biology* 12 (12): 236. https://doi.org/10.1186/gb-2011-12-12-236.
49. 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.
50. Devine, Scott E. 2023. 'Emerging Opportunities to Study Mobile Element Insertions and Their Source Elements in an Expanding Universe of Sequenced Human Genomes'. *Genes* 14 (10): 1923. https://doi.org/10.3390/genes14101923.
51. Díaz, S., A. Cao, A. Villalba, and M. J. Carballal. 2010. 'Expression of Mutant Protein P53 and Hsp70 and Hsp90 Chaperones in Cockles Cerastoderma Edule Affected by

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Neoplasia'. *Diseases of Aquatic Organisms* 90 (3): 215–22. https://doi.org/10.3354/dao02231.

52. Díaz, S., A. Villalba, A. Insua, P. Soudant, J. Fernández-Tajes, J. Méndez, and M. J. Carballal. 2013. 'Disseminated Neoplasia Causes Changes in Ploidy and Apoptosis Frequency in Cockles Cerastoderma Edule'. *Journal of Invertebrate Pathology* 113 (3): 214–19. https://doi.org/10.1016/j.jip.2013.03.010.
53. Dombroski, B. A., S. L. Mathias, E. Nanthakumar, A. F. Scott, and H. H. Kazazian. 1991. 'Isolation of an Active Human Transposable Element'. *Science (New York, N.Y.)* 254 (5039): 1805–8. https://doi.org/10.1126/science.1662412.
54. 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 (New York, N.Y.)* 372 (6537): eabf7117. https://doi.org/10.1126/science.abf7117.
55. Edgar, Robert C. 2004. 'MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput'. *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.
56. Eggertsson, Hannes P., Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T. Hardarson, Daniel F. Gudbjartsson, Kari Stefansson, Bjarni V. Halldorsson, and Pall Melsted. 2019. 'GraphTyper2 Enables Population-Scale Genotyping of Structural Variation Using Pangenome Graphs'. *Nature Communications* 10 (1): 5402. https://doi.org/10.1038/s41467-019-13341-9.
57. Elbarbary, Reyad A., Bronwyn A. Lucas, and Lynne E. Maquat. 2016. 'Retrotransposons as Regulators of Gene Expression'. *Science* 351 (6274): aac7247. https://doi.org/10.1126/science.aac7247.
58. Fanley, C. A. 1969. 'Probable Neoplastic Disease of the Hematopoietic System in Oysters, Crassostrea Virginica and Crassostrea Gigas'. *National Cancer Institute Monograph* 31 (July):541–55.
59. Faulkner, Geoffrey J., and Victor Billon. 2018. 'L1 Retrotransposition in the Soma: A Field Jumping Ahead'. *Mobile DNA* 9:22. https://doi.org/10.1186/s13100-018-0128-1.
60. Ferlay, Jacques, Murielle Colombet, Isabelle Soerjomataram, Donald M. Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. 2021. 'Cancer Statistics for the Year 2020: An Overview'. *International Journal of Cancer* 149 (4): 778–89. https://doi.org/10.1002/ijc.33588.
61. Foltz, Steven M., Wen-Wei Liang, Mingchao Xie, and Li Ding. 2017. 'MIRMMR: Binary Classification of Microsatellite Instability Using Methylation and Mutations'. *Bioinformatics (Oxford, England)* 33 (23): 3799–3801. https://doi.org/10.1093/bioinformatics/btx507.
62. Frank, John A., Manvendra Singh, Harrison B. Cullen, Raphael A. Kirou, Meriem Benkaddour-Boumzaouad, Jose L. Cortes, Jose Garcia Pérez, Carolyn B. Coyne, and Cédric Feschotte. 2022. 'Evolution and Antiviral Activity of a Human Protein of Retroviral Origin'. *Science (New York, N.Y.)* 378 (6618): 422–28. https://doi.org/10.1126/science.abq7871.
63. Fu, Yelin, Lishuang Qi, Wenbing Guo, Liangliang Jin, Kai Song, Tianyi You, Shuobo Zhang, Yunyan Gu, Wenyuan Zhao, and Zheng Guo. 2019. 'A Qualitative Transcriptional Signature for Predicting Microsatellite Instability Status of Right-Sided Colon Cancer'. *BMC Genomics* 20 (1): 769. https://doi.org/10.1186/s12864-019-6129-8.
64. Gandhi, Manish J., and D. Michael Strong. 2007. 'Donor Derived Malignancy

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Following Transplantation: A Review'. *Cell and Tissue Banking* 8 (4): 267–86. https://doi.org/10.1007/s10561-007-9036-1.

65. Ganesh, Karuna, and Joan Massagué. 2021. 'Targeting Metastatic Cancer'. *Nature Medicine* 27 (1): 34–44. https://doi.org/10.1038/s41591-020-01195-4.
66. Garcia-Perez, Jose L., Maria C. N. Marchetto, Alysson R. Muotri, Nicole G. Coufal, Fred H. Gage, K. Sue O'Shea, and John V. Moran. 2007. 'LINE-1 Retrotransposition in Human Embryonic Stem Cells'. *Human Molecular Genetics* 16 (13): 1569–77. https://doi.org/10.1093/hmg/ddm105.
67. 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.
68. Garcia-Souto, Daniel, Alicia L. Bruzos, Seila Diaz, Sara Rocha, Ana Pequeño-Valtierra, Camila F. Roman-Lewis, Juana Alonso, et al. 2022. 'Mitochondrial Genome Sequencing of Marine Leukaemias Reveals Cancer Contagion between Clam Species in the Seas of Southern Europe'. *eLife* 11 (January):e66946. https://doi.org/10.7554/eLife.66946.
69. Gardner, Eugene J., Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, and 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.
70. Gärtner, H. V., C. Seidl, C. Luckenbach, G. Schumm, E. Seifried, H. Ritter, and B. Bültmann. 1996. 'Genetic Analysis of a Sarcoma Accidentally Transplanted from a Patient to a Surgeon'. *The New England Journal of Medicine* 335 (20): 1494–96. https://doi.org/10.1056/NEJM199611143352004.
71. Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, et al. 2020. 'The Evolutionary History of 2,658 Cancers'. *Nature* 578 (7793): 122–28. https://doi.org/10.1038/s41586-019-1907-7.
72. Giardiello, Francis M., John I. Allen, Jennifer E. Axilbund, C. Richard Boland, Carol A. Burke, Randall W. Burt, James M. Church, et al. 2014. 'Guidelines on Genetic Evaluation and Management of Lynch Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer'. *The American Journal of Gastroenterology* 109 (8): 1159–79. https://doi.org/10.1038/ajg.2014.186.
73. Giersch, Rachael M., Samuel F. M. Hart, Satyatejas G. Reddy, Marisa A. Yonemitsu, María J. Orellana Rosales, Madelyn Korn, Brook M. Geleta, Peter D. Countway, José A. Fernández Robledo, and Michael J. Metzger. 2022. 'Survival and Detection of Bivalve Transmissible Neoplasia from the Soft-Shell Clam Mya Arenaria (MarBTN) in Seawater'. *Pathogens (Basel, Switzerland)* 11 (3): 283. https://doi.org/10.3390/pathogens11030283.
74. Gilbert, Nicolas, Sheila Lutz-Prigge, and John V. Moran. 2002. 'Genomic Deletions Created upon LINE-1 Retrotransposition'. *Cell* 110 (3): 315–25. https://doi.org/10.1016/S0092-8674(02)00828-0.
75. González-Tizón, A., A. Martínez-Lage, J. Ausio, and J. Méndez. 2000. 'Polyploidy in a Natural Population of Mussel, Mytilus Trossulus'. *Genome* 43 (2): 409–11. https://doi.org/10.1139/gen-43-2-409.
76. Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. 'Circlize Implements and Enhances Circular Visualization in R'. *Bioinformatics*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

(Oxford, England) 30 (19): 2811–12. https://doi.org/10.1093/bioinformatics/btu393.
77. Guallar, Diana, Xianju Bi, Jose Angel Pardavila, Xin Huang, Carmen Saenz, Xianle Shi, Hongwei Zhou, et al. 2018. 'RNA-Dependent Chromatin Targeting of TET2 for Endogenous Retrovirus Control in Pluripotent Stem Cells'. *Nature Genetics* 50 (3): 443–51. https://doi.org/10.1038/s41588-018-0060-9.
78. Haddrill, Penelope R. 2021. 'Developments in Forensic DNA Analysis'. *Emerging Topics in Life Sciences* 5 (3): 381–93. https://doi.org/10.1042/ETLS20200304.
79. Hahsler, Michael, Matthew Piekenbrock, and Derek Doran. 2019. 'Dbscan: Fast Density-Based Clustering with R'. *Journal of Statistical Software* 91 (October):1–30. https://doi.org/10.18637/jss.v091.i01.
80. Halo, Julia V., Amanda L. Pendleton, Abigail S. Jarosz, Robert J. Gifford, Malika L. Day, and Jeffrey M. Kidd. 2019. 'Origin and Recent Expansion of an Endogenous Gammaretroviral Lineage in Domestic and Wild Canids'. *Retrovirology* 16 (1): 6. https://doi.org/10.1186/s12977-019-0468-z.
81. Hammel, Maurine, Alexis Simon, Christine Arbiol, Antonio Villalba, Erika A. V. Burioli, Jean-François Pépin, Jean-Baptiste Lamy, et al. 2022. 'Prevalence and Polymorphism of a Mussel Transmissible Cancer in Europe'. *Molecular Ecology* 31 (3): 736–51. https://doi.org/10.1111/mec.16052.
82. Hancks, Dustin C., and Haig Kazazian. 2010. 'SVA Retrotransposons: Evolution and Genetic Instability'. *Seminars in Cancer Biology* 20 (4): 234–45. https://doi.org/10.1016/j.semcancer.2010.04.001.
83. Hart, Samuel F. M., Marisa A. Yonemitsu, Rachael M. Giersch, Fiona E. S. Garrett, Brian F. Beal, Gloria Arriagada, Brian W. Davis, Elaine A. Ostrander, Stephen P. Goff, and Michael J. Metzger. 2023. 'Centuries of Genome Instability and Evolution in Soft-Shell Clam, Mya Arenaria, Bivalve Transmissible Neoplasia'. *Nature Cancer*, October. https://doi.org/10.1038/s43018-023-00643-7.
84. Hause, Ronald J., Colin C. Pritchard, Jay Shendure, and Stephen J. Salipante. 2016. 'Classification and Characterization of Microsatellite Instability across 18 Cancer Types'. *Nature Medicine* 22 (11): 1342–50. https://doi.org/10.1038/nm.4191.
85. Helman, Elena, Michael S. Lawrence, Chip Stewart, Carrie Sougnez, Gad Getz, and Matthew Meyerson. 2014. 'Somatic Retrotransposition in Human Cancer Revealed by Whole-Genome and Exome Sequencing'. *Genome Research* 24 (7): 1053–63. https://doi.org/10.1101/gr.163659.113.
86. Holmes, S. E., B. A. Dombroski, C. M. Krebs, C. D. Boehm, and H. H. Kazazian. 1994. 'A New Retrotransposable Human L1 Element from the LRE2 Locus on Chromosome 1q Produces a Chimaeric Insertion'. *Nature Genetics* 7 (2): 143–48. https://doi.org/10.1038/ng0694-143.
87. Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. 'ART: A next-Generation Sequencing Read Simulator'. *Bioinformatics* 28 (4): 593–94. https://doi.org/10.1093/bioinformatics/btr708.
88. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. 'Pan-Cancer Analysis of Whole Genomes'. *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.
89. Insua, Ana, and Catherine Thiriot-Quiévreux. 1992. 'Karyotypes of Cerastoderma Edule, Venerupis Pullastra and Venerupis Rhomboides (Bivalvia, Veneroida)'. *Aquatic Living Resources* 5 (1): 1–8. https://doi.org/10.1051/alr:1992001.
90. Iskow, Rebecca C., Michael T. McCabe, Ryan E. Mills, Spencer Torene, W. Stephen Pittard, Andrew F. Neuwald, Erwin G. Van Meir, Paula M. Vertino, and Scott E.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Devine. 2010. ‘Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons’. *Cell* 141 (7): 1253–61. https://doi.org/10.1016/j.cell.2010.05.020.
91. Ivancevic, Atma M., R. Daniel Kortschak, Terry Bertozzi, and David L. Adelson. 2016. ‘LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life’. *Genome Biology and Evolution* 8 (11): 3301–22. https://doi.org/10.1093/gbe/evw243.
92. Jia, Peng, Xiaofei Yang, Li Guo, Bowen Liu, Jiadong Lin, Hao Liang, Jianyong Sun, Chengsheng Zhang, and Kai Ye. 2020. ‘MSIsensor-pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability’. *Genomics, Proteomics & Bioinformatics* 18 (1): 65–71. https://doi.org/10.1016/j.gpb.2020.02.001.
93. Jiricny, Josef. 2006. ‘The Multifaceted Mismatch-Repair System’. *Nature Reviews. Molecular Cell Biology* 7 (5): 335–46. https://doi.org/10.1038/nrm1907.
94. John, Christopher R. 2020. ‘MLeval: Machine Learning Model Evaluation’. https://cran.r-project.org/web/packages/MLeval/index.html.
95. Judd, Julius, Hayley Sanderson, and Cédric Feschotte. 2021. ‘Evolution of Mouse Circadian Enhancers from Transposable Elements’. *Genome Biology* 22 (1): 193. https://doi.org/10.1186/s13059-021-02409-9.
96. Jung, Hyunchul, Jung Kyoon Choi, and Eunjung Alice Lee. 2018. ‘Immune Signatures Correlate with L1 Retrotransposition in Gastrointestinal Cancers’. *Genome Research* 28 (8): 1136–46. https://doi.org/10.1101/gr.231837.117.
97. Kather, Jakob Nikolas, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, et al. 2019. ‘Deep Learning Can Predict Microsatellite Instability Directly from Histology in Gastrointestinal Cancer’. *Nature Medicine* 25 (7): 1054–56. https://doi.org/10.1038/s41591-019-0462-y.
98. Kautto, Esko A., Russell Bonneville, Jharna Miya, Lianbo Yu, Melanie A. Krook, Julie W. Reeser, and Sameek Roychowdhury. 2016. ‘Performance Evaluation for Rapid Detection of Pan-Cancer Microsatellite Instability with MANTIS’. *Oncotarget* 8 (5): 7452–63. https://doi.org/10.18632/oncotarget.13918.
99. Kazazian, H. H., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis. 1988. ‘Haemophilia A Resulting from de Novo Insertion of L1 Sequences Represents a Novel Mechanism for Mutation in Man’. *Nature* 332 (6160): 164–66. https://doi.org/10.1038/332164a0.
100. Kazazian, Haig H. 2004. ‘Mobile Elements: Drivers of Genome Evolution’. *Science (New York, N.Y.)* 303 (5664): 1626–32. https://doi.org/10.1126/science.1089670.
101. Kazazian, Haig H., and John V. Moran. 2017. ‘Mobile DNA in Health and Disease’. *The New England Journal of Medicine* 377 (4): 361–70. https://doi.org/10.1056/NEJMra1510092.
102. Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. 2002. ‘The Human Genome Browser at UCSC’. *Genome Research* 12 (6): 996–1006. https://doi.org/10.1101/gr.229102.
103. Khan, Hameed, Arian Smit, and Stéphane Boissinot. 2006. ‘Molecular Evolution and Tempo of Amplification of Human LINE-1 Retrotransposons since the Origin of Primates’. *Genome Research* 16 (1): 78–87. https://doi.org/10.1101/gr.4001406.
104. Krakau, Manuela, Sabine Jacobsen, Kurt Thomas Jensen, and Karsten Reise. 2012. ‘The Cockle Cerastoderma Edule at Northeast Atlantic Shores: Genetic Signatures of Glacial Refugia’. *Marine Biology* 159 (1): 221–30. https://doi.org/10.1007/s00227-011-1802-8.
105. Kuhn, Max. 2008. ‘Building Predictive Models in R Using the Caret Package’. *Journal*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

of Statistical Software 28 (November):1–26. https://doi.org/10.18637/jss.v028.i05.
106. Lanciano, Sophie, Claude Philippe, Arpita Sarkar, David Pratella, Cécilia Domrane, Aurélien J. Doucet, Dominic van Essen, et al. 2024. 'Locus-Level L1 DNA Methylation Profiling Reveals the Epigenetic and Transcriptional Interplay between L1s and Their Integration Sites'. *Cell Genomics* 0 (0). https://doi.org/10.1016/j.xgen.2024.100498.
107. Lawson, Andrew R. J., Federico Abascal, Tim H. H. Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, et al. 2020. 'Extensive Heterogeneity in Somatic Mutation and Selection in the Human Bladder'. *Science (New York, N.Y.)* 370 (6512): 75–82. https://doi.org/10.1126/science.aba8347.
108. Le Grand, Fabienne, Edouard Kraffe, Xavier de Montaudouin, Antonio Villalba, Yanic Marty, and Philippe Soudant. 2010. 'Prevalence, Intensity, and Aneuploidy Patterns of Disseminated Neoplasia in Cockles (Cerastoderma Edule) from Arcachon Bay: Seasonal Variation and Position in Sediment'. *Journal of Invertebrate Pathology* 104 (2): 110–18. https://doi.org/10.1016/j.jip.2010.01.011.
109. 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 (New York, N.Y.)* 337 (6097): 967–71. https://doi.org/10.1126/science.1222077.
110. Li, Heng. 2018. 'Minimap2: Pairwise Alignment for Nucleotide Sequences'. *Bioinformatics (Oxford, England)* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
111. Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 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.
112. Li, Lin, Qiushi Feng, and Xiaosheng Wang. 2020. 'PreMSIm: An R Package for Predicting Microsatellite Instability from the Expression Profiling of a Gene Panel in Cancer'. *Computational and Structural Biotechnology Journal* 18:668–75. https://doi.org/10.1016/j.csbj.2020.03.007.
113. Lim, J. K., and M. J. Simmons. 1994. 'Gross Chromosome Rearrangements Mediated by Transposable Elements in Drosophila Melanogaster'. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 16 (4): 269–75. https://doi.org/10.1002/bies.950160410.
114. López Bruzos, Alicia. 2022. 'Evolution of Bivalve Transmissible Cancers'. Doctoral thesis. https://minerva.usc.es/xmlui/handle/10347/29342.
115. López, Saioa, Emilia L. Lim, Stuart Horswell, Kerstin Haase, Ariana Huebner, Michelle Dietzen, Thanos P. Mourikis, et al. 2020. 'Interplay between Whole Genome Doubling and the Accumulation of Deleterious Alterations in Cancer Evolution'. *Nature Genetics* 52 (3): 283–93. https://doi.org/10.1038/s41588-020-0584-7.
116. Lowe, D. M., and M. N. Moore. 1978. 'Cytology and Quantitative Cytochemistry of a Poliferative Atypical Hemocytic Condition in Mytilus Edulis (Bivalvia, Mollusca)'. *Journal of the National Cancer Institute* 60 (6): 1455–59. https://doi.org/10.1093/jnci/60.6.1455.
117. 'LUMPY: A Probabilistic Framework for Structural Variant Discovery - PubMed'. n.d. Accessed 3 February 2022. https://pubmed.ncbi.nlm.nih.gov/24970577/.
118. Lutz, Sheila M., Bethaney J. Vincent, Haig H. Kazazian Jr., Mark A. Batzer, and John V. Moran. 2003. 'Allelic Heterogeneity in LINE-1 Retrotransposition Activity'.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

American Journal of Human Genetics 73 (6): 1431–37.
119. Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. 'Structural Variant Calling: The Long and the Short of It'. *Genome Biology* 20 (1): 246. https://doi.org/10.1186/s13059-019-1828-7.
120. Marchani, Elizabeth E., Jinchuan Xing, David J. Witherspoon, Lynn B. Jorde, and Alan R. Rogers. 2009. 'Estimating the Age of Retrotransposon Subfamilies Using Maximum Likelihood'. *Genomics* 94 (1): 78–82. https://doi.org/10.1016/j.ygeno.2009.04.002.
121. Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schöenhuth, and Tobias Marschall. 2016. 'WhatsHap: Fast and Accurate Read-Based Phasing'. bioRxiv. https://doi.org/10.1101/085050.
122. Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2017. 'Universal Patterns of Selection in Cancer and Somatic Tissues'. *Cell* 171 (5): 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
123. Martínez, Luisa, Ruth Freire, Alberto Arias-Pérez, Josefina Méndez, and Ana Insua. 2015. 'Patterns of Genetic Variation across the Distribution Range of the Cockle Cerastoderma Edule Inferred from Microsatellites and Mitochondrial DNA'. *Marine Biology* 162 (7): 1393–1406. https://doi.org/10.1007/s00227-015-2676-y.
124. Martínez-Cardús, Anna, Sebastian Moran, Eva Musulen, Cátia Moutinho, Jose L. Manzano, Eva Martinez-Balibrea, Montserrat Tierno, et al. 2016. 'Epigenetic Homogeneity Within Colorectal Tumors Predicts Shorter Relapse-Free and Overall Survival Times for Patients With Locoregional Cancer'. *Gastroenterology* 151 (5): 961–72. https://doi.org/10.1053/j.gastro.2016.08.001.
125. Martínez-Jiménez, Francisco, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, et al. 2020. 'A Compendium of Mutational Cancer Driver Genes'. *Nature Reviews. Cancer* 20 (10): 555–72. https://doi.org/10.1038/s41568-020-0290-x.
126. Martín-López, Juana V., and Richard Fishel. 2013. 'The Mechanism of Mismatch Repair and the Functional Analysis of Mismatch Repair Defects in Lynch Syndrome'. *Familial Cancer* 12 (2): 159–68. https://doi.org/10.1007/s10689-013-9635-x.
127. Matias, A. M., S. Diaz, J. Teixeira de Sousa, A. Villalba, M. J. Carballal, and A. Leitão. 2014. 'Karyotype Variation in Neoplastic Cells Associated to Severity of Disseminated Neoplasia in the Cockle Cerastoderma Edule'. *Aquaculture* 428–429:223–25. https://doi.org/10.1016/j.aquaculture.2014.03.017.
128. Mattesen, Trine B., Mads H. Rasmussen, Juan Sandoval, Halit Ongen, Sigrid S. Árnadóttir, Josephine Gladov, Anna Martinez-Cardus, et al. 2020. 'MethCORR Modelling of Methylomes from Formalin-Fixed Paraffin-Embedded Tissue Enables Characterization and Prognostication of Colorectal Cancer'. *Nature Communications* 11 (1): 2025. https://doi.org/10.1038/s41467-020-16000-6.
129. Merlo, Lauren M. F., Li-San Wang, John W. Pepper, Peter S. Rabinovitch, and Carlo C. Maley. 2010. 'Polyploidy, Aneuploidy and the Evolution of Cancer'. *Advances in Experimental Medicine and Biology* 676:1–13. https://doi.org/10.1007/978-1-4419-6199-0_1.
130. Metzger, Michael J., Carol Reinisch, James Sherry, and Stephen P. Goff. 2015. 'Horizontal Transmission of Clonal Cancer Cells Causes Leukemia in Soft-Shell Clams'. *Cell* 161 (2): 255–63. https://doi.org/10.1016/j.cell.2015.02.042.
131. Metzger, Michael J., Antonio Villalba, María J. Carballal, David Iglesias, James Sherry,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Carol Reinisch, Annette F. Muttray, Susan A. Baldwin, and Stephen P. Goff. 2016. 'Widespread Transmission of Independent Cancer Lineages within Multiple Bivalve Species'. *Nature* 534 (7609): 705–9. https://doi.org/10.1038/nature18599.

