TESE DE DOUTORAMENTO

IDENTIFICATION OF NEW GENOMIC DRIVERS AND PREDICTORS OF DISEASE EVOLUTION IN CHRONIC LYMPHOCYTIC LEUKEMIA

Adrián Mosquera Orgueira

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN MEDICINA CLÍNICA

SANTIAGO DE COMPOSTELA

ANO 2021

AUTORIZACIÓN DO DIRECTOR / TITOR DA TESE

“Identification of new genomic drivers and predictors of disease evolution in chronic lymphocytic leukemia”

D. José Luis Bello López

INFORMA:

Que a presente tese, correspóndese co traballo realizado por D. Adrián Mosquera Orgueira, baixo a miña dirección, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co artigo 37 do Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de COMPENDIO DE PUBLICACIÓNS, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Santiago de Compostela, 11 de Xaneiro de 2021

Asdo.

José Luis Bello López

DECLARACIÓN DO AUTOR/A DA TESE

D. Adrián Mosquera Orgueira

Título da Tese: **Identification of new genomic drivers and predictors of disease evolution in chronic lymphocytic leukemia**

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De selo caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En Santiago de Compostela, 11 de Xaneiro de 2021

Asdo

Adrián Mosquera Orgueira

INDEX

INDEX	**PAGE 2**
CHAPTER 1: *SUMMARY IN GALICIAN*	**PAGES 3-15**
CHAPTER 2: *INTRODUCTION*	**PAGES16-17**
CHAPTER 3: *OBJECTIVES*	**PAGE 18**
CHAPTER 4: *METHODOLOGY*	**PAGES 19-23**
CHAPTER 5: *DISCUSSION*	**PAGES 24-33**
CHAPTER 6: *CONCLUSIONS*	**PAGE 34**
CHAPTER 7: *REFERENCES*	**PAGES 35-46**
CHAPTER 8: *APPENDIX 1*	**PAGE 47**
CHAPTER 9: *APPENDIX 2*	**PAGE 48**
CHAPTER 10: *APPENDIX 3*	**PAGES 49-51**
CHAPTER 11: *APPENDIX 4: PUBLISHED OR ACCEPTED MANUSCRIPTS*	**PAGES 52-129**

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1 RESUMO

1.1 Introducción

A leucemia linfocítica crónica (LLC) e unha síndrome linfoproliferativa de células B cunha incidencia anual estimada en países occidentais de 6.9 casos por 100.000 habitantes/ano e unha importante variación entre razas.[1] A incidencia da LLC é maior nos homes que nas mulleres, incrementándose progresivamente a partir dos 35 anos de idade ata as últimas décadas da vida.[2] A LLC caracterízase pola súa heteroxeneidade clínica e xenética. *Döhner e cols.* (2000) describiron a ampliamente utilizada clasificación citoxenética da LLC baseada nas alteraciones cromosómicas máis frecuentes no xenoma desta patoloxía, que son a trisomía 12 e as delecións en 13q14.2–14.3, 11q22.3 e 17p13.1.[3] Dende aquel momento, os estudos de xenoma completo revelaron a existencia doutras altearcións moleculares recurrentes, tales como a trisomía do cromosoma 19, as amplificacións de 2p e 8q, e as delecións de 8p, 6q21, 18p, e 20p.[4,5] De xeito semellante, unha gran cantidade de alteracións xenéticas e epixenéticas modulan a agresividade clínica da LLC,[6] tal como as mutacións de *NOTCH1*, *SF3B1*, *ATM*, *TP53* e *POT1* e a ausencia de hipermutación somática na rexión variable do xen das cadeas pesadas das inmunoglobulinas (*IGHV*). Tamén se apreciou que as alteracións no número de copias (CNAs, polas súas siglas en inglés) no xenoma da LLC tenden a adquirirse precozmente e mantéñense estables no tempo, mentres que a heteroxeneidade mutacional tende a incrementarse paulatinamente.[6] De feito, unha cantidade crecente de evidencia científica indica que a acumulación destas alteracións modula a proliferación la LLC e a súa agresividade clínica dun xeito importante,[7,8] comportándose como directores da complexidade xenómica e evolución clonal,[9-11] e acumulándose en caso de recaída.[12,13] Ademais, a variación xenómica xerminal tamén pode predispor a padecer LLC, un feito que foi o foco de varios estudos de asociación de xenoma completo (GWAS, polas súas siglas en inglés). Neste sentido, doceas de variantes comúns en xenes como *BCL2*, *EOMES*, *CASP10* e *POT1* foron asociadas cun incremento do risco de desenvolver LLC.[14-16] Ademais, estudos GWAS noutras patoloxías linfoproliferativas como o linfoma folicular e o linfoma B difuso de células grandes evidenciaron que existe asociación entre a presenza de certas variantes xerminais e a supervivencia global ou libre de progresión dos doentes.[17,18] A pesar disto, a maioría das análisis sobre o xenoma da CLL centráronse ata o de agora nas alteracións somáticas adquiridas de xeito case exclusivo.

Unha meirande cantidade das mutacións no xenoma do cancro acontecen en rexións non codificantes de proteína, e a súa función aínda se está comenzado a atisbar nestes momentos.[19] O ADN non codificante de proteína supón o 98% do xenoma, pero estudos recentes amosan que a maioría destas secuencias forman parte de rexións reguladoras ou ben son activamente transcritas a ARN.[20,21] Estas mutacións poden inducir cambios funcionais a nivel xenómico a través da alteración da unión de factores de transcripción ou mediando reestructuracións da cromatina.[20,22] Por exemplo, as mutacións nas rexións non transcritas 5' e 3' poden alterar a estructura tridimensional do ARN, modificar a unión de microARN ou alterar sinais de poliadenilación.[20] De xeito semellante, as mutacións en xenes non codificantes de proteína tales como os microRNA e os RNA longos interxénicos (lincARNs) son recoñecidos mecanismos causantes de cancro.[20,23] Distintos estudos evidenciaron que a expresión de xenes tales como *BRCA1*, *CDH10*, *CCND1*, *MALAT1*, *PAX5*, *RB1*, *SDHD*, *TERT*, *TOX3* e *TAL1* se ve influída por mutacións somáticas en rexións non codificantes de proteína que actúan como reguladores da expresión xénica.[19,24,25] O proxecto *Pancancer Analysis of Whole Genomes* (PCAWG) revelou a existencia de rexións recurrentemente mutadas preto de importantes xenes oncoxénicos no xenoma do

cancro, podendo ser estas tanto comúns a diferentes tipos de cancro como específicos de cada tipo de