132. Michnowska, Alicja, Samuel F. M. Hart, Katarzyna Smolarz, Anna Hallmann, and Michael J. Metzger. 2022. 'Horizontal Transmission of Disseminated Neoplasia in the Widespread Clam Macoma Balthica from the Southern Baltic Sea'. *Molecular Ecology* 31 (11): 3128–36. https://doi.org/10.1111/mec.16464.
133. Miki, Y., I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya, K. W. Kinzler, B. Vogelstein, and Y. Nakamura. 1992. 'Disruption of the APC Gene by a Retrotransposal Insertion of L1 Sequence in a Colon Cancer'. *Cancer Research* 52 (3): 643–45.
134. Miroshnychenko, Daria, Etienne Baratchart, Meghan C. Ferrall-Fairbanks, Robert Vander Velde, Mark A. Laurie, Marilyn M. Bui, Aik Choon Tan, Philipp M. Altrock, David Basanta, and Andriy Marusyk. 2021. 'Spontaneous Cell Fusions as a Mechanism of Parasexual Recombination in Tumour Cell Populations'. *Nature Ecology & Evolution* 5 (3): 379–91. https://doi.org/10.1038/s41559-020-01367-y.
135. Mita, Paolo, Aleksandra Wudzinska, Xiaoji Sun, Joshua Andrade, Shruti Nayak, David J. Kahler, Sana Badri, et al. 2018. 'LINE-1 Protein Localization and Functional Dynamics during the Cell Cycle'. *eLife* 7 (January):e30058. https://doi.org/10.7554/eLife.30058.
136. Moran, J. V., R. J. DeBerardinis, and 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.
137. Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke, and H. H. Kazazian. 1996. 'High Frequency Retrotransposition in Cultured Mammalian Cells'. *Cell* 87 (5): 917–27. https://doi.org/10.1016/s0092-8674(00)81998-4.
138. Morris, Tiffany J., Lee M. Butcher, Andrew Feber, Andrew E. Teschendorff, Ankur R. Chakravarthy, Tomasz K. Wojdacz, and Stephan Beck. 2014. 'ChAMP: 450k Chip Analysis Methylation Pipeline'. *Bioinformatics (Oxford, England)* 30 (3): 428–30. https://doi.org/10.1093/bioinformatics/btt684.
139. Muehlenbachs, Atis, Blaine A. Mathison, and Peter D. Olson. 2016. 'Malignant Transformation of Hymenolepis Nana in a Human Host'. *The New England Journal of Medicine* 374 (13): 1293–94. https://doi.org/10.1056/NEJMc1600490.
140. Murchison, Elizabeth P. 2016. 'Cancer: Transmissible Tumours under the Sea'. *Nature* 534 (7609): 628–29. https://doi.org/10.1038/nature18455.
141. Muttray, Annette F., Teagan F. O'Toole, Wendy Morrill, Rebecca J. Van Beneden, and Susan A. Baldwin. 2010. 'An Invertebrate Mdm Homolog Interacts with P53 and Is Differentially Expressed Together with P53 and Ras in Neoplastic Mytilus Trossulus Haemocytes'. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology* 156 (4): 298–308. https://doi.org/10.1016/j.cbpb.2010.04.008.
142. Muttray, Annette F., Patricia M. Schulte, and Susan A. Baldwin. 2008. 'Invertebrate P53-like mRNA Isoforms Are Differentially Expressed in Mussel Haemic Neoplasia'. *Marine Environmental Research* 66 (4): 412–21. https://doi.org/10.1016/j.marenvres.2008.06.004.
143. Nam, Chang Hyun, Jeonghwan Youk, Jeong Yeon Kim, Joonoh Lim, Jung Woo Park, Soo A. Oh, Hyun Jung Lee, et al. 2023. 'Widespread Somatic L1 Retrotransposition in Normal Colorectal Epithelium'. *Nature* 617 (7961): 540–47. https://doi.org/10.1038/s41586-023-06046-z.
144. Namba, Shinichi, Kazuhito Sato, Shinya Kojima, Toshihide Ueno, Yoko Yamamoto,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Yosuke Tanaka, Satoshi Inoue, et al. 2019. 'Differential Regulation of CpG Island Methylation within Divergent and Unidirectional Promoters in Colorectal Cancer'. *Cancer Science* 110 (3): 1096–1104. https://doi.org/10.1111/cas.13937.

145. Nassar, Luis R., Galt P. Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, et al. 2023. 'The UCSC Genome Browser Database: 2023 Update'. *Nucleic Acids Research* 51 (D1): D1188–95. https://doi.org/10.1093/nar/gkac1072.
146. Newman, Aaron M., Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, et al. 2019. 'Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry'. *Nature Biotechnology* 37 (7): 773–82. https://doi.org/10.1038/s41587-019-0114-2.
147. Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies'. *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.
148. Ní Leathlobhair, Máire, Angela R. Perri, Evan K. Irving-Pease, Kelsey E. Witt, Anna Linderholm, James Haile, Ophelie Lebrasseur, et al. 2018. 'The Evolutionary History of Dogs in the Americas'. *Science (New York, N.Y.)* 361 (6397): 81–85. https://doi.org/10.1126/science.aao4776.
149. Niu, Beifang, Kai Ye, Qunyuan Zhang, Charles Lu, Mingchao Xie, Michael D. McLellan, Michael C. Wendl, and Li Ding. 2014. 'MSIsensor: Microsatellite Instability Detection Using Paired Tumor-Normal Sequence Data'. *Bioinformatics (Oxford, England)* 30 (7): 1015–16. https://doi.org/10.1093/bioinformatics/btt755.
150. Ono, Yukiteru, Kiyoshi Asai, and Michiaki Hamada. 2013. 'PBSIM: PacBio Reads Simulator--toward Accurate Genome Assembly'. *Bioinformatics (Oxford, England)* 29 (1): 119–21. https://doi.org/10.1093/bioinformatics/bts649.
151. Ostertag, E. M., and H. 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.
152. Ostrander, Elaine A., Brian W. Davis, and Gary K. Ostrander. 2016. 'Transmissible Tumors: Breaking the Cancer Paradigm'. *Trends in Genetics: TIG* 32 (1): 1–15. https://doi.org/10.1016/j.tig.2015.10.001.
153. Papaemmanuil, Elli, Inmaculada Rapado, Yilong Li, Nicola E. Potter, David C. Wedge, Jose Tubio, Ludmil B. Alexandrov, et al. 2014. 'RAG-Mediated Recombination Is the Predominant Driver of Oncogenic Rearrangement in ETV6-RUNX1 Acute Lymphoblastic Leukemia'. *Nature Genetics* 46 (2): 116–25. https://doi.org/10.1038/ng.2874.
154. Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. 'APE: Analyses of Phylogenetics and Evolution in R Language'. *Bioinformatics (Oxford, England)* 20 (2): 289–90. https://doi.org/10.1093/bioinformatics/btg412.
155. Peteiro, Laura G., Sarah A. Woodin, David S. Wethey, Damian Costas-Costas, Arantxa Martínez-Casal, Celia Olabarria, and Elsa Vázquez. 2018. 'Responses to Salinity Stress in Bivalves: Evidence of Ontogenetic Changes in Energetic Physiology on Cerastoderma Edule'. *Scientific Reports* 8 (1): 8329. https://doi.org/10.1038/s41598-018-26706-9.
156. Platt, Roy N., Michael W. Vandewege, and David A. Ray. 2018. 'Mammalian Transposable Elements and Their Impacts on Genome Evolution'. *Chromosome Research* 26 (1): 25–43. https://doi.org/10.1007/s10577-017-9570-z.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

157. Porubsky, David, Wolfram Höps, Hufsah Ashraf, PingHsun Hsieh, Bernardo Rodriguez-Martin, Feyza Yilmaz, Jana Ebler, et al. 2022. 'Recurrent Inversion Polymorphisms in Humans Associate with Genetic Instability and Genomic Disorders'. *Cell* 185 (11): 1986-2005.e26. https://doi.org/10.1016/j.cell.2022.04.017.
158. Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. 2012. 'DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis'. *Bioinformatics (Oxford, England)* 28 (18): i333–39. https://doi.org/10.1093/bioinformatics/bts378.
159. Rebbeck, Clare A., Armand M. Leroi, and Austin Burt. 2011. 'Mitochondrial Capture by a Transmissible Cancer'. *Science (New York, N.Y.)* 331 (6015): 303. https://doi.org/10.1126/science.1197696.
160. Ren, Xu, and Pei Fen Kuan. 2019. 'methylGSA: A Bioconductor Package and Shiny App for DNA Methylation Data Length Bias Adjustment in Gene Set Testing'. *Bioinformatics (Oxford, England)* 35 (11): 1958–59. https://doi.org/10.1093/bioinformatics/bty892.
161. Reno, Paul W., Marcia House, and Andrea Illingworth. 1994. 'Flow Cytometric and Chromosome Analysis of Softshell Clams, Mya Arenaria, with Disseminated Neoplasia'. *Journal of Invertebrate Pathology* 64 (3): 163–72. https://doi.org/10.1016/S0022-2011(94)90025-6.
162. Rizvi, Naiyer A., Matthew D. Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J. Havel, William Lee, et al. 2015. 'Cancer Immunology. Mutational Landscape Determines Sensitivity to PD-1 Blockade in Non-Small Cell Lung Cancer'. *Science (New York, N.Y.)* 348 (6230): 124–28. https://doi.org/10.1126/science.aaa1348.
163. 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.
164. Romalde, Jesús L., M. Luz Vilariño, Roxana Beaz, José M. Rodríguez, Seila Díaz, Antonio Villalba, and María J. Carballal. 2007. 'Evidence of Retroviral Etiology for Disseminated Neoplasia in Cockles (Cerastoderma Edule)'. *Journal of Invertebrate Pathology* 94 (2): 95–101. https://doi.org/10.1016/j.jip.2006.09.007.
165. Ruan, Jue, and Heng Li. 2020. 'Fast and Accurate Long-Read Assembly with Wtdbg2'. *Nature Methods* 17 (2): 155–58. https://doi.org/10.1038/s41592-019-0669-3.
166. Ruiz, Pamela, Seila Díaz, Amaia Orbea, Maria J. Carballal, Antonio Villalba, and Miren P. Cajaraville. 2013. 'Biomarkers and Transcription Levels of Cancer-Related Genes in Cockles Cerastoderma Edule from Galicia (NW Spain) with Disseminated Neoplasia'. *Aquatic Toxicology (Amsterdam, Netherlands)* 136–137 (July):101–11. https://doi.org/10.1016/j.aquatox.2013.03.018.
167. Ruiz-Bañobre, Juan, and Ajay Goel. 2019. 'DNA Mismatch Repair Deficiency and Immune Checkpoint Inhibitors in Gastrointestinal Cancers'. *Gastroenterology* 156 (4): 890–903. https://doi.org/10.1053/j.gastro.2018.11.071.
168. ———. 2021. 'Genomic and Epigenomic Biomarkers in Colorectal Cancer: From Diagnosis to Therapy'. *Advances in Cancer Research* 151:231–304. https://doi.org/10.1016/bs.acr.2021.02.008.
169. Saavedra, Carlos, and Evelyne Bachère. 2006. 'Bivalve Genomics'. *Aquaculture* 256 (1): 1–14. https://doi.org/10.1016/j.aquaculture.2006.02.023.
170. Saillard, Charlie, Rémy Dubois, Oussama Tchita, Nicolas Loiseau, Thierry Garcia,

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Aurélie Adriansen, Séverine Carpentier, et al. 2023. 'Validation of MSIntuit as an AI-Based Pre-Screening Tool for MSI Detection from Colorectal Cancer Histology Slides'. *Nature Communications* 14 (1): 6695. https://doi.org/10.1038/s41467-023-42453-6.

171. Salem, Abdel Halim, Jeremy S. Myers, Anthony C. Otieno, W. Scott Watkins, Lynn B. Jorde, and Mark A. Batzer. 2003. 'LINE-1 preTa Elements in the Human Genome'. *Journal of Molecular Biology* 326 (4): 1127–46. https://doi.org/10.1016/s0022-2836(03)00032-9.
172. Salipante, Stephen J., Sheena M. Scroggins, Heather L. Hampel, Emily H. Turner, and Colin C. Pritchard. 2014. 'Microsatellite Instability Detection by next Generation Sequencing'. *Clinical Chemistry* 60 (9): 1192–99. https://doi.org/10.1373/clinchem.2014.223677.
173. Salvatori, Silvia, Irene Marafini, Federica Laudisi, Giovanni Monteleone, and Carmine Stolfi. 2023. 'Helicobacter Pylori and Gastric Cancer: Pathogenetic Mechanisms'. *International Journal of Molecular Sciences* 24 (3): 2895. https://doi.org/10.3390/ijms24032895.
174. Sawaya, Sterling, Andrew Bagshaw, Emmanuel Buschiazzo, Pankaj Kumar, Shantanu Chowdhury, Michael A. Black, and Neil Gemmell. 2013. 'Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements'. *PLoS ONE* 8 (2): e54710. https://doi.org/10.1371/journal.pone.0054710.
175. Schliep, Klaus Peter. 2011. 'Phangorn: Phylogenetic Analysis in R'. *Bioinformatics (Oxford, England)* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.
176. Schlötterer, C. 2000. 'Evolutionary Dynamics of Microsatellite DNA'. *Chromosoma* 109 (6): 365–71. https://doi.org/10.1007/s004120000089.
177. Scott, Emma C., and Scott E. Devine. 2017. 'The Role of Somatic L1 Retrotransposition in Human Cancers'. *Viruses* 9 (6): 131. https://doi.org/10.3390/v9060131.
178. Scott, Emma C., Eugene J. Gardner, Ashiq Masood, Nelson T. Chuang, Paula M. Vertino, and Scott E. Devine. 2016. 'A Hot L1 Retrotransposon Evades Somatic Repression and Initiates Human Colorectal Cancer'. *Genome Research* 26 (6): 745–55. https://doi.org/10.1101/gr.201814.115.
179. Shankaranarayana, Akshatha Handattu, Bhagyalalitha Meduri, Gurubasavaraj Veeranna Pujar, Raghu Chandrashekar Hariharapura, Arun Kumar Sethu, Manisha Singh, and Durgesh Bidye. 2023. 'Restoration of P53 Functions by Suppression of Mortalin-P53 Sequestration: An Emerging Target in Cancer Therapy'. *Future Medicinal Chemistry* 15 (22): 2087–2112. https://doi.org/10.4155/fmc-2023-0061.
180. Shen, Zhiyuan. 2011. 'Genomic Instability and Cancer: An Introduction'. *Journal of Molecular Cell Biology* 3 (1): 1–3. https://doi.org/10.1093/jmcb/mjq057.
181. Shikano, Takahito, Jetty Ramadevi, Yukinori Shimada, and Juha Merilä. 2010. 'Utility of Sequenced Genomes for Microsatellite Marker Development in Non-Model Organisms: A Case Study of Functionally Important Genes in Nine-Spined Sticklebacks (Pungitius Pungitius)'. *BMC Genomics* 11 (May):334. https://doi.org/10.1186/1471-2164-11-334.
182. Skazina, Maria, Nelly Odintsova, Mariia Maiorova, Lidia Frolova, Irina Dolganova, Kira Regel, and Petr Strelkov. 2023. 'Two Lineages of Bivalve Transmissible Neoplasia Affect the Blue Mussel Mytilus Trossulus Gould in the Subarctic Sea of Okhotsk'. *Current Zoology* 69 (1): 91–102. https://doi.org/10.1093/cz/zoac012.
183. Smith, T. F., and M. S. Waterman. 1981. 'Identification of Common Molecular Subsequences'. *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

184. Snyder, Alexandra, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M. Zaretsky, Alexis Desrichard, Logan A. Walsh, et al. 2014. 'Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma'. *The New England Journal of Medicine* 371 (23): 2189–99. https://doi.org/10.1056/NEJMoa1406498.
185. Solyom, Szilvia, Adam D. Ewing, Eric P. Rahrmann, Tara Doucet, Heather H. Nelson, Michael B. Burns, Reuben S. Harris, et al. 2012. 'Extensive Somatic L1 Retrotransposition in Colorectal Tumors'. *Genome Research* 22 (12): 2328–38. https://doi.org/10.1101/gr.145235.112.
186. Sondka, Zbyslaw, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. 2018. 'The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers'. *Nature Reviews. Cancer* 18 (11): 696–705. https://doi.org/10.1038/s41568-018-0060-1.
187. Sondka, Zbyslaw, Nidhi Bindal Dhir, Denise Carvalho-Silva, Steven Jupe, Madhumita, Karen McLaren, Mike Starkey, et al. 2024. 'COSMIC: A Curated Database of Somatic Variants and Clinical Data for Cancer'. *Nucleic Acids Research* 52 (D1): D1210–17. https://doi.org/10.1093/nar/gkad986.
188. Song, Mingzhou, and Stéphane Boissinot. 2007. 'Selection against LINE-1 Retrotransposons Results Principally from Their Ability to Mediate Ectopic Recombination'. *Gene*, ASILOMAR 2006, 390 (1): 206–13. https://doi.org/10.1016/j.gene.2006.09.033.
189. Stammnitz, Maximilian R., Tim H. H. Coorens, Kevin C. Gori, Dane Hayes, Beiyuan Fu, Jinhong Wang, Daniel E. Martin-Herranz, et al. 2018. 'The Origins and Vulnerabilities of Two Transmissible Cancers in Tasmanian Devils'. *Cancer Cell* 33 (4): 607-619.e15. https://doi.org/10.1016/j.ccell.2018.03.013.
190. Stammnitz, Maximilian R., Kevin Gori, Young Mi Kwon, Edward Harry, Fergal J. Martin, Konstantinos Billis, Yuanyuan Cheng, et al. 2023. 'The Evolution of Two Transmissible Cancers in Tasmanian Devils'. *Science (New York, N.Y.)* 380 (6642): 283–93. https://doi.org/10.1126/science.abq6453.
191. Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, et al. 2011. 'Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development'. *Cell* 144 (1): 27–40. https://doi.org/10.1016/j.cell.2010.11.055.
192. Storchova, Zuzana, and David Pellman. 2004. 'From Polyploidy to Aneuploidy, Genome Instability and Cancer'. *Nature Reviews. Molecular Cell Biology* 5 (1): 45–54. https://doi.org/10.1038/nrm1276.
193. Strakova, Andrea, Thomas J. Nicholls, Adrian Baez-Ortega, Máire Ní Leathlobhair, Alexander T. Sampson, Katherine Hughes, Isobelle A. G. Bolton, et al. 2020. 'Recurrent Horizontal Transfer Identifies Mitochondrial Positive Selection in a Transmissible Cancer'. *Nature Communications* 11 (1): 3059. https://doi.org/10.1038/s41467-020-16765-w.
194. Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. 'The Cancer Genome'. *Nature* 458 (7239): 719–24. https://doi.org/10.1038/nature07943.
195. Sultana, Tania, Alessia Zamborlini, Gael Cristofari, and Pascale Lesage. 2017. 'Integration Site Selection by Retroviruses and Transposable Elements in Eukaryotes'. *Nature Reviews. Genetics* 18 (5): 292–308. https://doi.org/10.1038/nrg.2017.7.
196. Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

185 Countries'. *CA: A Cancer Journal for Clinicians* 71 (3): 209–49. https://doi.org/10.3322/caac.21660.
197. Tolar, Jakub, and Joseph P. Neglia. 2003. 'Transplacental and Other Routes of Cancer Transmission between Individuals'. *Journal of Pediatric Hematology/Oncology* 25 (6): 430–34. https://doi.org/10.1097/00043426-200306000-00002.
198. Tsakou-Ngouafo, Louis, Julien Paganini, Jim Kaufman, and Pierre Pontarotti. 2020. 'Origins of the RAG Transposome and the MHC'. *Trends in Immunology* 41 (7): 561–71. https://doi.org/10.1016/j.it.2020.05.002.
199. Tubio, Jose M. C. 2015. 'Somatic Structural Variation and Cancer'. *Briefings in Functional Genomics* 14 (5): 339–51. https://doi.org/10.1093/bfgp/elv016.
200. 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.
201. Twomey, Eamonn, and Maire F. Mulcahy. 1984. 'A Proliferative Disorder of Possible Hemic Origin in the Common Cockle, Cerastoderma Edule'. *Journal of Invertebrate Pathology* 44 (1): 109–11. https://doi.org/10.1016/0022-2011(84)90052-1.
202. Twomey, Eamonn, and Máire F. Mulcahy. 1988. 'Epizootiological Aspects of a Sarcoma in the Cockle Cerastoderma Edule'. *Diseases of Aquatic Organisms* 5:225–38.
203. Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. 'Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads'. *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.
204. Vassilenko, Ekaterina, and Susan A. Baldwin. 2014. 'Using Flow Cytometry to Detect Haemic Neoplasia in Mussels (Mytilus Trossulus) from the Pacific Coast of Southern British Columbia, Canada'. *Journal of Invertebrate Pathology* 117 (March):68–72. https://doi.org/10.1016/j.jip.2014.02.002.
205. Vera, Manuel, Francesco Maroso, Sophie B. Wilmes, Miguel Hermida, Andrés Blanco, Carlos Fernández, Emily Groves, et al. 2022. 'Genomic Survey of Edible Cockle (Cerastoderma Edule) in the Northeast Atlantic: A Baseline for Sustainable Management of Its Wild Resources'. *Evolutionary Applications* 15 (2): 262–85. https://doi.org/10.1111/eva.13340.
206. Walker, Charles, Stefanie Böttger, and Ben Low. 2006. 'Mortalin-Based Cytoplasmic Sequestration of P53 in a Nonmammalian Cancer Model'. *The American Journal of Pathology* 168 (5): 1526–30. https://doi.org/10.2353/ajpath.2006.050603.
207. Walker, Charles W., Rebecca J. Van Beneden, Annette F. Muttray, S. Anne Böttger, Melissa L. Kelley, Abraham E. Tucker, and W. Kelley Thomas. 2011. 'P53 Superfamily Proteins in Marine Bivalve Cancer and Stress Biology'. *Advances in Marine Biology* 59:1–36. https://doi.org/10.1016/B978-0-12-385536-7.00001-7.
208. Wang, Yunhao, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. 2021. 'Nanopore Sequencing Technology, Bioinformatics and Applications'. *Nature Biotechnology* 39 (11): 1348–65. https://doi.org/10.1038/s41587-021-01108-x.
209. Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. 'Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench'. *Bioinformatics (Oxford, England)* 25 (9): 1189–91. https://doi.org/10.1093/bioinformatics/btp033.
210. Wei, W., N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. D. Boeke, and J. V. Moran. 2001. 'Human L1 Retrotransposition: Cis Preference versus Trans

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Complementation'. *Molecular and Cellular Biology* 21 (4): 1429–39. https://doi.org/10.1128/MCB.21.4.1429-1439.2001.

211. Weisenberger, Daniel J., Kimberly D. Siegmund, Mihaela Campan, Joanne Young, Tiffany I. Long, Mark A. Faasse, Gyeong Hoon Kang, et al. 2006. 'CpG Island Methylator Phenotype Underlies Sporadic Microsatellite Instability and Is Tightly Associated with BRAF Mutation in Colorectal Cancer'. *Nature Genetics* 38 (7): 787–93. https://doi.org/10.1038/ng1834.
212. Xia, Bo, Weimin Zhang, Aleksandra Wudzinska, Emily Huang, Ran Brosh, Maayan Pour, Alexander Miller, et al. 2021. 'The Genetic Basis of Tail-Loss Evolution in Humans and Apes'. bioRxiv. https://doi.org/10.1101/2021.09.14.460388.
213. Yi, Kijong, and Young Seok Ju. 2018. 'Patterns and Mechanisms of Structural Variations in Human Cancer'. *Experimental & Molecular Medicine* 50 (8): 1–11. https://doi.org/10.1038/s12276-018-0112-3.
214. Yoder, J. A., C. P. Walsh, and T. H. Bestor. 1997. 'Cytosine Methylation and the Ecology of Intragenomic Parasites'. *Trends in Genetics: TIG* 13 (8): 335–40. https://doi.org/10.1016/s0168-9525(97)01181-5.
215. Yonemitsu, Marisa A., Rachael M. Giersch, Maria Polo-Prieto, Maurine Hammel, Alexis Simon, Florencia Cremonte, Fernando T. Avilés, et al. 2019. 'A Single Clonal Lineage of Transmissible Cancer Identified in Two Marine Mussel Species in South America and Europe'. *eLife* 8 (November):e47788. https://doi.org/10.7554/eLife.47788.
216. Zumalave Duro, Sonia. 2024. 'The Dynamics of Somatic Retrotransposition in Human Cancer'. Doctoral thesis. https://minerva.usc.es/xmlui/handle/10347/33039.
217. Zumalave, Sonia, Martin Santamarina, Nuria P. Espasandín, Daniel Garcia-Souto, Javier Temes, Toby M. Baker, Ana Pequeño-Valtierra, et al. 2024. 'Synchronous L1 Retrotransposition Events Promote Chromosomal Crossover Early in Human Tumorigenesis'. bioRxiv. https://doi.org/10.1101/2024.08.27.596794.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

APPENDIX

Appendix A: List of publications

Manuscript 1

Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Science 372, 6537 (2021). DOI: https://doi.org/10.1126/science.abf7117

Authors
Peter Ebert [# 1], Peter A Audano [# 2], Qihui Zhu [# 3], Bernardo Rodriguez-Martin [# 4], David Porubsky [2], Marc Jan Bonder [4 5], Arvis Sulovari [2], Jana Ebler [1], Weichen Zhou [6], Rebecca Serra Mari [1], Feyza Yilmaz [3], Xuefang Zhao [7 8], PingHsun Hsieh [2], Joyce Lee [9], Sushant Kumar [10], Jiadong Lin [11], Tobias Rausch [4], Yu Chen [12], Jingwen Ren [13], Martin Santamarina [14 15], Wolfram Höps [4], Hufsah Ashraf [1], Nelson T Chuang [16], Xiaofei Yang [17], Katherine M Munson [2], Alexandra P Lewis [2], Susan Fairley [18], Luke J Tallon [16], Wayne E Clarke [19], Anna O Basile [19], Marta Byrska-Bishop [19], André Corvelo [19], Uday S Evani [19], Tsung-Yu Lu [13], Mark J P Chaisson [13], Junjie Chen [20], Chong Li [20], Harrison Brand [7 8], Aaron M Wenger [21], Maryam Ghareghani [22 23 1], William T Harvey [2], Benjamin Raeder [4], Patrick Hasenfeld [4], Allison A Regier [24], Haley J Abel [24], Ira M Hall [25], Paul Flicek [18], Oliver Stegle [4 5], Mark B Gerstein [10], Jose M C Tubio [14 15], Zepeng Mu [26], Yang I Li [27], Xinghua Shi [20], Alex R Hastie [9], Kai Ye [11 28], Zechen Chong [12], Ashley D Sanders [4], Michael C Zody [19], Michael E Talkowski [7 8], Ryan E Mills [6 28], Scott E Devine [16], Charles Lee [# 29 30 31], Jan O Korbel [# 32 18], Tobias Marschall [# 33], Evan E Eichler [# 34 35]

Affiliations
[1] Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Moorenstraße 20, 40225 Düsseldorf, Germany.
[2] Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Avenue NE, Seattle, WA 98195-5065, USA.
[3] The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA.
[4] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany.
[5] Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.
[6] Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA.
[7] Center for Genomic Medicine, Massachusetts General Hospital, Department of Neurology, Harvard Medical School, Boston, MA 02114, USA.
[8] Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
[9] Bionano Genomics, San Diego, CA 92121, USA.
[10] Program in Computational Biology and Bioinformatics, Yale University, BASS 432 and 437, 266 Whitney Avenue, New Haven, CT 06520, USA.
[11] School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.

USC DE SANTIAGO DE COMPOSTELA

[12] Department of Genetics and Informatics Institute, School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA.
[13] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA.
[14] Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[15] Department of Zoology, Genetics, and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[16] Institute for Genome Sciences, University of Maryland School of Medicine, 670 W Baltimore Street, Baltimore, MD 21201, USA.
[17] School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.
[18] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[19] New York Genome Center, New York, NY 10013, USA.
[20] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA.
[21] Pacific Biosciences of California, Menlo Park, CA 94025, USA.
[22] Max Planck Institute for Informatics, Saarland Informatics Campus E1.4, 66123 Saarbrücken, Germany.
[23] Saarbrücken Graduate School of Computer Science, Saarland University, Saarland Informatics Campus E1.3, 66123 Saarbrücken, Germany.
[24] Department of Medicine, Washington University, St. Louis, MO 63108, USA.
[25] Department of Genetics, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA.
[26] Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL 60637, USA.
[27] Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL 60637, USA.
[28] Department of Human Genetics, University of Michigan, 1241 E. Catherine Street, Ann Arbor, MI 48109, USA.
[29] The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA. eee@gs.washington.edu tobias.marschall@hhu.de jan.korbel@embl.org charles.lee@jax.org.
[30] Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, 277 West Yanta Road, Xi'an, 710061, Shaanxi, China.
[31] Department of Graduate Studies-Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul 120-750, South Korea.
[32] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany. eee@gs.washington.edu tobias.marschall@hhu.de jan.korbel@embl.org charles.lee@jax.org.
[33] Heinrich Heine University, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Moorenstraße 20, 40225 Düsseldorf, Germany. eee@gs.washington.edu tobias.marschall@hhu.de jan.korbel@embl.org charles.lee@jax.org.
[34] Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Avenue NE, Seattle, WA 98195-5065, USA. eee@gs.washington.edu tobias.marschall@hhu.de jan.korbel@embl.org charles.lee@jax.org.
[35] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
[#] Contributed equally.

USC
DE COMPOSTELA

Specific contributions to the publication
I contributed to the section of the manuscript concerning mobile element insertions (MEIs). My major contribution consisted in the estimation of L1Hs phylogenies based on the reconstructed sequences of active FL-L1 elements. I generated multiple sequence alignments based on polished FL-L1 sequences and compared alternative substitution models to optimize ML tree inferences. I developed and implemented code to systematically assess the integrity of L1 open reading frames (ORF1 and ORF2), determine L1Hs subfamilies (*pre-Ta*, *Ta-0* and *Ta-1*) and obtain L1 age estimations based on sequence divergence. I performed statistical analyses assessing L1Hs subfamily age distributions. I analysed the phylogenetic distribution of L1 source elements with somatic activity in cancer. I created circos plots to visualize the genomic distribution of retrotransposition events. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work.

Main text:
Section "Mobile element insertions" Figure 3 (A, B, D)

Supplementary text:
- Figure S17, S18, S19, S20.

Quality metrics
The journal Science has an impact factor of 56.9 (2022 Journal Citation Reports), a Scopus CiteScore of 59.0 (26th of December 2023) and is in the first quartile (Q1) in Multidisciplinary according to Scimago (SJR 2022 13.3).

Journal authorization
This is an open access article under the AAAS Author License to Publish Policy.
Link to the AAAS Author License to Publish is provided **here**

Manuscript 2

Somatic evolution of marine transmissible leukemias in the common cockle, Cerastoderma edule
Nature cancer 4, 1575-1591 (2023). DOI: https://doi.org/10.1038/s43018-023-00641-9

Authors
Alicia L Bruzos [#1 2 3], Martín Santamarina [#1 2 3], Daniel García-Souto [#1 2 3 4], Seila Díaz [#1 5], Sara Rocha [#6], Jorge Zamora [1], Yunah Lee [7], Alejandro Viña-Feás [1 3], Michael A Quail [4], Iago Otero [1 2 3], Ana Pequeño-Valtierra [1 3], Javier Temes [1 3], Jorge Rodriguez-Castro [1 3], Leyre Aramburu [1], André Vidal-Capón [8], Antonio Villanueva [9], Damián Costas [9], Rosana Rodríguez [9], Tamara Prieto [6 10 11], Laura Tomás [6 10], Pilar Alvariño [6 10], Juana Alonso [6 10], Asunción Cao [12], David Iglesias [12], María J Carballal [12], Ana M Amaral [13], Pablo Balseiro [14 15], Ricardo Calado [5], Bouchra El Khalfi [16], Urtzi Izagirre [17 18], Xavier de Montaudouin [19], Nicolas G Pade [20], Ian Probert [21], Fernando Ricardo [5], Pamela Ruiz [17 18], Maria Skazina [22], Katarzyna Smolarz [23], Juan J Pasantes [8 9], Antonio Villalba [12 17 24], Zemin Ning [4], Young Seok Ju [7], David Posada [6 8 10], Jonas Demeulemeester [25 26 27], Adrian Baez-Ortega [28 29], Jose M C Tubio [30 31 32]

Affiliations
Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[2] Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
[3] Instituto de Investigaciones Sanitarias de Santiago de Compostela (IDIS), Santiago de Compostela, Spain.
[4] Wellcome Sanger Institute, Hinxton, UK.
[5] ECOMARE, Centre for Environmental and Marine Studies (CESAM) & Department of Biology, University of Aveiro, Aveiro, Portugal.
[6] CINBIO, Universidade de Vigo, Vigo, Spain.
[7] Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea.
[8] Department of Biochemistry, Genetics and Immunology, Universidade de Vigo, Vigo, Spain.
[9] Centro de Investigación Mariña (CIM-ECIMAT), Universidade de Vigo, Vigo, Spain.
[10] Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain.
[11] New York Genome Center, New York, NY, USA.
[12] Centro de Investigacións Mariñas (CIMA), Consellería do Mar, Xunta de Galicia, Vilanova de Arousa, Spain.
[13] Centro de Ciencias do Mar do Algarve (CCMAR), Universidade do Algarve, Faro, Portugal.
[14] Department of Biological Sciences, University of Bergen, Bergen, Norway.
[15] NORCE AS, Bergen, Norway.
[16] Laboratory of Physiopathology, Molecular Genetics & Biotechnology, Faculty of Sciences Ain Chock, Health and Biotechnology Research Centre, Hassan II University of Casablanca, Casablanca, Morocco.
[17] Research Centre for Experimental Marine Biology and Biotechnology (PiE-UPV/EHU), University of the Basque Country (UPV/EHU), Plenzia-Bitzkaia, Spain.
[18] Cell Biology in Environmental Toxicology Research Group, University of the Basque Country (UPV/EHU), Leioa-Bizkaia, Spain.
[19] UMR EPOC5805, University of Bordeaux, Arcachon, France.
[20] European Marine Biology Resources Centre (EMBRC-ERIC), Paris, France.
[21] FR2424 Station Biologique de Roscoff, Sorbonne University/CNRS, Roscoff, France.
[22] Department of Applied Ecology, St Petersburg State University, St Petersburg, Russia.
[23] Department of Marine Ecosystem Functioning, University of Gdańsk, Gdynia, Poland.
[24] Department of Life Sciences, Universidad de Alcalá, Alcalá de Henares, Spain.
[25] VIB-KU Leuven Center for Cancer Biology, Leuven, Belgium.
[26] Department of Oncology, KU Leuven, Leuven, Belgium.
[27] The Francis Crick Institute, London, UK.
[28] Wellcome Sanger Institute, Hinxton, UK. ab2324@cam.ac.uk.
[29] Magdalene College, University of Cambridge, Cambridge, UK. ab2324@cam.ac.uk.
[30] Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain. jose.mc.tubio@usc.es.
[31] Department of Zoology, Genetics and Physical Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. jose.mc.tubio@usc.es.
[32] Instituto de Investigaciones Sanitarias de Santiago de Compostela (IDIS), Santiago de Compostela, Spain. jose.mc.tubio@usc.es.

USC DE COMPOSTELA

[#] Contributed equally.

Specific contributions to the publication
My contributions to this work were focused on the investigation of structural variation patterns and clonal somatic evolution of the *Cerastoderma edule* bivalve transmissible neoplasia. I conducted a genome-wide structural variation calling in CedBTN, genotyping the identified alterations in both tumours and a diverse panel of common cockles. I discriminated between germline and somatic variation in CedBTN and analysed the genomic distribution of somatic alterations. To further characterize the clonal structure of the neoplasia, I estimated ML and MP phylogenies based on large sets of SNVs and SVs and explored CedBTN variation in the context of the host species genetic diversity. I performed a comprehensive screening of high-impact somatic mutations – including SNVs, INDELs, SVs and CNVs - in CedBTN1 and CedBTN2, identifying candidate driver alterations in several cancer genes. I identified molecular markers to discriminate CedBTN1 and CedBTN2 tumours. I examined the distribution of repetitive content within the reference genome of the common cockle. I created circos plots to visualize genomic rearrangements affecting CedBTN chromosomes. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work.

Main text:
- Section "Reference genome and transcriptome of the common cockle"
- Section "Two transmissible cancers propagate through cockle populations" Figure 1 (D)
- Section "Pervasive genomic instability drives CedBTN evolution" Figure 4 (C, D).

Manuscript 3

MSIMEP: Predicting microsatellite instability from microarray DNA methylation tumor profiles
iScience 26(3), 106127 (2023). DOI: https://doi.org/10.1016/j.isci.2023.106127

Authors
Martín Santamarina-García,[1] Jenifer Brea-Iglesias,[1,2] Jesper Bertram Bramsen,[3] Mar Fuentes-Losada,[4,5] Francisco Javier Caneiro-Gómez,[6] José Ángel Vázquez-Bueno,[7] Héctor Lázare-Iglesias,[6] Natalia Fernández-Díaz,[4,5] Laura Sánchez-Rivadulla,[8] Yoel Z. Betancor,[1,5] Miriam Ferreiro-Pantín,[1,5] Pablo Conesa-Zamora,[9] José Ramón Antúnez-López,[6] Masahito Kawazu,[10,11] Manel Esteller,[12,13,14,15] Claus Lindbjerg Andersen,[3] Jose M.C. Tubio,[1] Rafael López-López,[4,5,15,*] and Juan Ruiz-Bañobre[1,4,5,15,16,**]

Affiliations
[1] Genomes and Disease, Centre for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain.
[2] Translational Oncology Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Álvaro Cunqueiro Hospital, 36213 Vigo, Spain.
[3] Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus, Denmark.
[4] Department of Medical Oncology, University Clinical Hospital of Santiago de Compostela (SERGAS), University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

[5] Translational Medical Oncology Group (ONCOMET), Health Research Institute of Santiago de Compostela (IDIS), University Clinical Hospital of Santiago de Compostela, University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain.
[6] Department of Pathology, University Clinical Hospital of Santiago de Compostela, University of Santiago de Compostela (USC), 15706 Santiago de Compostela, Spain.
[7] Department of Pathology, Complejo Hospitalario Universitario de Ferrol, 15405 Ferrol, Spain.
[8] Department of Gynaecology and Obstetrics, Complejo Hospitalario Universitario de Ferrol, 15405 Ferrol, Spain.
[9] Department of Clinical Analysis, Santa Lucía University Hospital, 30202 Cartagena, Spain.
[10] Chiba Cancer Center, Research Institute, 260-0801 Chiba, Japan.
[11] Division of Cellular Signaling, National Cancer Center Research Institute, 104-0045 Tokyo, Japan.
[12] Josep Carreras Leukaemia Research Institute (IJC), 08916 Badalona, Barcelona, Spain.
[13] Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain.
[14] Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), 08907 Barcelona, Spain.
[15] Centro de Investigación Biomédica en Red Cáncer (CIBERONC), 28029 Madrid, Spain.

Specific contributions to the publication

My contributions to this work were focused on the development and validation of a computational predictor of microsatellite instability (MSIMEP). I performed differential methylation analyses using microarray TCGA tumour datasets, followed by variable selection to optimize MSI prediction in colorectal cancer. I tuned and evaluated alternative algorithms for MSI prediction in colorectal cancer. I investigated the biological features of MSIMEP, conducting pathway enrichment analysis and determining the genomic context of critical CpG sites within the classification model. I assessed the performance of MSIMEP across multiple colorectal cancer cohorts and tested its applicability to other tumour types. I compared MSIMEP with a *MLH1* methylation-based model and characterized the immune cell context of MSIMEP predictions. I also contributed with interpretation of results, prepared figures and drafted manuscript sections related to my work.

Main text:

- Section “Model development and optimization”. Figure 1.
- Section “Underlying features of the MSIMEP model”. Figure 2.
- Section “External validation of MSIMEP models”. Figure 3 and Figure 4.
- Section “Comparison of MSIMEP models with a *MLH1* promoter methylation-based model”. Figure 5.

Supplementary text:

- Figure S1, S2, S3, S5, S6, S7, S8, S9, S10, S11, S12.

Quality metrics

The journal iScience has an impact factor of 5.8 (2022 Journal Citation Reports), a Scopus CiteScore of 6.9 (26th of December 2023) and is in the first quartile (Q1) in Multidisciplinary according to Scimago (SJR 2022 1.624).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Journal authorization

This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license.

Link to the CC BY-NC-ND license is provided here

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Appendix B: List of figures and tables

Index of main figures

Figure 1: The role of somatic mutation in cancer development. 26
Figure 2: Transposable element biology. 34
Figure 3: The L1 retrotransposition cycle. 34
Figure 4: The landscape of somatic retrotransposition across human cancers. 38
Figure 5: Genomic distribution of active L1 source elements in the human germline. 43
Figure 6: Characterization of germline active FL-L1s with long-read sequencing. 44
Figure 7: Phylogenetic tracing of L1Hs source elements. 46
Figure 8: The workflow of MEIsimulator. 49
Figure 9: IGV view of a simple retrotransposition event generated with MEIsimulator. 50
Figure 10: IGV view of a complex retrotransposition generated with MEIsimulator. 51
Figure 11: The workflow of the Source Inference method 53
Figure 12: Evaluation of the Source Inference method with somatic retrotranspositions. 56
Figure 13: Evaluation of the Source Inference method with in-silico retrotranspositions. 56
Figure 14: Overview of the L1 source element activity in MEIGA tumours. 59
Figure 15: Patterns of mobilization of dominant source L1 elements in MEIGA tumours. 60
Figure 16: Tracing ambiguous transductions. 61
Figure 17: The landscape of hot L1 source element activity in MEIGA tumours. 62
Figure 18: The role of polyadenylation signals in L1 source element transduction rates. 63
Figure 19: Active source elements with L1 mobilizations traced at allelic resolution. 65
Figure 20: Inference of L1 somatic activity traced at allelic resolution. 65
Figure 21: Bivalve species studied for being affected by transmissible cancer. 70
Figure 22: Illustration of bivalve transmissible neoplasias (BTNs) transmission cycle. 72
Figure 23: Common cockles are affected by the Cerastoderma edule BTN. 73
Figure 24: CedBTN NJ phylogenetic tree estimated from mtDNA microsatellite markers. 75
Figure 25: Neoplastic metaphases in the cockle Cerastoderma edule. 78
Figure 26: Structural variation detection from short-read paired end sequencing data. 82
Figure 27: IGV visualization of CedBTN somatic structural variant breakpoints. 82
Figure 28: Landscape of somatic structural variants across the CedBTN genomes. 83
Figure 29: Phylogenetic analysis from CedBTN single nucleotide variants. 86

USC DE SANTIAGO DE COMPOSTELA

Figure 30: Analysis of germline CedBTN SNVs. 87
Figure 31: Phylogenetic analysis from CedBTN structural variants. 88
Figure 32: Genomic locations of the marker variants included in our diagnostic assay. 91
Figure 33: Illustration of the Scuba Cancers driver search strategy. 92
Figure 34: Impact of CedBTN short variation in cancer genes. 94
Figure 35: Impact of CedBTN structural variation in cancer genes. 96
Figure 36: Analysis of cancer genes impacted by copy-number alterations in CedBTN 98
Figure 37: Analysis of CedBTN copy number alterations on cancer genes. 99
Figure 38: Copy number profiles of the MGMT gene. 100
Figure 39: Expression of cancer genes with potential driver copy number alterations. 101
Figure 40: Microsatellite DNA overview. 106
Figure 41: Development of the MSIMEP classifier. 111
Figure 42: Features of the MSIMEP model. 114
Figure 43: MSIMEP evaluation in CRC external cohort I. 115
Figure 44: MSIMEP evaluation in CRC external cohorts II-IV. 116
Figure 45: Comparison of the performance of three GLM models in a pooled CRC cohort. 117

Index of supplementary figures

Supp. Figure 1: Association between evolutionary age and L1Hs subfamily. 165
Supp. Figure 2: Association between evolutionary age and FL-L1 features. 166
Supp. Figure 3: Complete phylogeny for all active sequence-resolved FL-L1s. 167
Supp. Figure 4: Identification of MEIGA hot L1 source elements. 168
Supp. Figure 5: Frequency of classifiable repeats along the reference cockle genome. 169
Supp. Figure 6: Maximum likelihood CedBTN nuclear phylogeny from genotyped SNVs. 170
Supp. Figure 7: Molecular cytogenetic results from metaphases of healthy and neoplastic specimens. 171
Supp. Figure 8: Copy number profiles of CedBTN samples. 173
Supp. Figure 9: Genomic GC content composition of representative bivalve species. 173
Supp. Figure 10: Volcano plot representing methylation differences between MSI and MSS TCGA-COAD/READ patients. 174
Supp. Figure 11: Recursive Feature Elimination for MSIMEP development. 175

USC DE SANTIAGO DE COMPOSTELA

Supp. Figure 12: Violin plots representing MANTIS score differences between those cases correctly and incorrectly classified by MSIMEP compared with PCR-based MSI status in TCGA- COAD/READ cohort. 176
Supp. Figure 13: MSIMEP optimized model evaluation in TCGA COAD/READ cohort stratified by main clinical and molecular characteristics. 177
Supp. Figure 14: MSIMEP reduced model evaluation in TCGA COAD/READ cohort stratified by main clinical and molecular characteristics. 178
Supp. Figure 15: Exploratory evaluation of MSIMEP in TCGA-STAD and TCGA-UCEC cohorts. 179
Supp. Figure 16: Pathway enrichment analysis showing additional overrepresented KEGG terms in MSIMEP, ranked by B-H FDR-adjusted p-values. 180
Supp. Figure 17: Comparison of CIBERSORTx estimated immune cell context of predicted MSI by MSIMEP vs MSI tumors defined by standard PCR-based assays in the TCGA COAD/READ cohort. 181
Supp. Figure 18: MSIMEP optimized model evaluation in pooled CRC external cohort stratified by main clinical and molecular characteristic. 182
Supp. Figure 19: MSIMEP reduced model evaluation in pooled CRC external cohort stratified by main clinical and molecular characteristics. 183
Supp. Figure 20: MSIMEP and MLH1 promoter methylation-based model evaluation in pooled CRC external cohort stratified by MLH1 promoter methylation status. 184

Index of main tables

Table 1: Hallmarks of Genomic instability and Structural Variation in Human Cancer. 29
Table 2: Computational monitoring of MEIsimulator executions. 51
Table 3: Inference of L1 somatic activity traced to source loci at allelic resolution. 66
Table 4: General features of transmissible cancers identified in nature. 69
Table 5: Bivalve species known to be affected by transmissible bivalve neoplasia (BTN). 71
Table 6: Bivalve species with published reference genomes. ..**¡Error! Marcador no definido.**
Table 7: Size distribution of detected somatic structural variants in the CedBTN genomes. .. 84
Table 8: Primer sequences designed for the detection of CedBTN1 and CedBTN2. 91
Table 9: Candidate Cancer Genes impacted by high-impact short variants in CedBTN. 93
Table 10: Candidate Cancer Genes impacted by high-impact structural variants in CedBTN. 95

Table 11: Candidate ancestral drivers. 100
Table 12: Frequency of MSI in tumour samples from The Cancer Genome Atlas. 108
Table 13: MSIMEP optimized model performance in different cohorts. 111
Table 14: Description of the 25 CpG probes of the MSIMEP reduced model. 113
Table 15: MSIMEP reduced model performance in different cohorts. 117

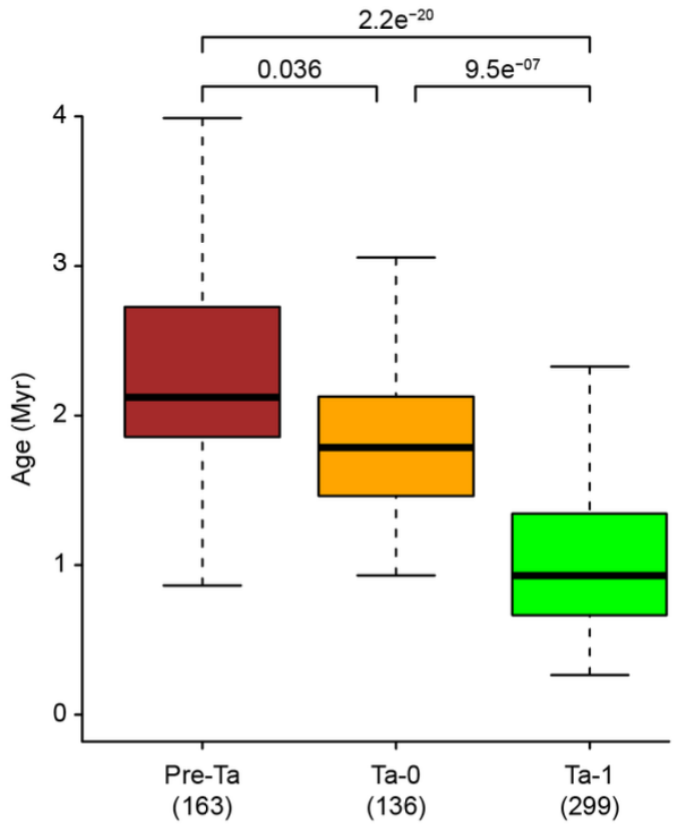
Index of supplementary tables

Supp. Table 1: Evaluation of Source Inference with real data. 185
Supp. Table 2: Evaluation of Source Inference with simulated data. 186
Supp. Table 3: Cancer genes potentially impacted by somatic alterations in CedBTN. 188
Supp. Table 4: Features of the CpGs from the MLH1 promoter methylation model. 192

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

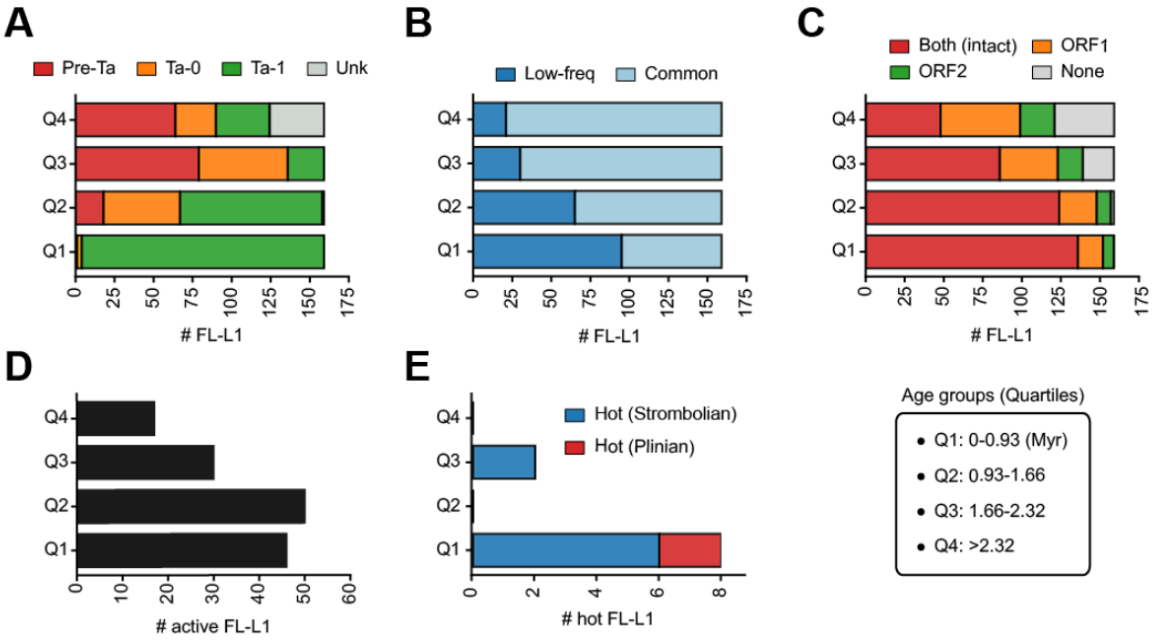
Appendix C: Supplementary material

Supplementary figures



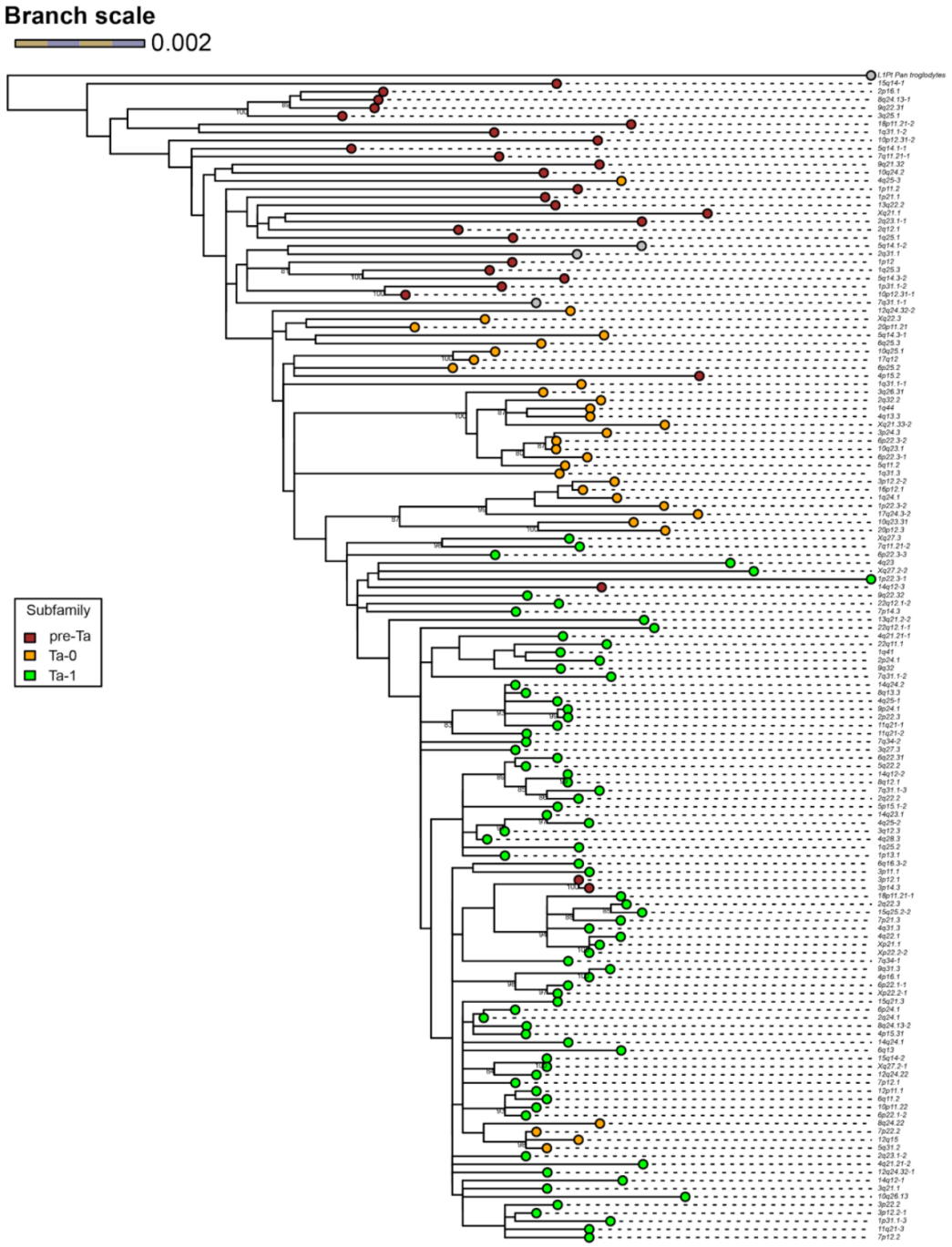
Supp. Figure 1: Association between evolutionary age and L1Hs subfamily.
Age estimates in million years (Myr) for FL-L1 grouped according to subfamily. Subfamily assignments (*Pre-Ta*, *Ta-0*, *Ta-1*) are based on diagnostic nucleotide positions. Subfamily distributions were compared pairwise via a Student's t-test, and the p-values are displayed for each comparison.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



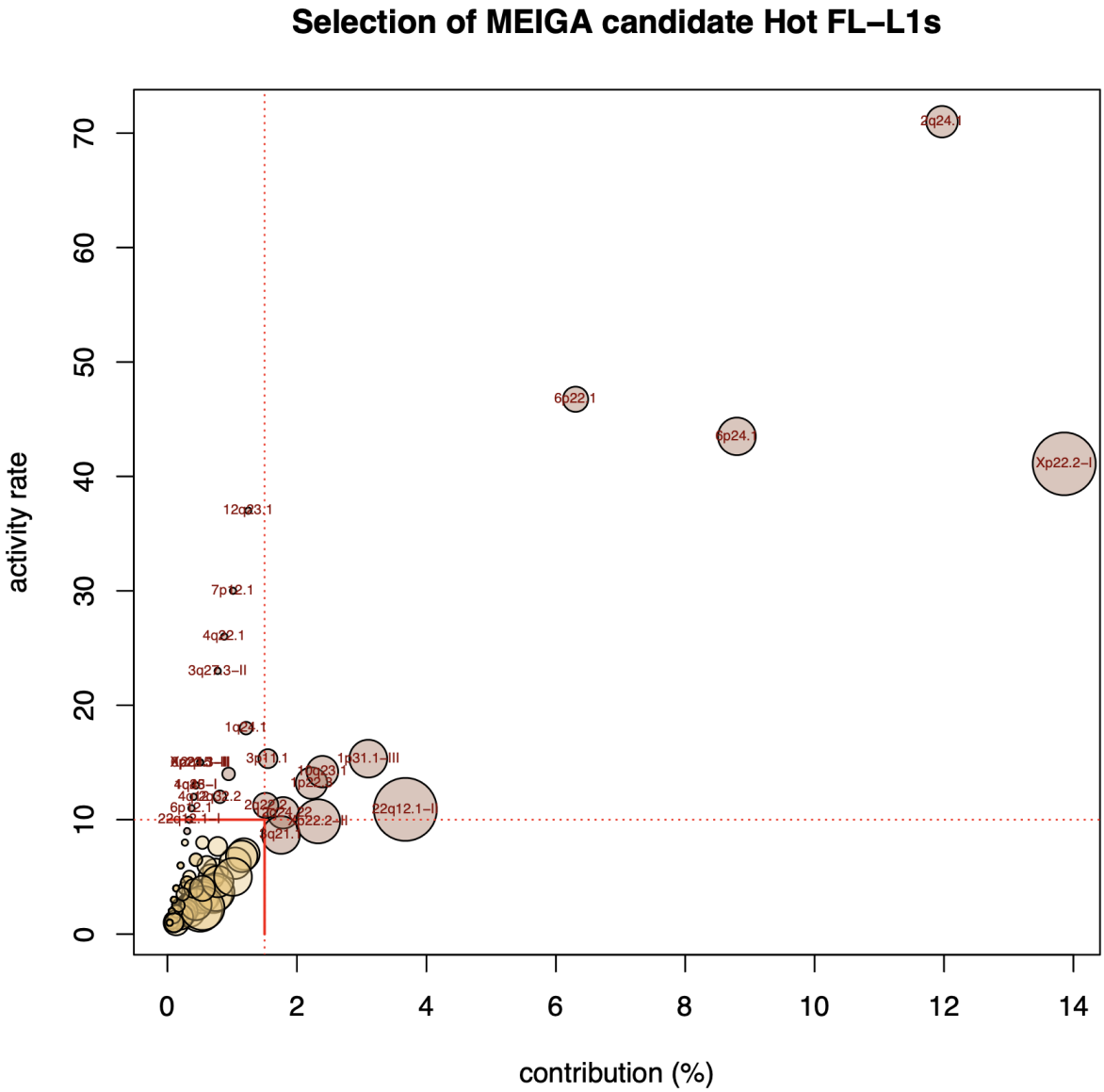
Supp. Figure 2: Association between evolutionary age and FL-L1 features.
A) Number of FL-L1s belonging to each L1 subfamily at age quartiles (Q1-Q4). FL-L1s with no subfamily diagnostic nucleotides found are classified as unknown. B) Number of low-frequency (<5% MAF) and common (≥5% MAF) FL-L1s per quartile. C) Number of FL-L1s with both ORFs intact (Both), one ORF intact (ORF1 or ORF2), or both disrupted (None) per quartile. D) Number of FL-L1 known to be active in the population, in cancer or in vitro per quartile. E) Number of FL-L1s reported to be "hot" in cancer genomes per quartile. Credit: Bernardo Rodríguez-Martín.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



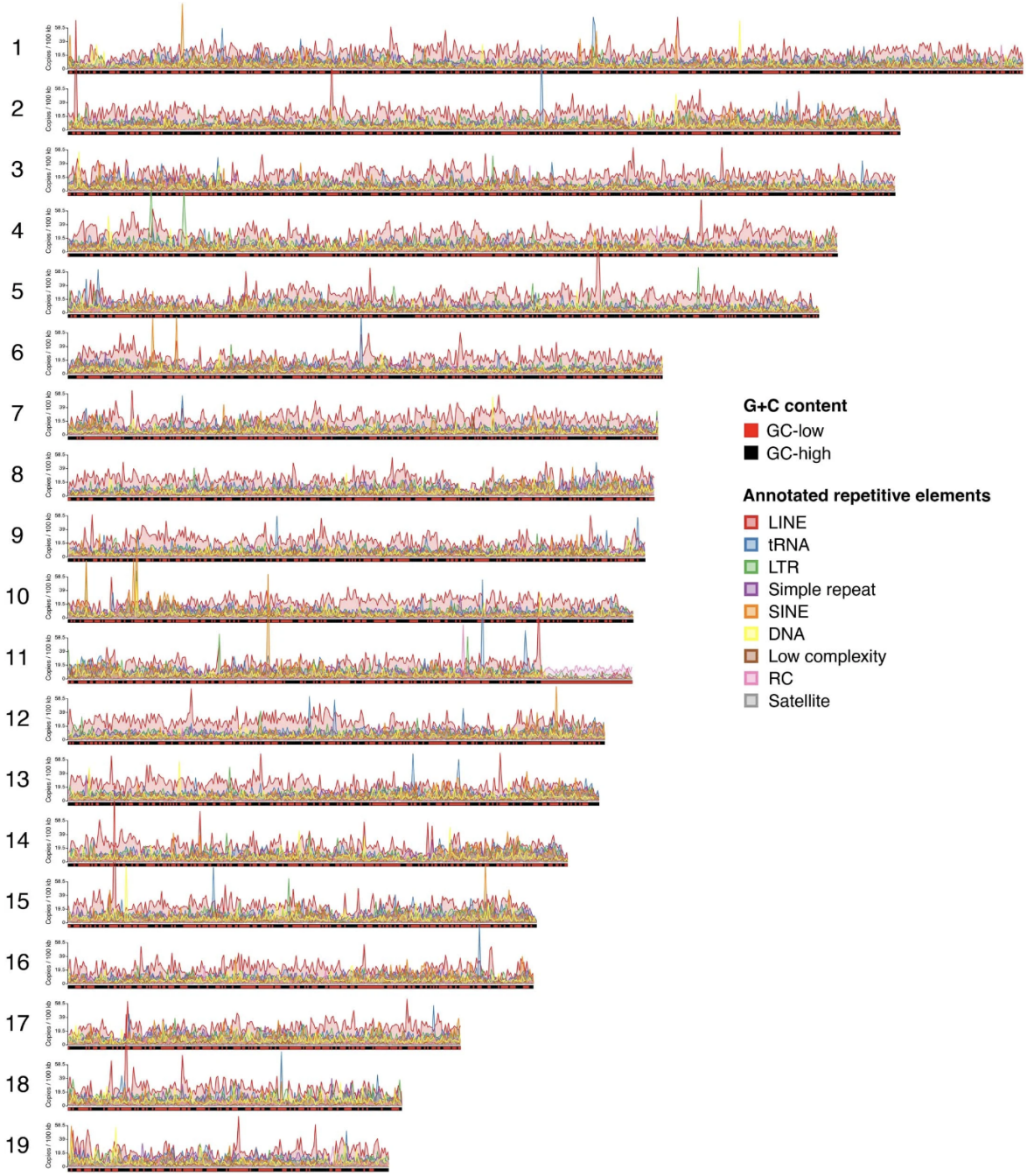
Supp. Figure 3: Complete phylogeny for all active sequence-resolved FL-L1s.
Tree branch lengths are scaled according to the average number of substitutions per base position. Active L1s are named according to the cytoband where they reside. If multiple L1 are located over the same cytoband, a numerical index is appended. Subfamily assignments based on diagnostic nucleotides for each FL-L1 are indicated as colored circles on the tip of each branch. L1 *Pan troglodytes* (L1Pt) is included as an outgroup. Bootstrap support values ≥80% are indicated.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



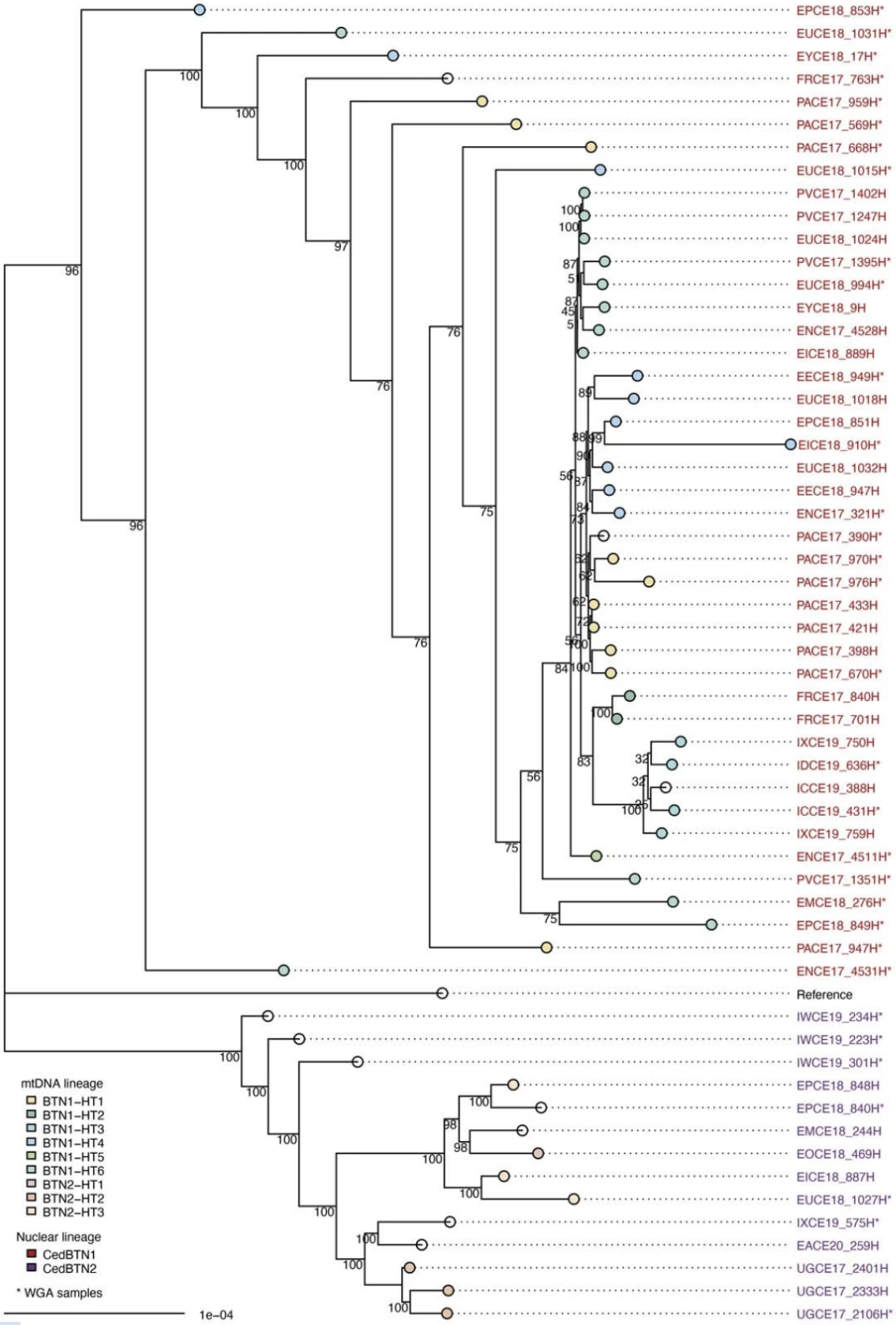
Supp. Figure 4: Identification of MEIGA hot L1 source elements.
Scatter plot showing the distribution of MEIGA source elements according to the following two variables: L1 activity rate (y axis) and relative contribution to traced somatic mobilizations (x-axis). Each circle corresponds to an active L1 source element, coloured according to a density-based clustering identity. Circle size is proportional to the number of MEIGA genomes where the source element is present (1-10). Red lines mark the threshold values applied for hot element selection (L1 activity rate ≥10 and / or contribution ≥1.5%). Hot L1 elements meeting this criterion are denoted with GRCh38 cytoband identifiers.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



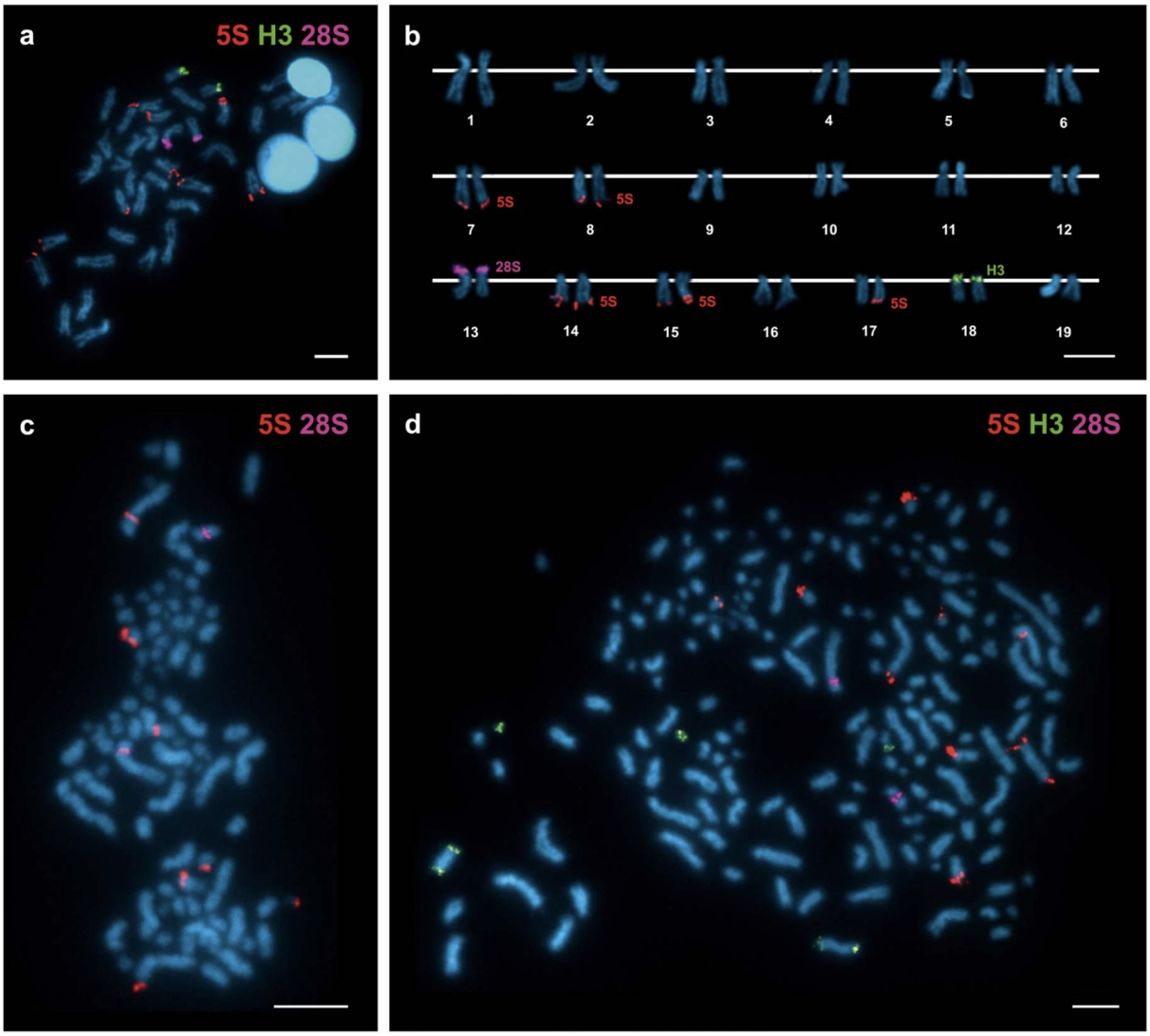
Supp. Figure 5: Frequency of classifiable repeats along the reference cockle genome.
Repeats are displayed in terms of number of copies per 100-kb genomic segment. Repetitive element types with more than 1000 annotated copies are represented: long interspersed nuclear elements (LINE, 172722 copies, 33.0%), transfer RNA repeats (tRNA, 81766 copies, 15.6%), long terminal repeats (LTR, 78009 copies, 14.9%), simple repeats (70016 copies, 13.3%), short interspersed nuclear elements (SINE, 55434 copies, 10.6%), DNA repeat elements (42917 copies, 8.2%), low complexity repeats (12171 copies, 2.3%), rolling circle repeats (RC, 8843 copies, 1.7%), satellite repeats (2100 copies, 0.4%). Genomic segments along the ideogram are classified as GC-low or GC-high based on whether their average nucleotide content is below or above the estimated average genomic G+C content (35.6%).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



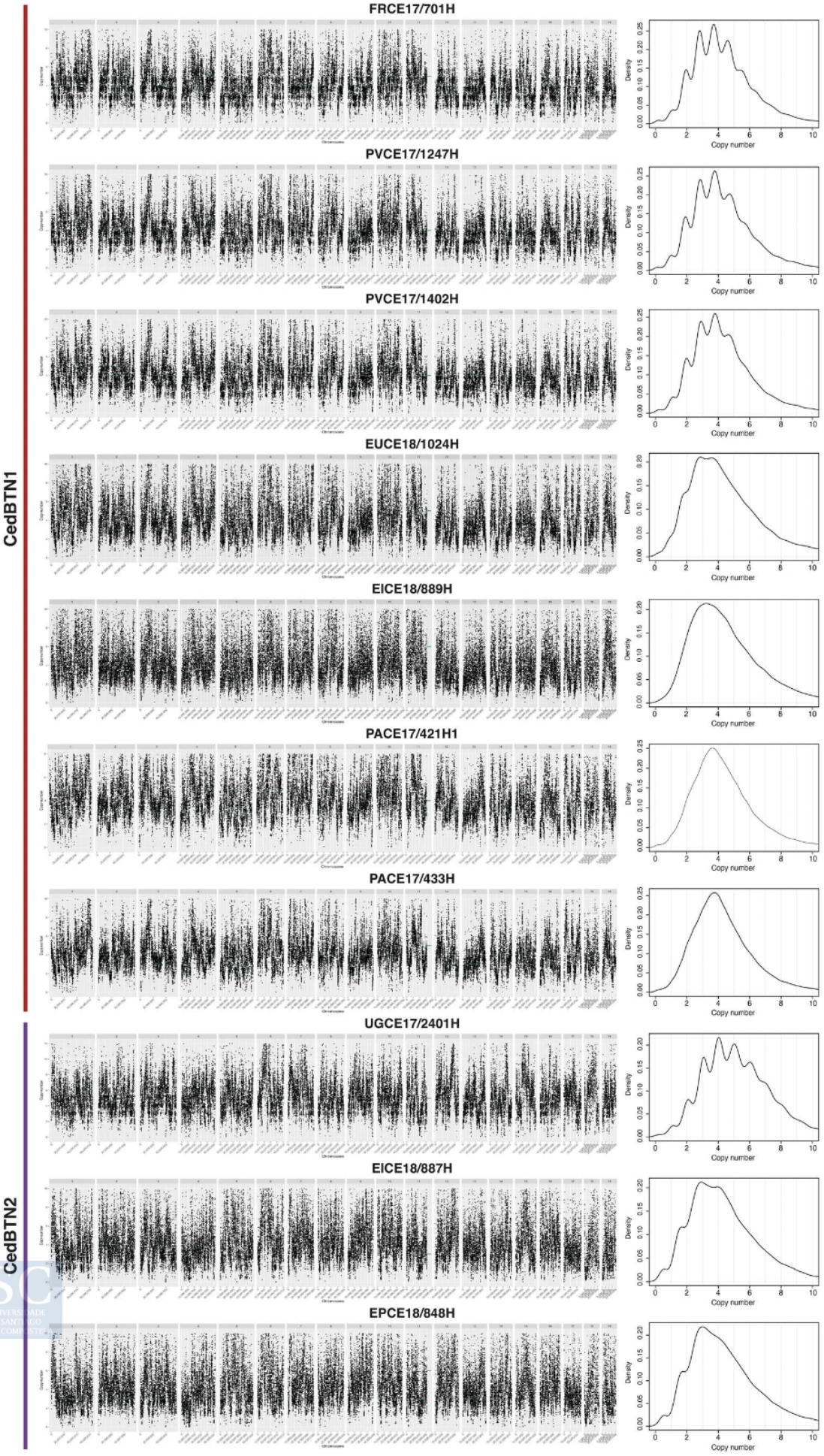
Supp. Figure 6: Maximum likelihood CedBTN nuclear phylogeny from genotyped SNVs.
Tree inferred from a subset of 833007 BTN-specific SNVs, including 300000 randomly selected SNVs from each of the ancestral variant sets ('A0','A1','A2') and all the non-ancestral (post-divergence) SNV in each nuclear

lineage, which were genotyped across 61 tumour samples. Tips are coloured according to mtDNA lineage (where information is available); sample labels are coloured according to nuclear CedBTN lineage. Bootstrap support values (n = 1000 replicates) are shown for all nodes. Samples subjected to whole-genome amplification (WGA) are indicated by asterisks. Sample EICE18_910H is a case of co-infection by cells from mtDNA lineages BTN1-HT4 and BTN2-HT2.



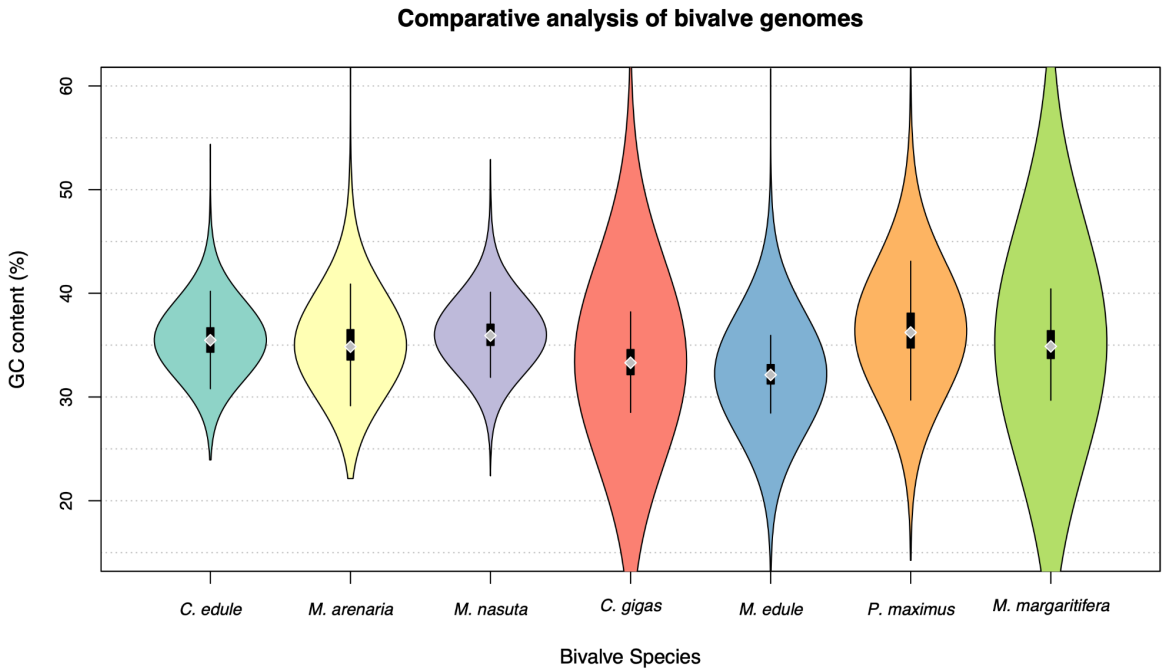
Supp. Figure 7: Molecular cytogenetic results from metaphases of healthy and neoplastic specimens. a-b, FISH of 28S ribosomal DNA (rDNA; violet), 5S rDNA (red) and H3 histone gene (green) probes mapped onto a mataphase plate of a healthy specimen of *C. edule* and its corresponding karyotype with 2n=38 chromosomes. As previously described, up to five chromosome pairs hold subtelomeric clusters of 5S rDNA on their long arms, while 28S rDNA and histone H3 probes hybridize to the short arm of subtelocentric chromosomes. c-d, FISH mapping of the probes above onto example neoplastic metaphases, revealing abnormal location and number of these clusters. Scale bars, 10 µm. Credit: Daniel García-Souto.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



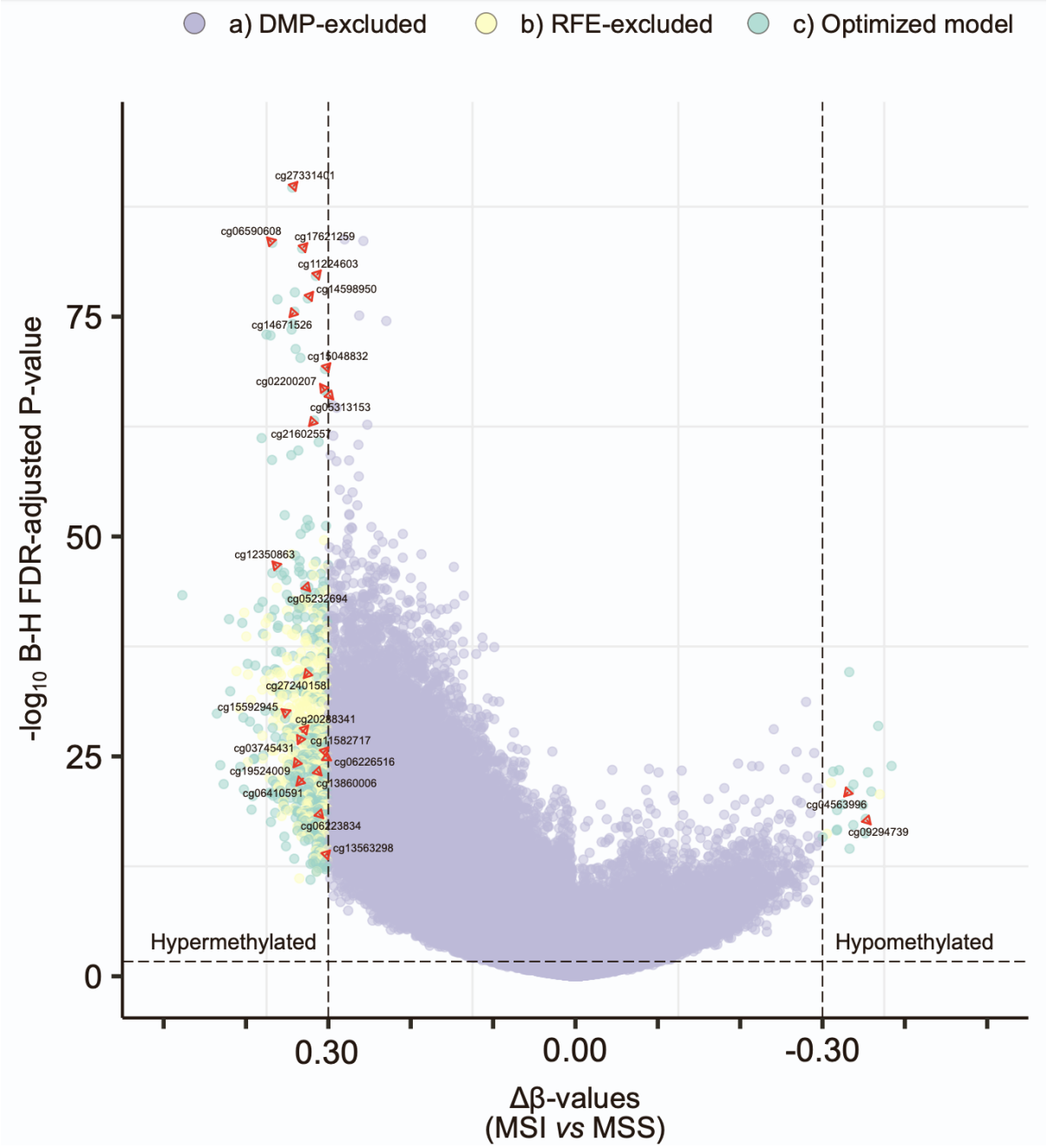
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Supp. Figure 8: Copy number profiles of CedBTN samples.
Plots of unrounded copy number along the reference genome (left) and copy number density (right) for each sample in the 'golden set' of high-purity tumours, grouped by CedBTN lineage. Each dot represents a genomic bin containing 10 Kb of mappable sequence. Credit: Jonas Demeulemeester.



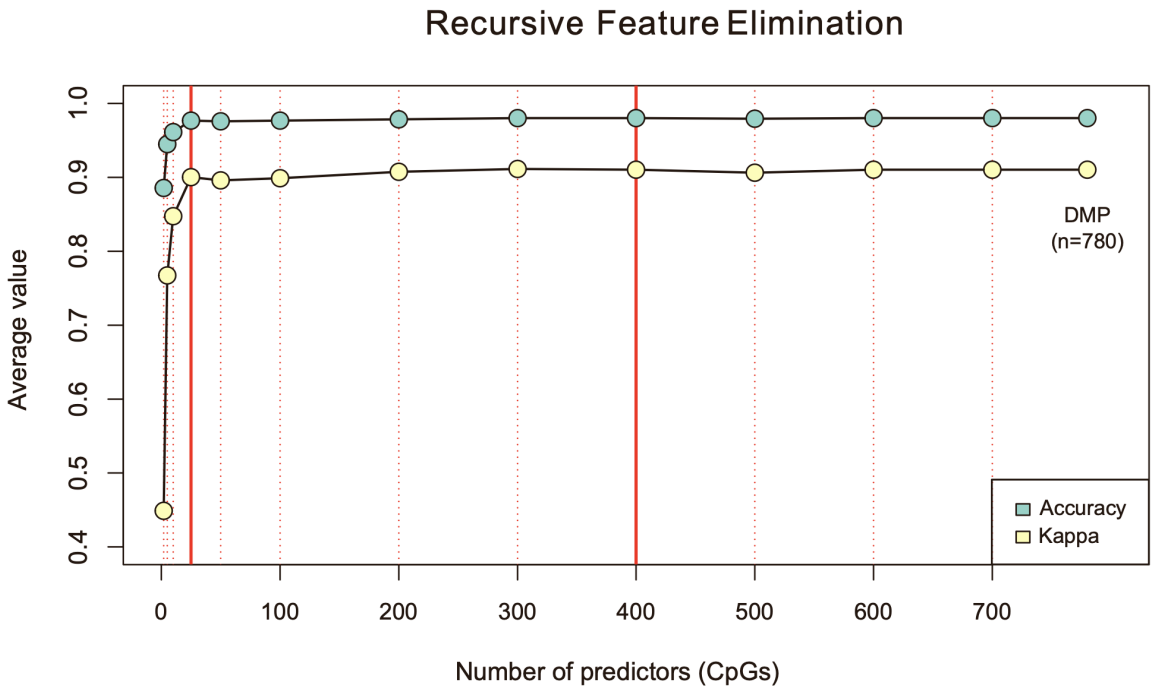
Supp. Figure 9: Genomic GC content composition of representative bivalve species.
The GC content was computed at 10KB sliding genomic windows. Reference genomes were downloaded from: https://www.ncbi.nlm.nih.gov/datasets/genome/

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



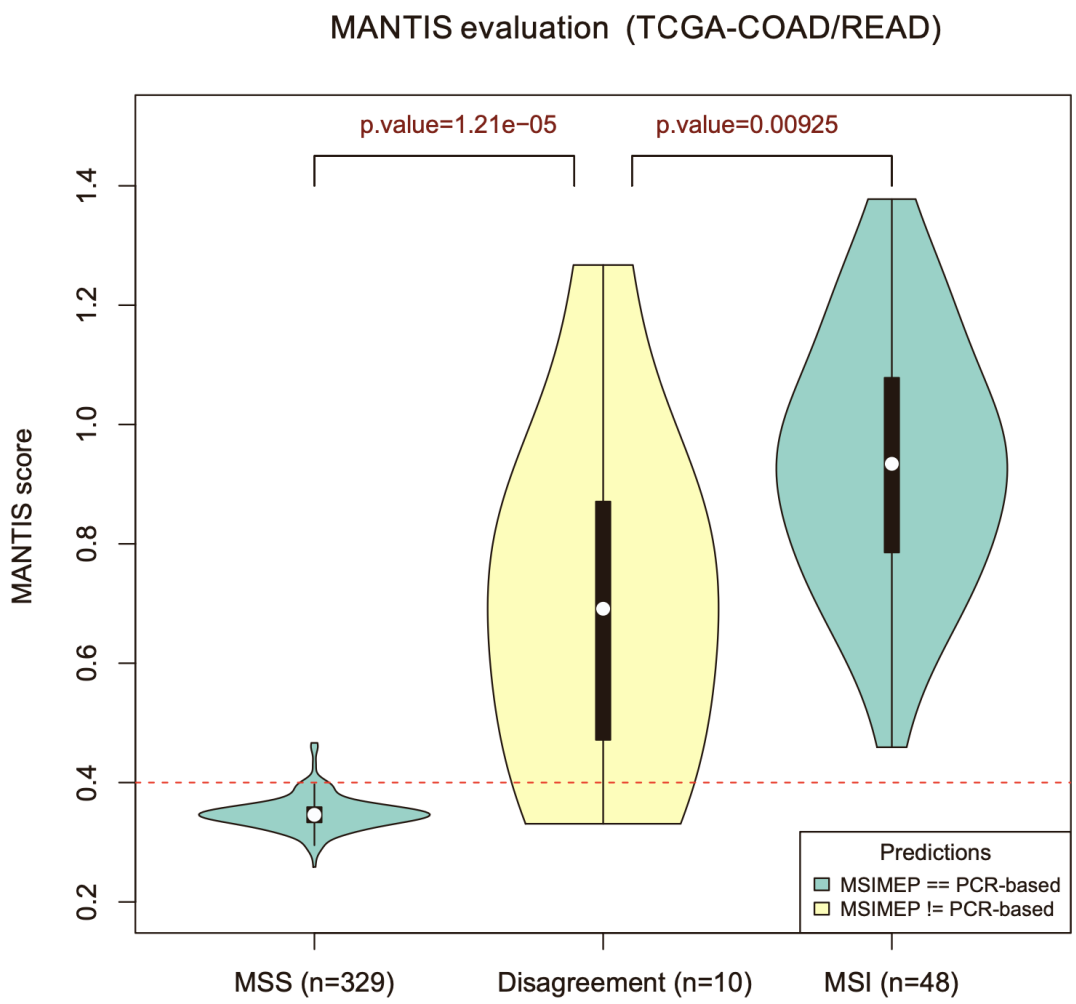
Supp. Figure 10: Volcano plot representing methylation differences between MSI and MSS TCGA-COAD/READ patients.
The X-axis represents the difference in average methylation (Δß values) between MSI and MSS cases, highlighting hypermethylated (left, Δß>0.3) and hypomethylated (right, Δß<-0.3) CpG probes. The Y-axis shows the B-H FDR-adjusted p-values in logarithmic scale. Dashed lines indicate significance thresholds. Purple: CpG probes excluded during differential methylation analysis. Yellow: probes excluded during recursive feature elimination analysis. Green: 400 CpG probes selected for MSIMEP model. Red flags highlight the 25 probes included in MSIMEP reduced model version.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



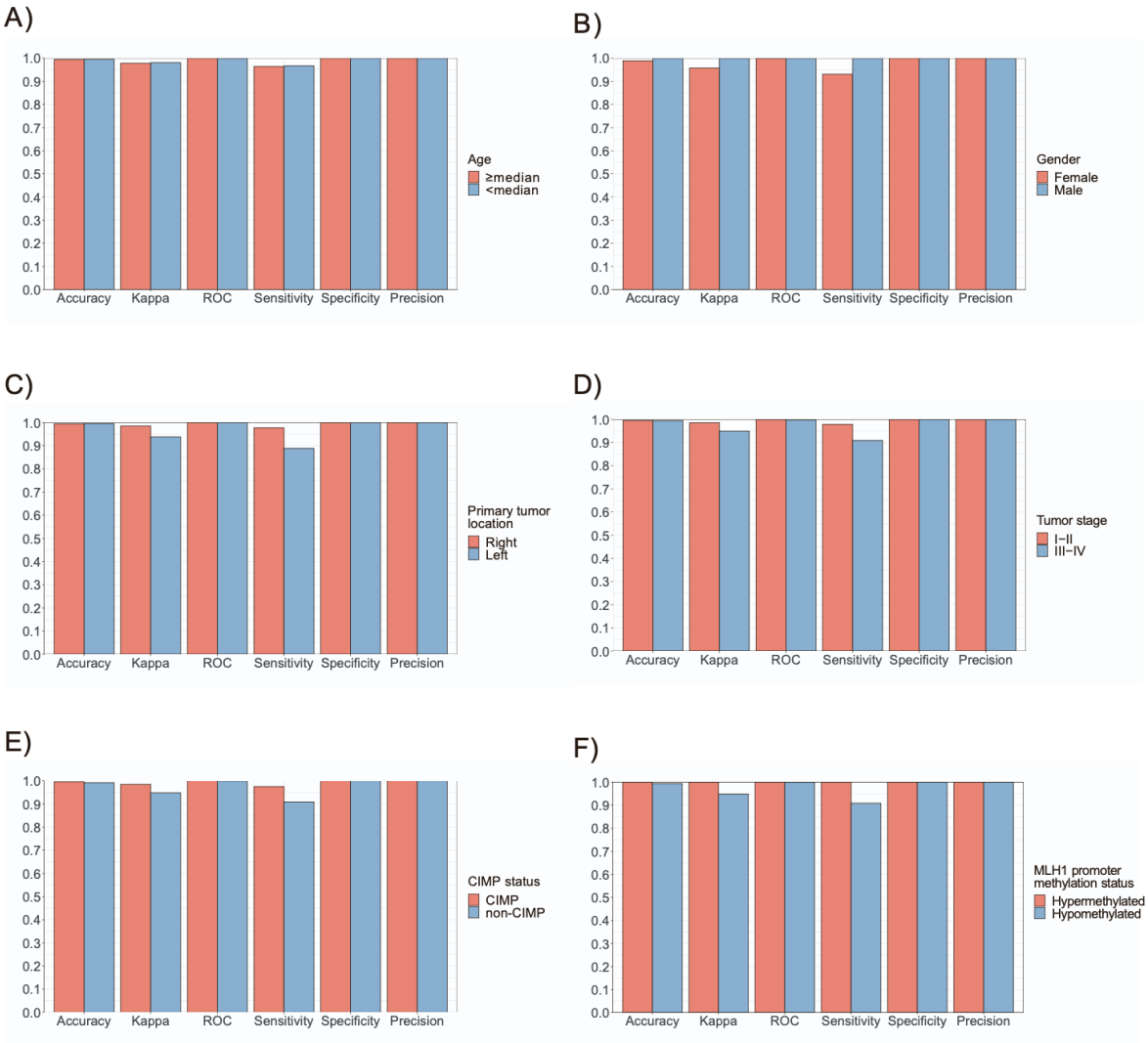
Supp. Figure 11: Recursive Feature Elimination for MSIMEP development.
Backwards feature selection was performed starting from the basal model of 780 CpG probes, testing 700, 600, 500, 400, 300, 200, 100, 50, 25, 10, 5 and 2 probes reduced model versions. Model evaluations were performed based on 3 repeats of 2-fold cross validation. MSIMEP optimized and reduced models (solid red lines) were chosen by selecting the less complex model with higher accuracy estimate (400 CpG probes) and the less complex model with accuracy loss inferior to 1% (25 CpG probes), respectively.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



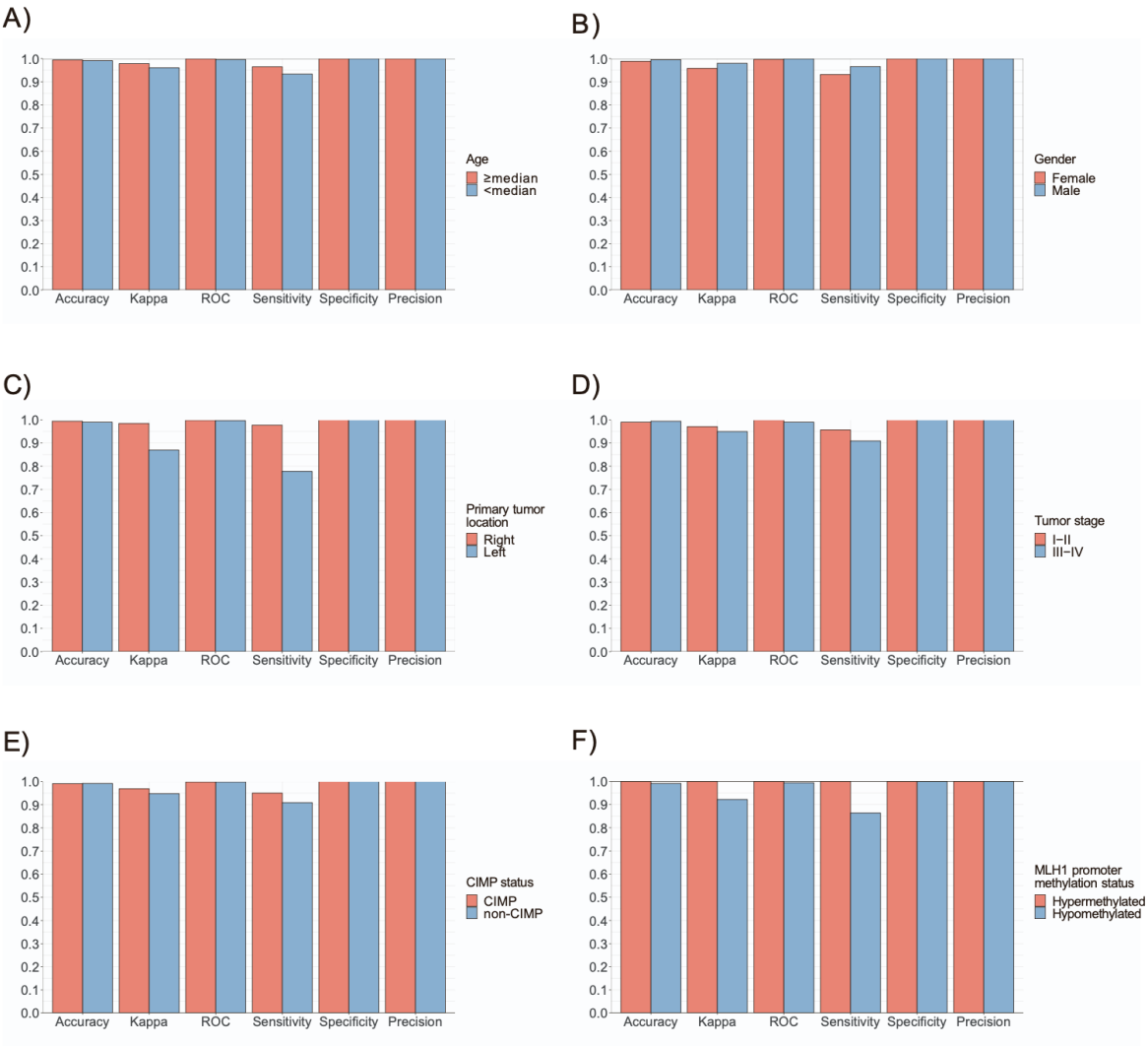
Supp. Figure 12: Violin plots representing MANTIS score differences between those cases correctly and incorrectly classified by MSIMEP compared with PCR-based MSI status in TCGA- COAD/READ cohort. MSIMEP predictions were set as the consensus phenotype (MSI/MSS) derived from the 5 repeats of 10-fold cross validation. PCR-based predictions were the ones used for training MSIMEP. The Y-axis quantifies microsatellite instability, computed in terms of MANTIS score. Dashed red line indicates MANTIS default threshold (0.4). Green: cases where MSIMEP prediction matches PCR-based (reference). Yellow: cases where MSIMEP and PCR-based prediction disagree. Dark boxes display the interquartile range (IQR), with white dots showing the median value of each distribution. Statistical significance in MANTIS score differences were assessed through the Wilcoxon rank sum test.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



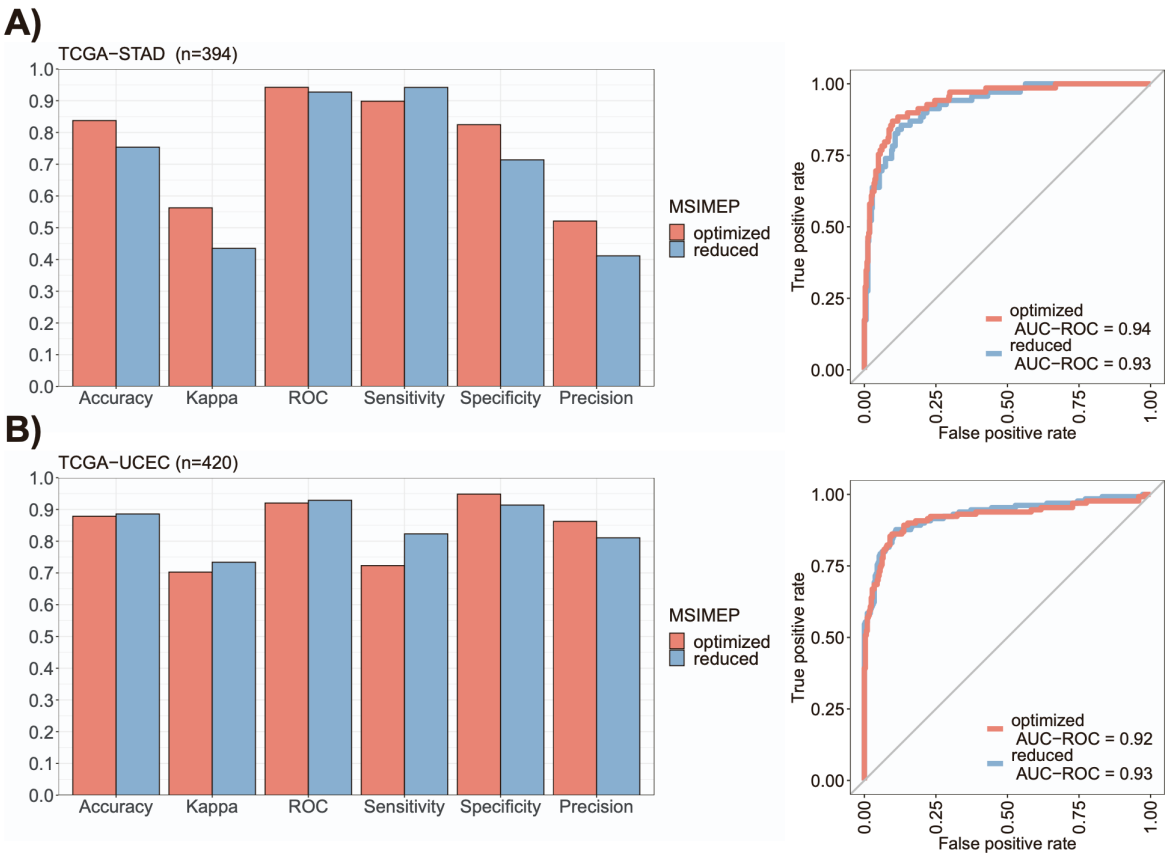
Supp. Figure 13: MSIMEP optimized model evaluation in TCGA COAD/READ cohort stratified by main clinical and molecular characteristics.
Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



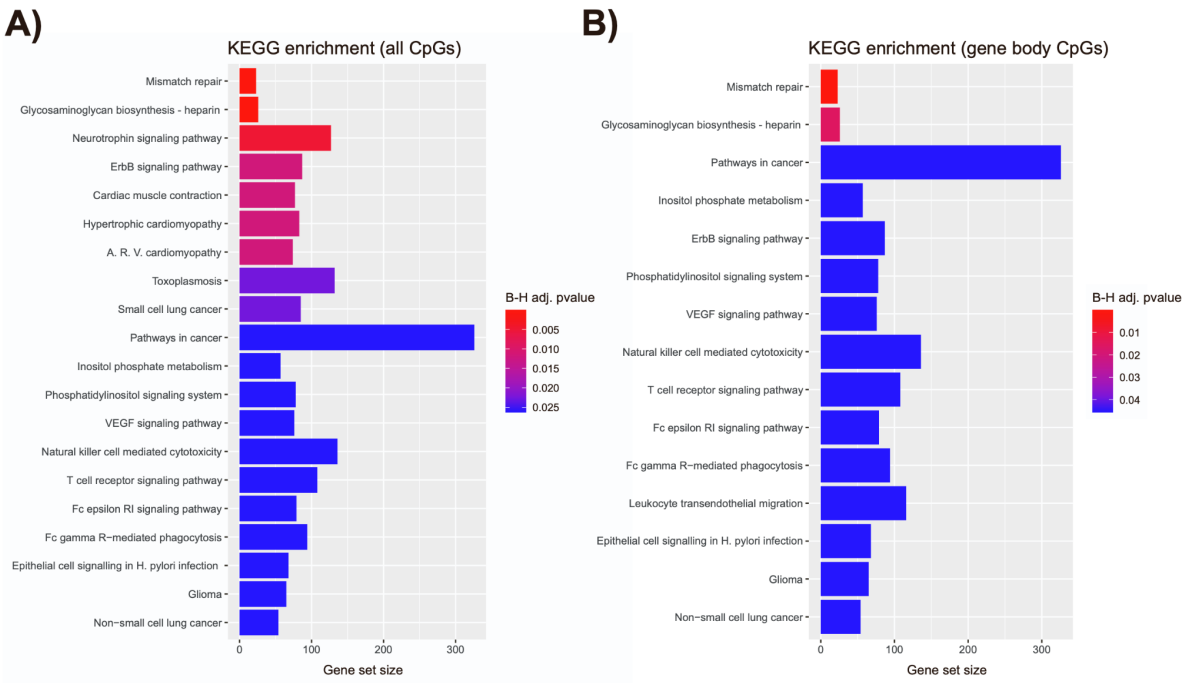
Supp. Figure 14: MSIMEP reduced model evaluation in TCGA COAD/READ cohort stratified by main clinical and molecular characteristics.
Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



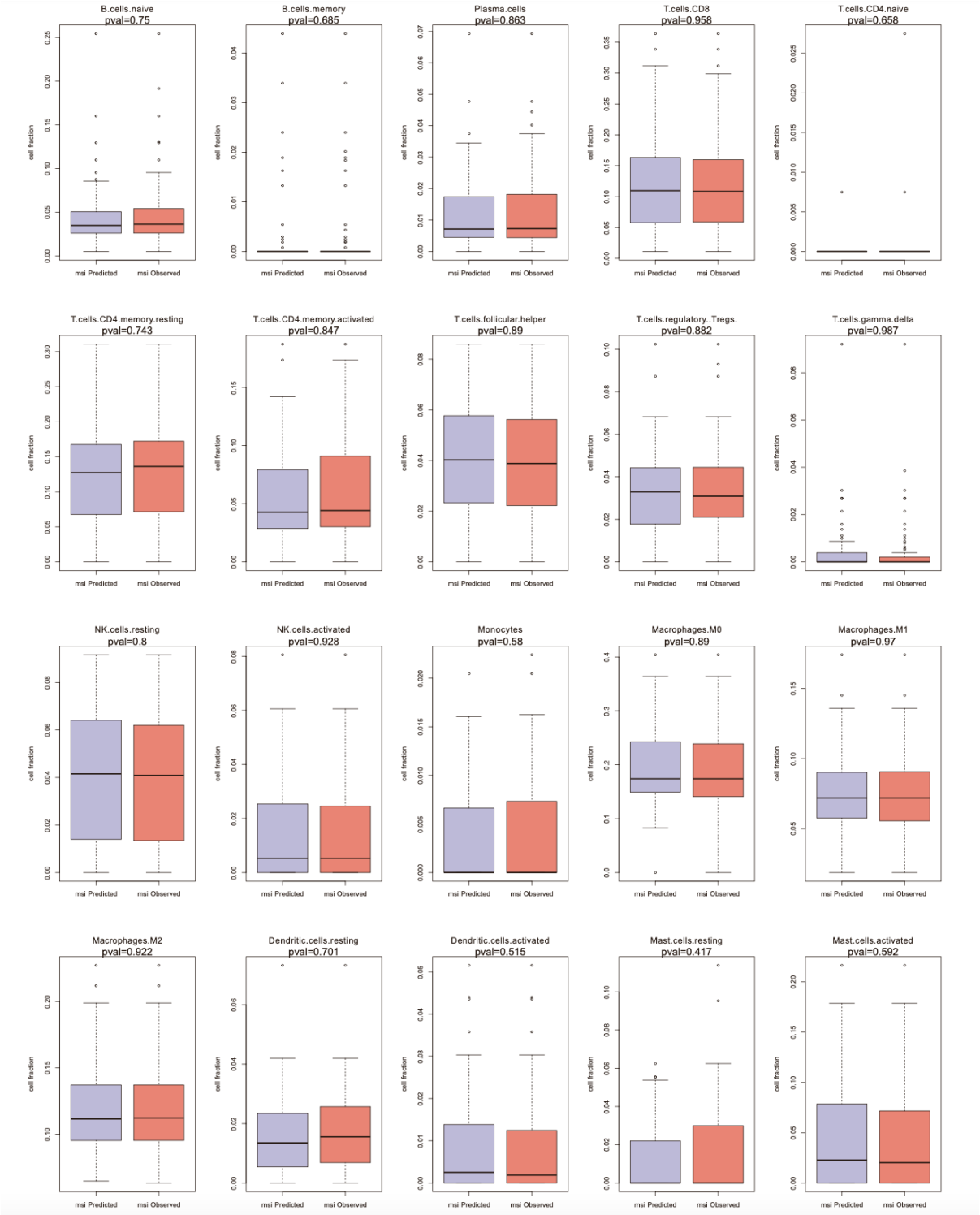
Supp. Figure 15: Exploratory evaluation of MSIMEP in TCGA-STAD and TCGA-UCEC cohorts.
Exploratory evaluation of MSIMEP in TCGA-STAD and TCGA-UCEC cohorts. Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients. ROC curves showing MSIMEP classification performance at alternative predictive thresholds. Models: MSIMEP optimized (red) and MSIMEP reduced (blue). (A) Cohort TCGA-STAD (n=394, 18% MSI), and (B) Cohort TCGA-UCEC (n=420, 31% MSI).

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



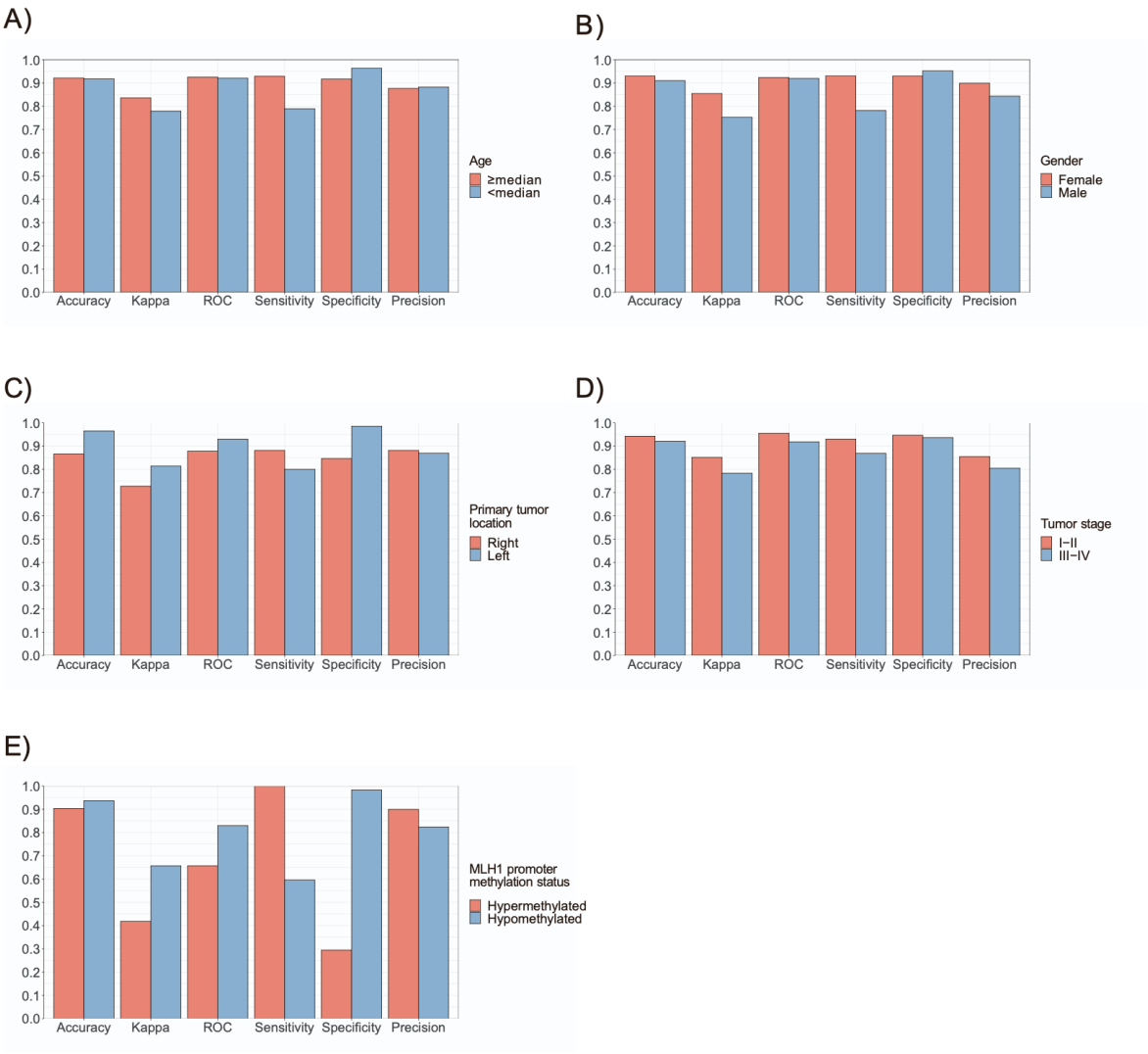
Supp. Figure 16: Pathway enrichment analysis showing additional overrepresented KEGG terms in MSIMEP, ranked by B-H FDR-adjusted p-values.
Pathway enrichment analysis showing additional overrepresented KEGG terms in MSIMEP, ranked by B-H FDR-adjusted p-values. (A) Enriched KEGG terms considering all MSIMEP CpGs. (B) Enriched KEGG terms considering gene body MSIMEP CpGs. The X-axis represents the size of gene sets associated with significantly enriched terms.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



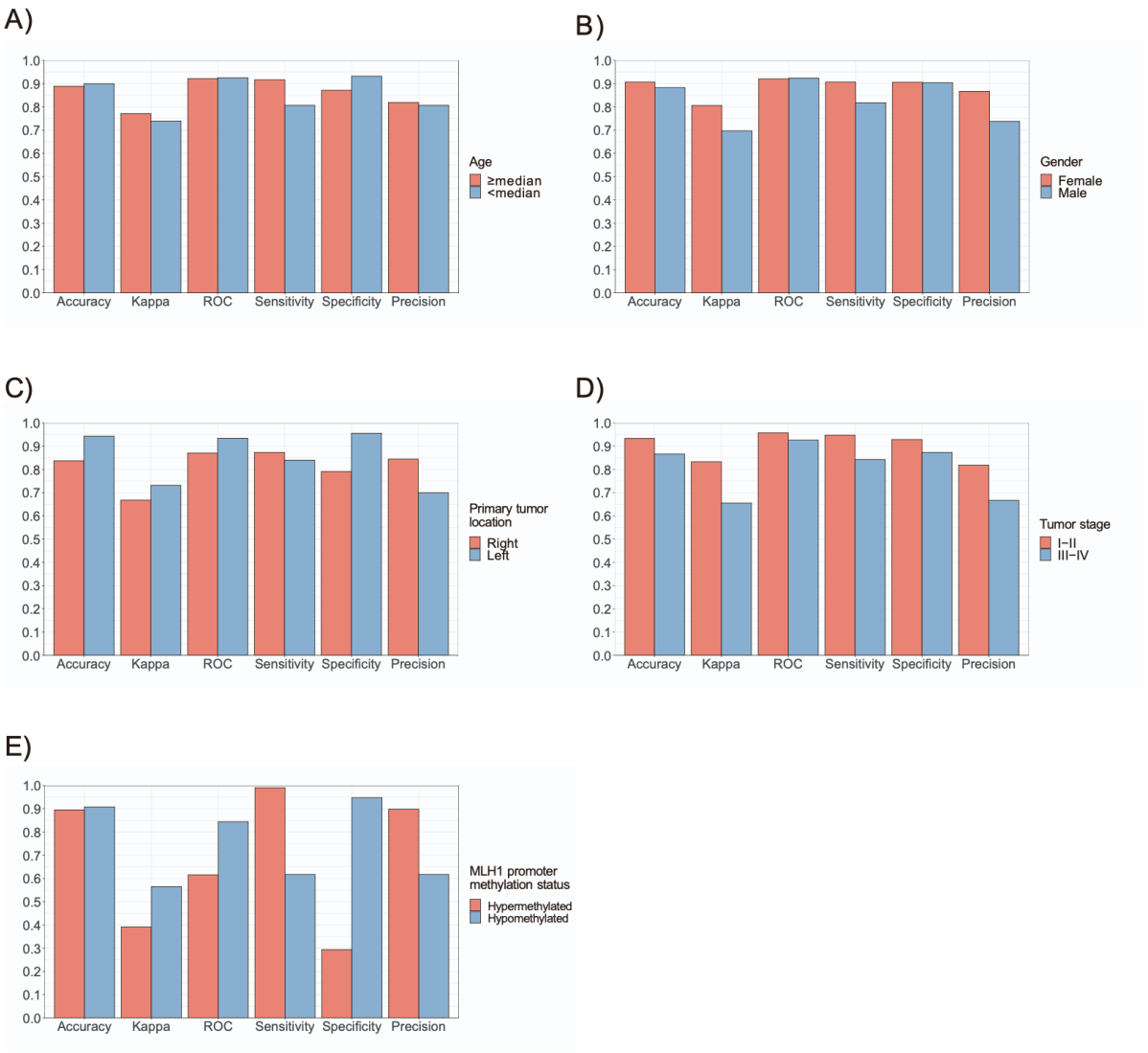
Supp. Figure 17: Comparison of CIBERSORTx estimated immune cell context of predicted MSI by MSIMEP vs MSI tumors defined by standard PCR-based assays in the TCGA COAD/READ cohort.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



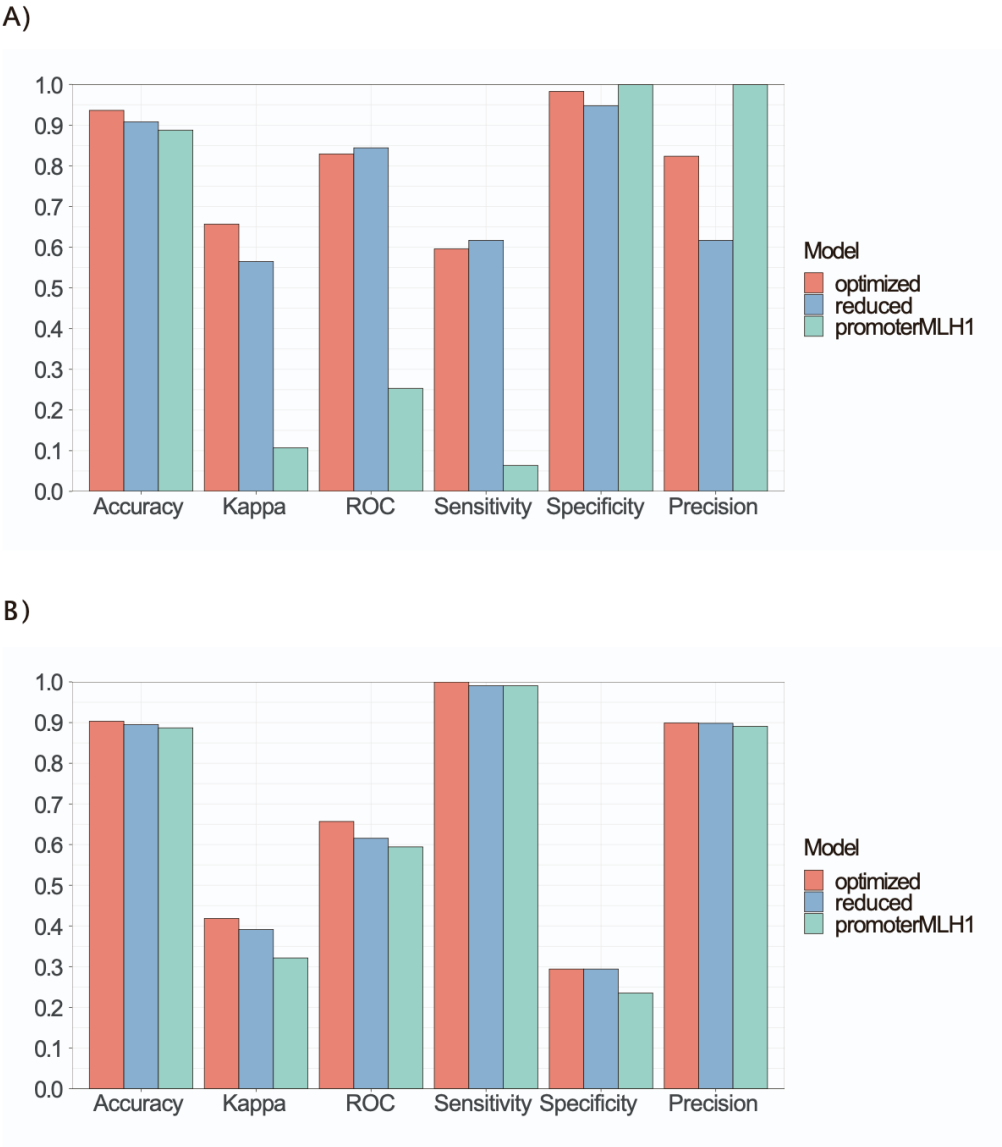
Supp. Figure 18: MSIMEP optimized model evaluation in pooled CRC external cohort stratified by main clinical and molecular characteristic.
Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Supp. Figure 19: MSIMEP reduced model evaluation in pooled CRC external cohort stratified by main clinical and molecular characteristics.
Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA



Supp. Figure 20: MSIMEP and MLH1 promoter methylation-based model evaluation in pooled CRC external cohort stratified by MLH1 promoter methylation status.
Bar plots showing complementary metrics (Accuracy, Kappa, AUROC, Sensitivity, Specificity, and Precision) for the evaluation of MSIMEP classification capacity of MSI vs MSS patients.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Supplementary tables

Supp. Table 1: Evaluation of Source Inference with real data.

Sample	MEI	Type	TDLEN	Source TD	Source Inference	Status
PD0270	chr1:190050368	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr1:243203719	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr1:246015938	partnered	800	chr5_39788000	chr5_39788000	Concordant
PD0270	chr10:4388709	partnered	357	chr6_29952000	chr6_29952000	Concordant
PD0270	chr10:10390240	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr10:54079253	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr11:4862888	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr11:5456029	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr11:90369962	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr12:72525244	partnered	763	chr12_117377000	chr12_117377000	Concordant
PD0270	chr13:42526888	partnered	151	chr6_2418000	chr6_2418000	Concordant
PD0270	chr13:43271764	partnered	427	chr22_28663000	chr22_28663000	Concordant
PD0270	chr13:88231467	partnered	111	chr1_84052000	chr11_95436000	Discordant
PD0270	chr14:43497050	partnered	608	chr22_28663000	chr22_28663000	Concordant
PD0270	chr14:43927595	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr14:46453934	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr14:78325140	partnered	317	chr4_90676000	chr4_90676000	Concordant
PD0270	chr14:86483993	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr16:13266349	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr17:53507876	partnered	316	chr6_29952000	chr6_29952000	Concordant
PD0270	chr17:56500577	partnered	413	chr6_29952000	chr22_28663000	Discordant
PD0270	chr18:29494928	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr19:12463437	partnered	427	chr22_28663000	chr22_28663000	Concordant
PD0270	chr2:29390561	partnered	130	chr1_84052000	chr11_95436000	Discordant
PD0270	chr2:99956600	partnered	361	chr6_29952000	chr6_29952000	Concordant
PD0270	chr2:117229338	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr2:131342651	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr2:146039244	partnered	119	chr6_29952000	chr6_29952000	Concordant
PD0270	chr20:6626531	partnered	494	chr1_113497000	chr1_113497000	Concordant
PD0270	chr20:11726050	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr22:10946736	partnered	376	chr5_39788000	chr5_39788000	Concordant
PD0270	chr22:48032848	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr3:4176867	partnered	64	chr1_86679000	chr1_86679000	Concordant
PD0270	chr3:117242916	partnered	818	chr22_28663000	chr22_28663000	Concordant
PD0270	chr3:135238025	partnered	420	chr6_29952000	chr6_29952000	Concordant
PD0270	chr3:187363036	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr4:11971382	partnered	62	chr1_86679000	chr1_86679000	Concordant
PD0270	chr4:63519587	partnered	233	chr10_89958000	chr10_89958000	Concordant
PD0270	chr4:65446270	partnered	410	chr6_29952000	chr6_29952000	Concordant
PD0270	chr5:7454315	partnered	27	chrX_11707000	chrX_11707000	Concordant
PD0270	chr5:12360865	partnered	400	chrX_11935000	chrX_11935000	Concordant
PD0270	chr5:13335431	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr5:109164889	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr6:23874925	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr6:66515860	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr6:95506731	partnered	818	chr22_28663000	chr22_28663000	Concordant
PD0270	chr6:96784175	partnered	1783	chr6_29952000	chr6_29952000	Concordant
PD0270	chr6:130158258	partnered	494	chr1_113497000	chr1_113497000	Concordant
PD0270	chr6:135690898	partnered	130	chr1_84052000	chr11_95436000	Discordant
PD0270	chr6:153384595	partnered	111	chr1_84052000	chr11_95436000	Discordant
PD0270	chr7:40510110	partnered	413	chr6_29952000	chr6_29952000	Concordant

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PD0270	chr7:51545090	partnered	71	chr4_90676000	chr4_90676000	Concordant
PD0270	chr7:61759067	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:67742862	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:78472498	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:85030765	partnered	517	chr5_39788000	chr5_39788000	Concordant
PD0270	chr7:90071356	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:110139889	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:113432665	partnered	25	chrX_11707000	chrX_11707000	Concordant
PD0270	chr7:124633732	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:126512011	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr7:126739551	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr8:3325113	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chr8:34842090	partnered	1756	chrX_11935000	chrX_11935000	Concordant
PD0270	chr8:52129641	partnered	650	chr22_28663000	chr22_28663000	Concordant
PD0270	chr8:55638874	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr8:85993973	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr9:5119871	partnered	418	chr6_29952000	chr6_29952000	Concordant
PD0270	chr9:18985405	partnered	22	chrX_119435000	chrX_119435000	Concordant
PD0270	chr9:23345245	partnered	568	chr22_28663000	chr22_28663000	Concordant
PD0270	chr9:28081708	partnered	413	chr6_29952000	chr6_29952000	Concordant
PD0270	chrX:91721484	partnered	418	chr6_29952000	chr6_29952000	Concordant

For each L1-partnered event traced in sample PD0270 with complementary strategies (3' transduction and diagnostic nucleotide profiling), candidate sources are compared, and inference status is assigned (concordant / discordant) based on their agreement.

Supp. Table 2: Evaluation of Source Inference with simulated data.

MEI	Length	Subfamily	Source Inference	Source element	Status
chr2:238543449	1500	Ta-1	Not assigned	chr11_78677772	untraced
chr2:46636438	218	Ta-0	Not assigned	chr6_19764892	untraced
chr2:117120272	2648	Ta-0	Not assigned	chr10_85355506	untraced
chr10:44576984	316	Ta-0	Not assigned	chr4_119948726	untraced
chr7:146694500	1967	pre-Ta	chr3_108749400	chr3_108749400	traced
chr3:152421379	293	Ta-0	Not assigned	chr18_75846851	untraced
chr17:33312429	1059	Ta-0	chrX_106469285	chrX_106469285	traced
chr13:102932137	837	Ta-1	Not assigned	chr7_25041860	untraced
chr4:142754706	680	Ta-1	chr3_132946006	chr3_132946006	traced
chrX:25605150	3410	pre-Ta	chrX_83542396	chrX_83542396	traced
chr4:19421426	451	Not assigned	Not assigned	chr3_26398017	untraced
chr8:92154222	159	Ta-0	Not assigned	chrX_106469285	untraced
chr15:96776839	975	pre-Ta	chr18_51689848	chr18_51689848	traced
chr21:22849712	1399	pre-Ta	chr5_146609485	chr5_146609485	traced
chrX:155912961	2570	Ta-0	chr11_90966271	chr11_90966271	traced
chrX:65612883	331	Ta-0	chr2_16593725	chr2_16593725	traced
chr12:122796147	385	Ta-0	chr11_31315654	chr11_31315654	traced
chr7:7475099	118	Not assigned	Not assigned	chr6_72988654	untraced
chr12:35421717	572	Ta-1	chrX_69477026	chrX_69477026	traced
chr18:42182340	3357	Ta-1	Not assigned	chr2_155671334	untraced
chr3:106386835	1252	Ta-0	chr13_31302314	chr13_31302314	traced
chr2:72602701	454	pre-Ta	Not assigned	chr5_133583288	untraced
chr14:77264092	552	pre-Ta	chr1_187597671	chr1_187597671	traced
chr17:11601798	364	pre-Ta	chr5_39787653	chr5_39787653	traced
chr7:152270639	6068	pre-Ta	chrX_66180696	chrX_66180696	traced
chr3:150221744	852	Ta-0	chr7_108188803	chr7_108188803	traced
chr15:88472426	213	Ta-1	Not assigned	chr1_95263900	untraced
chr10:60116311	212	pre-Ta	Not assigned	chr5_102131356	untraced
chr7:86240036	312	Ta-1	chr3_79167153	chr3_79167153	traced
chr13:22189421	212	pre-Ta	chr1_71513698	chr1_71513698	traced
chr2:84807076	178	Ta-1	chr6_24811657	chr6_24811657	traced

chr2:5871171	374	Ta-1	Not assigned	chr2_10996106	untraced
chr17:13588678	378	Not assigned	chr8_91522091	chr8_91522091	traced
chr12:100822908	119	Ta-1	Not assigned	chr4_93638307	untraced
chr12:4573147	382	Ta-1	Not assigned	chr12_126299023	untraced
chr3:15955706	137	Ta-1	Not assigned	chr22_28663283	untraced
chr7:55672594	4826	Ta-1	chrX_69477026	chrX_69477026	traced
chr3:71326208	151	pre-Ta	Not assigned	chr9_110791097	untraced
chr6:136479786	2591	Ta-0	chr14_62535851	chr14_62535851	traced
chr1:203003320	665	pre-Ta	chr16_54042096	chr16_54042096	traced
chr1:228001803	333	Ta-0	Not assigned	chr13_31302314	untraced
chr13:94165046	528	pre-Ta	Not assigned	chr1_102103355	untraced
chr20:16275297	188	Ta-1	chr4_98592435	chr4_98592435	traced
chr10:77197150	872	pre-Ta	chr18_41775014	chr18_41775014	traced
chr15:77065768	1096	Ta-1	Not assigned	chr7_70197328	untraced
chr7:121722982	705	Ta-1	chr4_93638307	chr4_93638307	traced
chr9:125091052	1355	pre-Ta	chr18_50343959	chr18_50343959	traced
chr6:108328887	855	Not assigned	Not assigned	chr1_146717192	untraced
chr6:37803066	2116	pre-Ta	chr3_163236941	chr3_163236941	traced
chr3:116474313	860	Ta-0	chr14_62535851	chr14_62535851	traced
chr7:128735021	6057	Ta-1	chr9_112798107	chr9_112798107	traced
chr1:43169043	331	Ta-0	Not assigned	chr15_87509891	untraced
chr1:3410257	374	Ta-1	Not assigned	chr2_112503812	untraced
chr1:117205715	1015	pre-Ta	chr7_97613656	chr7_97613656	traced
chr1:57283945	668	pre-Ta	Not assigned	chr1_102103355	untraced
chr22:16750420	2229	Ta-1	chr16_83637252	chr16_83637252	traced
chr21:40379202	376	Ta-0	Not assigned	chr4_169515501	untraced
chr2:43928496	1062	Ta-0	chr10_39508152	chr10_39508152	traced
chr6:103968788	3537	Not assigned	chr12_33864390	chr12_33864390	traced
chr13:31252924	224	Ta-1	chrX_56695884	chrX_56695884	traced
chr11:31458884	491	pre-Ta	Not assigned	chr14_61055116	untraced
chr4:107990507	1042	Not assigned	Not assigned	chr7_7465092	untraced
chr1:86381446	613	Ta-1	chr5_109259387	chr5_109259387	traced
chrY:5756012	2435	Ta-0	chr2_213567231	chr2_213567231	traced
chr18:79788336	460	pre-Ta	Not assigned	chr16_68583448	untraced
chr8:25730657	760	pre-Ta	chr1_85748519	chr1_85748519	traced
chr6:4773531	817	pre-Ta	chrX_26314417	chrX_26314417	traced
chr19:28821116	287	pre-Ta	Not assigned	chr14_30684809	untraced
chr11:33377712	266	Ta-0	Not assigned	chr3_46783105	untraced
chrX:64901139	1543	Ta-1	chr5_109259387	chr5_109259387	traced
chr12:30101013	5406	Ta-1	chr6_112703745	chr6_112703745	traced
chr2:81951742	474	Ta-0	Not assigned	chr10_85355506	untraced
chr2:174471703	537	Ta-1	chr13_34480262	chr13_34480262	traced
chr14:76760914	280	Ta-0	Not assigned	chr7_108188803	untraced
chr6:113057120	1972	Ta-0	Not assigned	chr18_70746549	untraced
chr4:13832548	399	Not assigned	chr12_74874868	chr12_74874868	traced
chr6:128507903	715	Ta-1	chr3_109199872	chr3_109199872	traced
chr8:97878227	284	pre-Ta	Not assigned	chr11_109177494	untraced
chr12:99147163	1231	pre-Ta	chr17_9615985	chr17_9615985	traced
chr3:81551978	315	Ta-0	Not assigned	chr1_95336298	untraced
chrX:53880633	329	pre-Ta	Not assigned	chr3_136963693	untraced
chr17:78111750	3832	Ta-0	chr18_70746549	chr18_70746549	traced
chr15:18212512	292	Ta-1	Not assigned	chr5_34147845	untraced
chr8:139113939	212	Ta-1	Not assigned	chr12_117376636	untraced
chr5:21681210	2014	Ta-1	chr6_112703745	chr6_112703745	traced
chr17:67160455	453	Ta-1	Not assigned	chr4_21159390	untraced
chr12:87682173	2056	pre-Ta	chr3_136963693	chr3_136963693	traced
chr7:132547500	192	Ta-0	Not assigned	chr4_61939927	untraced
chr12:67360121	292	Ta-1	chr2_153007766	chr2_153007766	traced
chr3:4298626	593	pre-Ta	chr1_237019467	chr1_237019467	traced
chr14:47286778	523	Ta-0	chr5_156061919	chr5_156061919	traced
chr2:145445537	208	pre-Ta	chr18_51689848	chr18_51689848	traced
chr8:64966543	709	Ta-1	Not assigned	chr2_175481951	untraced
chr2:197961225	294	Ta-1	Not assigned	chr2_155671334	untraced
chr7:11294396	310	Not assigned	Not assigned	chr7_110707004	untraced

chr6:159682253	820	pre-Ta	chr17_70544788	chr17_70544788	traced
chr2:135744688	986	pre-Ta	chrX_87821407	chrX_87821407	traced
chr11:126403923	542	pre-Ta	chr1_193717837	chr1_193717837	traced
chr5:6918196	120	Ta-1	Not assigned	chr7_70197328	untraced
chrX:117857687	163	Ta-0	chr5_156061919	chr5_156061919	traced

For each in silico solo-L1 event simulated from PD0270 sources, the traced source based on diagnostic nucleotide profiling is shown, and inference status is indicated (traced / untraced) based on if a correct determination of the element-of-origin is done. A random selection of 100 events is represented due to space constraints.

Supp. Table 3: Cancer genes potentially impacted by somatic alterations in CedBTN.

Lineage	Gene	Cancer role	CN range	Mutation	Coordinates	Change	Consequence
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21334105	C>A	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21334148	A>G	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21341394	G>A	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21341811	C>T	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21342122	G>A	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	MNV	1:21350429	CA>TG	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21364757	C>T	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	SNV	1:21371517	T>C	missense variant
CedBTN1	BIRC6	oncogene	5.1-8.2	MNV	1:21385057	TC>AT	missense variant
CedBTN2	BIRC6	oncogene	2.4-3.8	SNV	1:21369956	G>A	missense variant
CedBTN1	CRNKL1	Not assigned	3.3-4.7	SNV	1:62355162	T>G	missense variant
CedBTN1	DICER1	TSG	6.3-7.9	SNV	1:55267826	C>A	missense variant
CedBTN1	DICER1	TSG	6.3-7.9	SNV	1:55270266	G>A	missense variant
CedBTN2	DICER1	TSG	3.3-4	SNV	1:55268107	C>T	missense variant
CedBTN2	DICER1	TSG	3.3-4	SNV	1:55268260	C>G	missense variant
CedBTN2	DICER1	TSG	3.3-4	MNV	1:55268977	TA>AT	missense variant
CedBTN2	DICER1	TSG	3.3-4	SNV	1:55269085	A>T	missense variant
CedBTN2	DICER1	TSG	3.3-4	SNV	1:55269906	G>A	missense variant
CedBTN1	ESR1	oncogene, TSG	6.4-6.9	SNV	1:40275848	G>A	missense variant
CedBTN1	GNA11	oncogene	4.3-5.3	SNV	1:35870201	T>G	missense variant
CedBTN1	GNA11	oncogene	4.3-5.3	SNV	1:35870306	T>G	missense variant
CedBTN2	GNA11	oncogene	3.4-5.3	SNV	1:35832691	T>A	missense variant
CedBTN2	GNA11	oncogene	3.4-5.3	SNV	1:35870264	T>C	missense variant
CedBTN2	MAX	TSG	4.3-6.2	SNV	1:26598446	T>C	missense variant
CedBTN2	MEN1	TSG	3.6-5.7	SNV	1:60325745	G>A	stop gained
CedBTN2	MEN1	TSG	3.6-5.7	Indel	1:60325762	T>TC	frameshift
CedBTN2	MEN1	TSG	3.6-5.7	Deletion	1:60294218	DEL	gene ablation
CedBTN1	TSHR	oncogene	3.9-6.1	SNV	1:2092598	A>C	missense variant
CedBTN1	TSHR	oncogene	3.9-6.1	SNV	1:2093039	A>G	missense variant
CedBTN2	TSHR	oncogene	6.5-8.2	SNV	1:2094436	C>T	missense variant
CedBTN2	ERCC4	TSG	2.8-4.9	SNV	2:4484025	A>G	missense variant
CedBTN1	KEAP1	TSG	2.6-3.5	SNV	2:26222431	A>G	missense variant
CedBTN1	KEAP1	TSG	2.6-3.5	SNV	2:26223117	T>G	missense variant
CedBTN1	KEAP1	TSG	2.6-3.5	SNV	2:26223552	G>C	missense variant
CedBTN1	KEAP1	TSG	2.6-3.5	SNV	2:26223683	A>G	missense variant
CedBTN2	KEAP1	TSG	3.3-6.4	SNV	2:26223677	A>C	missense variant
CedBTN1	MAP3K1	oncogene, TSG	2.1-4.2	SNV	2:47421418	A>C	missense variant
CedBTN1	MAP3K1	oncogene, TSG	2.1-4.2	SNV	2:47429872	T>G	missense variant
CedBTN2	MAP3K1	oncogene, TSG	5.1-6.5	SNV	2:47405893	A>T	missense variant
CedBTN2	MAP3K1	oncogene, TSG	5.1-6.5	SNV	2:47406335	C>T	stop gained
CedBTN2	MAP3K1	oncogene, TSG	5.1-6.5	SNV	2:47406747	G>A	missense variant
CedBTN2	MAP3K1	oncogene, TSG	5.1-6.5	SNV	2:47429032	T>C	missense variant
CedBTN2	MAP3K1	oncogene, TSG	5.1-6.5	SNV	2:47429733	G>C	missense variant
CedBTN1	RAD17	TSG	3.6-5	Deletion	2:41968381	DEL	gene ablation
CedBTN2	CASP3	TSG	5.9-7.4	Deletion	3:52944272	DEL	gene ablation
CedBTN2	CASP3	TSG	5.9-7.4	Deletion	3:54995168	DEL	coding sequence
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33624041	A>G	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33625716	T>A	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33625788	G>C	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33638215	T>C	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33647241	A>T	missense variant

CedBTN1	FAT4	TSG	4.5-5	MNV	3:33647955	CG>GA	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33647958	G>C	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33668477	G>T	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33685593	C>T	missense variant
CedBTN1	FAT4	TSG	4.5-5	SNV	3:33689869	G>A	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33608869	T>G	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33613231	C>A	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33613336	T>C	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33613343	A>C	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33663605	G>T	missense variant
CedBTN2	FAT4	TSG	4.6-7.2	SNV	3:33689756	C>G	missense variant
CedBTN1	PMS2	TSG	3.2-4.4	SNV	3:8219901	C>T	missense variant
CedBTN2	PMS2	TSG	5.4-5.8	SNV	3:8238237	C>A	missense variant
CedBTN1	PRKAR1A	oncogene, TSG	3.9-4.5	SNV	3:37318120	C>T	missense variant
CedBTN1	PRKAR1A	oncogene, TSG	3.9-4.5	SNV	3:37346249	G>A	splice region
CedBTN2	PRKAR1A	oncogene, TSG	2.5-5.1	SNV	3:37339808	T>C	missense variant
CedBTN1	PTEN	TSG	3.5-4.9	SNV	3:38726195	A>C	missense variant
CedBTN2	PTEN	TSG	5.1-5.6	SNV	3:38726493	T>A	missense variant
CedBTN2	PTEN	TSG	5.1-5.6	Deletion	3:38730971	DEL	coding sequence
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4357437	C>A	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4359342	C>T	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4359482	C>T	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4374971	G>A	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4375011	T>G	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4375047	G>C	missense variant
CedBTN1	PTPN13	TSG	1.9-2.5	SNV	3:4390348	A>G	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4359411	A>G	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4372554	G>T	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4389597	G>A	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4389670	A>T	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4389961	C>T	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4396198	C>T	missense variant
CedBTN2	PTPN13	TSG	2.6-3	SNV	3:4397793	G>C	missense variant
CedBTN2	SUFU	TSG	5.6-9.3	Deletion	3:52944272	DEL	gene ablation
CedBTN1	TSC2	TSG	5.8-10.3	SNV	3:40972285	A>T	stop gained
CedBTN1	TSC2	TSG	5.8-10.3	SNV	3:40976941	C>T	stop gained
CedBTN1	TSC2	TSG	5.8-10.3	SNV	3:40977458	T>C	missense variant
CedBTN2	FANCG	TSG	1.3-4.6	MNV	4:42032505	GC>AT	missense variant
CedBTN2	FANCG	TSG	1.3-4.6	SNV	4:42039281	A>T	missense variant
CedBTN1	FAT1	TSG	4.9-7.3	MNV	4:47922375	TT>CC	missense variant
CedBTN1	FAT1	TSG	4.9-7.3	SNV	4:47922393	G>A	missense variant
CedBTN2	LATS1	TSG	4-5.4	SNV	4:1442140	G>C	missense variant
CedBTN1	MSH2	TSG	3.1-5.2	SNV	4:23899887	G>T	splice region
CedBTN2	MSH2	TSG	3-3.5	SNV	4:23915213	C>T	missense variant
CedBTN1	POLQ	oncogene, TSG	4.8-5.6	SNV	4:49342148	G>A	missense variant
CedBTN1	POLQ	oncogene, TSG	4.8-5.6	SNV	4:49342186	A>T	missense variant
CedBTN1	POLQ	oncogene, TSG	4.8-5.6	SNV	4:49342220	C>G	missense variant
CedBTN1	POLQ	oncogene, TSG	4.8-5.6	SNV	4:49348146	C>T	missense variant
CedBTN2	POLQ	oncogene, TSG	1.6-6	MNV	4:49344866	AG>CA	missense variant
CedBTN1	PTPRD	TSG	0.7-1.8	SNV	4:48235587	G>C	missense variant
CedBTN2	PTPRD	TSG	1.6-1.9	SNV	4:48235917	C>T	missense variant
CedBTN2	PTPRD	TSG	1.6-1.9	MNV	4:48235960	TG>CA	missense variant
CedBTN1	USP8	oncogene	6.3-7.3	SNV	4:51526885	G>A	missense variant
CedBTN1	ATRX	TSG	1.8-2.3	SNV	5:23872631	C>G	missense variant
CedBTN1	ATRX	TSG	1.8-2.3	SNV	5:23884972	G>A	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24024683	T>C	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24047687	C>G	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24047882	T>A	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24048206	T>C	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24048536	C>G	missense variant
CedBTN1	ATRX	TSG	2.3-2.7	SNV	5:24048717	C>A	missense variant
CedBTN2	ATRX	TSG	5.2-6.6	SNV	5:23883938	T>G	missense variant
CedBTN2	ATRX	TSG	5.2-6.6	SNV	5:23883960	C>T	missense variant
CedBTN2	ATRX	TSG	5.2-6.6	SNV	5:23889373	C>G	missense variant

CedBTN2	ATRX	TSG	5.2-6.6	SNV	5:23896435	C>A	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24024654	C>A	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24042566	C>T	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24045581	C>T	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24047644	T>C	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24047759	C>T	missense variant
CedBTN2	ATRX	TSG	3.3-5.4	SNV	5:24050751	C>G	missense variant
CedBTN1	FAT4	TSG	0.9-1.3	SNV	5:24387846	G>C	missense variant
CedBTN1	FAT4	TSG	0.9-1.3	SNV	5:24407535	C>G	missense variant
CedBTN2	FBXW7	TSG	4.8-10.6	MNV	5:46365231	GT>AA	missense variant
CedBTN1	PTPRT	TSG	3.5-6.1	SNV	5:40051331	G>A	missense variant
CedBTN1	PTPRT	TSG	1.2-3.6	SNV	5:48726729	C>A	missense variant
CedBTN2	PTPRT	TSG	2.6-5.5	MNV	5:40039753	GG>AA	missense variant
CedBTN1	ATP1A1	oncogene, TSG	3.9-5.1	SNV	6:7030441	C>T	missense variant
CedBTN2	ATP1A1	oncogene, TSG	6.6-8	SNV	6:7019664	A>G	missense variant
CedBTN2	ATP1A1	oncogene, TSG	6.6-8	SNV	6:7028593	C>T	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	Indel	6:34253431	A-T>A	splice donor
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34256426	A>G	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34256553	C>T	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34256612	C>T	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34256802	A>G	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34261196	A>C	missense variant
CedBTN1	CIC	oncogene, TSG	2.2-2.9	SNV	6:34264255	A>T	missense variant
CedBTN2	CIC	oncogene, TSG	2.4-4.6	SNV	6:34253427	C>T	missense variant
CedBTN2	CIC	oncogene, TSG	2.4-4.6	SNV	6:34256615	T>C	missense variant
CedBTN2	CIC	oncogene, TSG	2.4-4.6	SNV	6:34257026	C>G	missense variant
CedBTN2	CIC	oncogene, TSG	2.4-4.6	SNV	6:34262046	G>A	missense variant
CedBTN2	CIC	oncogene, TSG	2.4-4.6	SNV	6:34263791	G>T	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1962474	G>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1970110	A>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1970495	G>A	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1977246	G>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1979057	C>T	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1979456	G>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1979499	T>A	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:1980611	G>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:2002054	C>T	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:2004762	G>C	missense variant
CedBTN1	KMT2D	oncogene, TSG	4.5-5.1	SNV	6:2005095	C>A	stop gained
CedBTN2	KMT2D	oncogene, TSG	1.8-4	SNV	6:1962467	C>T	missense variant
CedBTN2	POLE	TSG	3-4.7	SNV	6:3299533	C>T	missense variant
CedBTN1	U2AF1	oncogene	7.5-9	SNV	6:16618541	C>T	missense variant
CedBTN1	CNBD1	Not assigned	3.9-4.7	SNV	7:31339189	A>T	missense variant
CedBTN1	CNBD1	Not assigned	3.9-4.7	SNV	7:31349689	A>C	missense variant
CedBTN1	NBN	TSG	3.4-4.1	SNV	7:23936675	G>A	missense variant
CedBTN1	NBN	TSG	3.4-4.1	SNV	7:23938662	C>T	missense variant
CedBTN1	NBN	TSG	3.4-4.1	Indel	7:23938706	T>TC	frameshift
CedBTN1	NBN	TSG	3.4-4.1	SNV	7:23938968	G>A	missense variant
CedBTN1	NBN	TSG	3.4-4.1	SNV	7:23939037	A>T	missense variant
CedBTN2	NBN	TSG	3.8-6.5	SNV	7:23938793	G>A	missense variant
CedBTN1	RECQL4	oncogene, TSG	2.6-4	MNV	7:9192850	TA>AT	stop gained
CedBTN1	BRIP1	TSG	2.9-3.5	SNV	8:1345269	G>T	missense variant
CedBTN1	BRIP1	TSG	2.9-3.5	SNV	8:1352181	T>G	missense variant
CedBTN1	BRIP1	TSG	2.9-3.5	SNV	8:1352212	C>T	missense variant
CedBTN2	BRIP1	TSG	2-2.6	MNV	8:1358956	AT>TC	missense variant
CedBTN2	BRIP1	TSG	2-2.6	SNV	8:1374878	T>G	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14799629	A>C	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14803793	A>T	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14828439	C>T	splice region
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14836845	C>T	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14848009	A>T	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14855716	T>C	missense variant
CedBTN1	NF1	TSG	3.3-3.9	SNV	8:14863482	C>A	missense variant
CedBTN2	NF1	TSG	2.2-3.2	Deletion	8:14575309	DEL	stop lost

CedBTN1	PPM1D	oncogene	3.2-6	SNV	8:32727106	C>G	missense variant
CedBTN2	PPM1D	oncogene	3.7-8.4	MNV	8:32727221	GG>AA	stop gained
CedBTN1	TBX3	oncogene, TSG	4.6-11.2	SNV	8:20416901	C>T	missense variant
CedBTN1	TBX3	oncogene, TSG	4.6-11.2	SNV	8:20417277	C>T	missense variant
CedBTN2	TBX3	oncogene, TSG	2.7-7.2	SNV	8:20406776	C>T	missense variant
CedBTN2	TBX3	oncogene, TSG	2.7-7.2	SNV	8:20406866	A>G	missense variant
CedBTN2	TBX3	oncogene, TSG	2.7-7.2	Indel	8:20417234	TC>T	frameshift
CedBTN2	TBX3	oncogene, TSG	2.7-7.2	SNV	8:20417390	T>C	missense variant
CedBTN1	ATR	TSG	1.6-1.8	SNV	9:17143057	C>T	missense variant
CedBTN1	ATR	TSG	1.6-1.8	SNV	9:17177581	T>C	missense variant
CedBTN1	ATR	TSG	1.6-1.8	SNV	9:17179623	C>T	missense variant
CedBTN2	ATR	TSG	5.2-5.6	SNV	9:17122610	T>C	missense variant
CedBTN2	ATR	TSG	5.2-5.6	SNV	9:17156868	C>T	missense variant
CedBTN1	BRCA2	TSG	2.4-4	SNV	9:14036918	C>A	missense variant
CedBTN1	BRCA2	TSG	2.4-4	SNV	9:14037193	C>A	missense variant
CedBTN1	BRCA2	TSG	2.4-4	SNV	9:14037220	C>T	missense variant
CedBTN1	BRCA2	TSG	2.4-4	Deletion	9:6645348	DEL	gene ablation
CedBTN2	BRCA2	TSG	5-7.7	Deletion	9:14018716	DEL	gene ablation
CedBTN2	XPA	TSG	2.8-3.7	SNV	9:25655898	G>A	missense variant
CedBTN2	XPA	TSG	2.8-3.7	SNV	9:25655971	T>C	missense variant
CedBTN2	XPA	TSG	2.8-3.7	SNV	9:25656017	G>C	missense variant
CedBTN1	FANCD2	TSG	2.8-3.9	SNV	10:26762590	G>A	missense variant
CedBTN1	MTOR	oncogene	1.9-2.7	SNV	10:4046236	A>G	missense variant
CedBTN1	MTOR	oncogene	1.9-2.7	SNV	10:4119257	G>A	missense variant
CedBTN2	FGFR3	oncogene	2.6-3.7	SNV	11:28711704	G>A	missense variant
CedBTN2	FGFR3	oncogene	2.6-3.7	SNV	11:28713106	C>T	missense variant
CedBTN1	PMS1	Not assigned	5.3-10.4	SNV	11:14458957	A>G	missense variant
CedBTN1	PMS1	Not assigned	5.3-10.4	SNV	11:14458977	C>G	missense variant
CedBTN2	PMS1	Not assigned	6.2-7.6	SNV	11:14454925	A>G	missense variant
CedBTN1	SF3B1	oncogene	5-7.2	SNV	11:12855416	C>T	missense variant
CedBTN2	SF3B1	oncogene	5.2-6	SNV	11:12843201	A>T	missense variant
CedBTN2	SF3B1	oncogene	5.2-6	SNV	11:12856281	G>T	missense variant
CedBTN2	SF3B1	oncogene	5.2-6	SNV	11:12857887	A>G	missense variant
CedBTN1	TERT	oncogene, TSG	3.4-4.5	SNV	11:30162268	C>T	missense variant
CedBTN1	TERT	oncogene, TSG	3.4-4.5	SNV	11:30171951	T>C	missense variant
CedBTN2	TERT	oncogene, TSG	3.1-4	SNV	11:30171836	C>T	missense variant
CedBTN2	TERT	oncogene, TSG	3.1-4	SNV	11:30171862	T>G	missense variant
CedBTN2	TRRAP	oncogene	5.1-8.5	SNV	11:8543190	G>C	missense variant
CedBTN2	TRRAP	oncogene	5.1-8.5	SNV	11:8551672	C>T	missense variant
CedBTN2	TRRAP	oncogene	5.1-8.5	SNV	11:8553118	G>A	missense variant
CedBTN2	TRRAP	oncogene	5.1-8.5	SNV	11:8567448	G>T	missense variant
CedBTN2	TRRAP	oncogene	5.1-8.5	SNV	11:8569761	C>A	missense variant
CedBTN2	FGFR4	oncogene	6.1-7.6	SNV	12:9295971	G>T	missense variant
CedBTN2	FGFR4	oncogene	6.1-7.6	SNV	12:9296828	C>T	missense variant
CedBTN2	FGFR4	oncogene	6.1-7.6	SNV	12:9296896	C>T	missense variant
CedBTN2	KAT7	oncogene	4.6-8.5	SNV	12:4719668	A>T	missense variant
CedBTN1	ISX	Not assigned	2.6-4	SNV	13:18490264	G>A	missense variant
CedBTN1	ISX	Not assigned	2.6-4	SNV	13:18493770	C>T	missense variant
CedBTN2	ISX	Not assigned	5.9-7.8	SNV	13:18494415	G>A	missense variant
CedBTN1	POLD1	TSG	3.8-4.8	SNV	13:32088491	T>C	missense variant
CedBTN1	SETDB1	oncogene	3.3-4.1	SNV	13:13680003	G>T	missense variant
CedBTN2	SETDB1	oncogene	2.7-3	SNV	13:13680050	G>C	missense variant
CedBTN2	DDB2	oncogene, TSG	3.2-4.8	SNV	14:1678910	A>G	missense variant
CedBTN2	PIK3CA	oncogene	2.3-3.7	SNV	14:2760775	G>C	missense variant
CedBTN2	PIK3CA	oncogene	2.3-3.7	SNV	14:2761390	G>A	missense variant
CedBTN1	TP63	oncogene, TSG	1.7-3.6	SNV	14:28661054	C>T	missense variant
CedBTN2	TP63	oncogene, TSG	1.9-2	SNV	14:28661617	G>A	missense variant
CedBTN1	BIRC3	oncogene, TSG	2-3.7	SNV	15:14239759	G>A	missense variant
CedBTN1	BIRC3	oncogene, TSG	2-3.7	SNV	15:14239852	C>G	missense variant
CedBTN1	BIRC3	oncogene, TSG	2-3.7	SNV	15:14242690	G>A	missense variant
CedBTN1	DAXX	oncogene, TSG	3.8-5.7	SNV	15:23059000	C>T	missense variant
CedBTN1	DAXX	oncogene, TSG	3.8-5.7	SNV	15:23060577	T>G	missense variant
CedBTN2	DAXX	oncogene, TSG	3-3.4	SNV	15:23072714	A>C	missense variant
CedBTN2	DAXX	oncogene, TSG	3-3.4	SNV	15:23072946	G>A	missense variant

CedBTN1	DNM2	TSG	3.9-4.2	SNV	15:12433191	G>A	splice acceptor
CedBTN1	DNM2	TSG	3.9-4.2	SNV	15:12434922	A>G	missense variant
CedBTN1	DROSHA	TSG	4.1-5.1	SNV	15:19845019	C>T	missense variant
CedBTN2	DROSHA	TSG	5.9-9.6	SNV	15:19844319	G>A	missense variant
CedBTN2	DROSHA	TSG	5.9-9.6	SNV	15:19845061	A>G	missense variant
CedBTN2	DROSHA	TSG	5.9-9.6	SNV	15:19863132	C>T	missense variant
CedBTN1	TSC1	TSG	4.7-5.6	SNV	15:3441073	G>A	missense variant
CedBTN1	TSC1	TSG	4.7-5.6	SNV	15:3441173	G>A	missense variant
CedBTN2	TSC1	TSG	3.4-5.5	SNV	15:3435708	T>C	missense variant
CedBTN2	TSC1	TSG	3.4-5.5	Deletion	15:3422532	DEL	coding sequence
CedBTN1	BCORL1	oncogene, TSG	2.5-4.4	SNV	16:23056769	T>A	missense variant
CedBTN1	LRP1B	TSG	1.5-2.3	SNV	16:1048313	G>T	missense variant
CedBTN2	LRP1B	TSG	2.5-3.8	SNV	16:1058426	A>C	missense variant
CedBTN2	LRP1B	TSG	2.5-3.8	SNV	16:1072961	C>A	missense variant
CedBTN2	MUC16	oncogene	3.4-4.8	SNV	16:22116340	G>A	missense variant
CedBTN1	RB1	TSG	3.6-5.1	SNV	16:13768100	T>A	splice donor
CedBTN1	STAG2	TSG	6.4-9.8	Indel	16:18521827	GA>G	frameshift
CedBTN2	STAG2	TSG	4.2-5	SNV	16:18523654	G>A	splice donor
CedBTN2	STAG2	TSG	4.2-5	SNV	16:18528225	C>G	missense variant
CedBTN2	STAG2	TSG	4.2-5	SNV	16:18529380	T>C	missense variant
CedBTN1	TBL1XR1	oncogene, TSG	3.4-4.9	SNV	16:15097616	C>A	missense variant
CedBTN1	KEAP1	TSG	1.6-3	Indel	17:15474229	GC>G	frameshift
CedBTN1	KEAP1	TSG	1.6-3	Indel	17:15474231	AC>A	frameshift
CedBTN2	KEAP1	TSG	3.2-5.1	SNV	17:15474100	G>A	missense variant
CedBTN2	KEAP1	TSG	3.2-5.1	SNV	17:15479357	T>A	missense variant
CedBTN2	RAD21	oncogene, TSG	3-4.5	SNV	17:8927171	T>A	missense variant
CedBTN1	SDHA	TSG	1.6-2	SNV	17:18138725	G>C	missense variant
CedBTN1	PTPRT	TSG	1.6-3.1	SNV	18:2134066	G>T	stop gained
CedBTN1	PTPRT	TSG	1.6-3.1	SNV	18:2137873	A>T	missense variant
CedBTN2	SMAD4	TSG	1.8-3.8	SNV	18:980938	G>A	missense variant
CedBTN2	ACVR1	oncogene	3.6-4	SNV	19:15051947	T>C	missense variant
CedBTN1	ERBB2	oncogene	3.8-4.3	SNV	19:14290182	C>A	missense variant
CedBTN1	ERBB2	oncogene	3.8-4.3	SNV	19:14301143	G>A	missense variant
CedBTN1	ERBB2	oncogene	3.8-4.3	SNV	19:14301235	T>C	missense variant
CedBTN1	ERBB2	oncogene	3.8-4.3	SNV	19:14301331	C>T	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14277245	G>T	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14277957	G>C	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14286097	C>G	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14299912	T>C	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14301307	T>C	missense variant
CedBTN2	ERBB2	oncogene	2.8-5.4	SNV	19:14312262	G>A	missense variant
CedBTN1	SETD2	TSG	3.5-4.6	SNV	19:19027545	C>T	missense variant
CedBTN1	SETD2	TSG	3.5-4.6	SNV	19:19029312	C>A	missense variant
CedBTN1	SETD2	TSG	3.5-4.6	SNV	19:19029720	C>T	missense variant
CedBTN1	SETD2	TSG	3.5-4.6	SNV	19:19031036	G>T	stop gained
CedBTN1	SETD2	TSG	3.5-4.6	SNV	19:19041064	C>T	missense variant
CedBTN2	SETD2	TSG	1.7-6.5	SNV	19:19024170	C>G	missense variant
CedBTN2	SETD2	TSG	1.7-6.5	SNV	19:19028314	T>A	missense variant

Abbreviations: TSG=tumour suppressor gene; SNV=single nucleotide variant; MNV=multi-nucleotide variant.

Supp. Table 4: Features of the CpGs from the MLH1 promoter methylation model.

CpG probe	Chromosome	Position	Strand	Gene	Region	Methylation status*
cg17641046	3	37034473	F	MLH1	TSS1500	Hypermethylated
cg07101782	3	37034495	F	MLH1	TSS1500	Hypermethylated
cg07064226	3	37034997	R	MLH1	TSS1500	Hypermethylated
cg06108510	3	37035063	R	MLH1	TSS1500	Hypermethylated
cg24985459	3	37035090	R	MLH1	TSS200	Hypermethylated
cg25202636	3	37035158	R	MLH1	TSS200	Hypermethylated
cg17621259	3	37035168	F	MLH1	TSS200	Hypermethylated
cg14671526	3	37035200	R	MLH1	TSS200	Hypermethylated
cg27331401	3	37035207	F	MLH1	TSS200	Hypermethylated
cg06590608	3	37035228	F	MLH1	TSS200	Hypermethylated

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Abbreviations: F=forward; R=reverse; TSS=transcription start site.
*Methylation status is described based on the comparison of MSI vs MSS tumours from TCGA-COAD/READ.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Appendix D: Declarations

Declaration of conflicts of interest

The doctoral candidate and his PhD supervisor freely agreed to present the results of this study and declare no conflicts of interest in relation to it.

Santiago de Compostela, September 29th, 2024

Martín Santamarina García

Declaration of image use rights

Permission to reproduce and/or adapt any images for which the doctoral candidate does not own the copyright have been requested.

A table containing image source, copyright holder and permission is provided next:

Image	Use	Source	Copyright holder	Permission
Figure 1	Adaptation	Stratton, Campbell, and Futreal 2009	Macmillan Publishers Limited	Granted (24/09/2024)
Figure 2	Adaptation	Haig H. Kazazian and Moran 2017	Massachusetts Medical Society	Not needed (thesis). Credit was given
Figure 3	Adaptation	Devine 2023	The Author	Not needed (CC BY). Credit was given
Figure 4	Adaptation	Rodriguez-Martin et al. 2020	The Authors, some rights reserved	Not needed (CC BY). PhD candidate is co-author
Figure 5	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 6	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 7	Own work	Ebert et al. 2021	The Authors, some rights reserved; exclusive licensee AAAS	Not needed (thesis). PhD candidate is co-author
Figure 8	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 9	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 10	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 11	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 12	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figure 13	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 14	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 15	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 16	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 17	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 18	Own work / reproduction	bioRxiv 2024.08.27.596794	The Authors	Requested
Figure 19	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 20	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 21	Own design	Wikimedia Commons inaturalist.org	Not applicable	Not applicable
Figure 22	Own design	BioRender.com	Martín Santamarina	Granted (CC-BY)
Figure 23	Reproduction	Scuba Cancers ERC project	Alicia L. Bruzos	Not needed (CC BY-SA). Credit was given
Figure 24	Adaptation	Metzger et al. 2016	Springer Nature Limited	Granted (24/09/2024)
Figure 25	Adaptation	Matias et al. 2014	Elsevier	Granted (24/09/2024)
Figure 26	Adaptation	Cosenza et al. 2022	Annual Reviews	Not needed (CC BY). Credit was given
Figure 27	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 28	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Figure 29	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Figure 30	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Figure 31	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figure 32	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 33	Own design	BioRender.com	Martín Santamarina	Granted (CC-BY)
Figure 34	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 35	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 36	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 37	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 38	Own work / reproduction	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not applicable
Figure 39	Reproduction	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not applicable
Figure 40	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Figure 41	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Figure 42	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Figure 43	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Figure 44	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Figure 45	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 1	Own work	Ebert et al. 2021	The Authors, some rights reserved; exclusive licensee AAAS	Not needed (thesis). PhD candidate is co-author
Supp. Figure 2	Reproduction	Ebert et al. 2021	The Authors, some rights reserved; exclusive licensee AAAS	Not needed (thesis) PhD candidate is co-author
Supp. Figure 3	Own work	Ebert et al. 2021	The Authors, some rights reserved; exclusive licensee AAAS	Not needed (thesis) PhD candidate is co-author

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Supp. Figure 4	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Supp. Figure 5	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Supp. Figure 6	Own work	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Supp. Figure 7	Reproduction	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Supp. Figure 8	Reproduction	Bruzos et al. 2023	The Authors, under exclusive licence to Springer Nature America, Inc.	Not needed (thesis) PhD candidate is co-author
Supp. Figure 9	Own work	This thesis	Universidade de Santiago de Compostela	Not applicable
Supp. Figure 10	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 11	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 12	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 13	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 14	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 15	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 16	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 17	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 18	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 19	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author
Supp. Figure 20	Own work	Santamarina-García et al. 2023	The Authors	Not needed (thesis) PhD candidate is co-author

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix E: Funding

The author of this thesis's contract was funded by the Xunta de Galicia, within the framework of its predoctoral contracts program, in the 2017 call (Reference: **ED481A-2017/306**).

Additionally, his research was also funded through the following projects:

- "Convenio de Colaboración entre a Consellería de Cultura, Educación e Ordenación Universitaria, a UDC e a USC para completar as axudas ao persoal investigador principal do programa Starting Grants do Consello Europeo de Investigación" (01/02/2021 - 30/06/2021)

- "Axudas para a acreditación, estruturación e mellora dos centros de investigación do SUG - CENTROS SINGULARES 2019 - CIMUS" (26/11/2021 - 30/06/2022)

- "El papel de la inestabilidad genómica temprana promovida por las integraciones del ADN del virus de la hepatitis B (VHB) en los orígenes del cáncer de hígado - GENERACIÓN DE CONOCIMIENTO 2021 - Proyectos de Investigación orientada" (10/01/2023 - 09/01/2024)

- "CONSOLIDACIÓN E ESTRUTURACIÓN 2023 - Proxectos de Excelencia (liña emerxente) - Estudio de la inestabilidad genómica promovida por retrotransposición en tumores humanos" (09/02/2024 - 08/02/2025)

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Appendix F: List of collaborators

The following section list the names and affiliations from scientific and tecnichal staff who somehow contributed to this work, enabling the completion of this PhD thesis. I would like to express my deep gratitude for your scientific contributions to our research. Additionally, I extend my sincere apologies to any collaborator who migth have been accidentally omited from this list.

Chapter 1 contributors:

The MEIGA research team

Sonia Zumalave	*Universidade de Santiago de Compostela*
Nuria P. Espasandín	*Universidade de Santiago de Compostela*
Daniel García-Souto	*Universidade de Santiago de Compostela*
Javier Temes	*Universidade de Santiago de Compostela*
Toby M. Baker	*The Francis Crick Institute*
Ana Pequeño-Valtierra	*Universidade de Santiago de Compostela*
Iago Otero	*Universidade de Santiago de Compostela*
Jorge Rodríguez-Castro	*Universidade de Santiago de Compostela*
Ana Oitabén	*Universidade de Santiago de Compostela*
Eva G. Álvarez	*Universidade de Santiago de Compostela*
Paula Otero	*Universidade de Santiago de Compostela*
Iria Díaz-Arias	*Universidade de Santiago de Compostela*
Mónica Martínez-Fernández	*Universidade de Vigo*
Peter Van Loo	*The Francis Crick Institute*
Gael Cristofari	*University Cote d'Azur*
Bernardo Rodríguez-Martin	*Universidade de Santiago de Compostela*
Jose M. C. Tubio	*Universidade de Santiago de Compostela*

The HGSVC2 research team

Peter Ebert	*Heinrich Heine University*
Peter A. Audano	*University of Washington School of Medicine*
Qihui Zhu	*The Jackson Laboratory for Genomic Medicine*
Bernardo Rodriguez-Martín	*European Molecular Biology Laboratory*
David Porubsky	*University of Washington School of Medicine*
Marc Jan Bonder	*European Molecular Biology Laboratory*
Arvis Sulovari	*University of Washington School of Medicine*
Jana Ebler	*Heinrich Heine University*
Weichen Zhou	*University of Michigan Medical School*
Rebecca Serra Mari	*Heinrich Heine University*
Feyza Yilmaz	*The Jackson Laboratory for Genomic Medicine*
Xuefang Zhao	*Harvard Medical School*
PingHsun Hsieh	*University of Washington School of Medicine*
Joyce Lee	*Bionano Genomics*
Sushant Kumar	*Yale University*
Jiadong Lin	*Xi'an Jiaotong University*
Tobias Rausch	*European Molecular Biology Laboratory*

Yu Chen	*University of Alabama at Birmingham*
Jingwen Ren	*University of Southern California*
Wolfram Höps	*European Molecular Biology Laboratory*
Hufsah Ashraf	*Heinrich Heine University*
Nelson T. Chuang	*University of Maryland School of Medicine*
Xiaofei Yang	*Xi'an Jiaotong University*
Katherine M. Munson	*University of Washington School of Medicine*
Alexandra P. Lewis	*University of Washington School of Medicine*
Susan Fairley	*European Molecular Biology Laboratory*
Luke J. Tallon	*University of Maryland School of Medicine*
Wayne E. Clarke	*New York Genome Center*
Anna O. Basile	*New York Genome Center*
Marta Byrska-Bishop	*New York Genome Center*
André Corvelo	*New York Genome Center*
Uday S. Evani	*New York Genome Center*
Tsung-Yu Lu	*University of Southern California*
Mark J.P. Chaisson	*University of Southern California*
Junjie Chen	*Temple University*
Chong Li	*Temple University*
Harrison Brand	*Harvard Medical School*
Aaron M. Wenger	*Pacific Biosciences of California*
Maryam Ghareghani	*Saarland University*
William T. Harvey	*University of Washington School of Medicine*
Benjamin Raeder	*European Molecular Biology Laboratory*
Patrick Hasenfeld	*European Molecular Biology Laboratory*
Allison A. Regier	*Washington University*
Haley J. Abel	*Washington University*
Ira M. Hall	*Yale School of Medicine*
Paul Flicek	*European Molecular Biology Laboratory*
Oliver Stegle	*European Molecular Biology Laboratory*
Mark B. Gerstein	*Yale University*
Jose M.C. Tubio	*Universidade de Santiago de Compostela*
Zepeng Mu	*University of Chicago*
Yang I. Li	*University of Chicago*
Xinghua Shi	*Temple University*
Alex R. Hastie	*Bionano Genomics*
Kai Ye	*Xi'an Jiaotong University*
Zechen Chong	U*niversity of Alabama at Birmingham*
Ashley D. Sanders	*European Molecular Biology Laboratory*
Michael C. Zody	*New York Genome Center*
Michael E. Talkowski	*Harvard Medical School*
Ryan E. Mills	*University of Michigan Medical School*
Scott E. Devine	*University of Maryland School of Medicine*
Charles Lee	*Xi'an Jiaotong University*
Jan O. Korbel	*European Molecular Biology Laboratory*
Tobias Marschall	*Heinrich Heine University*
Evan E. Eichler	*University of Washington School of Medicine*

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Chapter 2 contributors:

The SCUBA CANCERS research team

Alicia L. Bruzos	*Universidade de Santiago de Compostela*
Daniel García-Souto	*Universidade de Santiago de Compostela*
Seila Díaz	*Universidade de Santiago de Compostela*
Sara Rocha	*Universidade de Vigo*
Jorge Zamora	*Universidade de Santiago de Compostela*
Yunah Lee	*Korea Advanced Institute of Science and Technology*
Alejandro Viña-Feás	*Universidade de Santiago de Compostela*
Michael A. Quail	*Wellcome Sanger Institute*
Iago Otero	*Universidade de Santiago de Compostela*
Ana Pequeño-Valtierra	*Universidade de Santiago de Compostela*
Javier Temes	*Universidade de Santiago de Compostela*
Jorge-Rodríguez-Castro	*Universidade de Santiago de Compostela*
Leyre Aramburu	*Universidade de Santiago de Compostela*
André Vidal-Capón	*Universidade de Vigo*
Antonio Villanueva	*Universidade de Vigo*
Damián Costas	*Universidade de Vigo*
Rosana Rodríguez	*Universidade de Vigo*
Tamara Prieto	*Universidade de Vigo*
Laura Tomás	*Universidade de Vigo*
Pilar Alvariño	*Universidade de Vigo*
Juana Alonso	*Universidade de Vigo*
Asunción Cao	*Centro de Investigacións Mariñas*
David Iglesias	*Centro de Investigacións Mariñas*
Maria J. Carballal	*Centro de Investigacións Mariñas*
Ana M. Amaral	*Universidade do Algarve*
Pablo Balseiro	*University of Bergen*
Ricardo Calado	*Universidade de Aveiro*
Bouchra El Khalfi	*Hassan II University of Casablanca*
Urtzi Izagirre	*University of the Basque Country*
Xavier de Montadouin	*University of Bordeaux*
Nicolas G. Pade	*European Marine Biology Resources Centre*
Ian Probert	*Sorbonne University*
Fernando Ricardo	*University of Aveiro*
Pamela Ruiz	*University of the Basque Country*
Maria Skazina	*St Petersburg State University*
Kataryna Smolarz	*University of Gdańsk*
Juan J. Pasantes	*Universidade de Vigo*
Antonio Villalba	*Centro de Investigacións Mariñas*
Zemin Ning	*Wellcome Sanger Institute*
Young Seok Ju	*Korea Advanced Institute of Science and Technology*
David Posada	Universidade de Vigo
Jonas Demeulemeester	*The Francis Crick Institute*
Adrián Baez-Ortega	*University of Cambridge*
José M.C. Tubio	*Universidade de Santiago de Compostela*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Chapter 3 contributors:

The MSIMEP research team

Jenifer Brea-Iglesias	*Universidade de Santiago de Compostela*
Jesper Bertram Bramsen	*Aarhus University Hospital*
Mar Fuentes-Losada	*Universidade de Santiago de Compostela*
Francisco Javier Caneiro	*Universidade de Santiago de Compostela*
Jose Ángel Vázquez-Bueno	*Complejo Hospitalario Universitario de Ferrol*
Héctor Lázare-Iglesias	*Universidade de Santiago de Compostela*
Natalia Fernández-Díaz	*Universidade de Santiago de Compostela*
Laura Sánchez-Ribadulla	*Complejo Hospitalario Universitario de Ferrol*
Yoel Z. Betancor	*Universidade de Santiago de Compostela*
Miriam Ferreiro-Pantín	*Universidade de Santiago de Compostela*
Pablo Conesa-Zamora	*Santa Lucía University Hospital*
Jose Ramón Antúnez-López	*Universidade de Santiago de Compostela*
Masahito Kawazu	*Chiba Cancer Center*
Manel Esteller	*University of Barcelona*
Claus Lindbjerg Andersen	*Aarhus University Hospital*
Jose M. C. Tubio	*Universidade de Santiago de Compostela*
Rafael López-López	*Universidade de Santiago de Compostela*
Juan Ruiz-Bañobre	*Universidade de Santiago de Compostela*

Other relevant contributors

Elizabeth P. Murchison	*University of Cambridge*
Kevin Gori	*University of Cambridge*
Kunal Sharan	*CSIR-Central Food Technological Research Institute*
Luis Llaneza	*A.RE.NA, Asesores en Recursos Naturales, S.L.*
Marta Ferrón	*Universidade de Santiago de Compostela*
Mónica Martínez	*Universidade de Santiago de Compostela*
Richard Möller	*Universidad de la República de Uruguay*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA





Genomic instability is a hallmark of cancer and a key driver of tumor somatic evolution. However, the molecular determinants underlying this process remain poorly understood. In this thesis, we investigated three distinct mechanisms driving genomic instability across different cancer models. Using novel bioinformatic approaches, we analyzed genetic variants in these tumors. First, we employed long-read sequencing to uncover new insights into L1 retrotransposition activity in human cancers. Next, we examined patterns of structural variation to explore the evolution of bivalve transmissible cancer genomes. Finally, we developed a GLM predictor to identify microsatellite instability in colorectal cancer. Collectively, these findings enhance our understanding of the molecular basis of genomic instability in cancer.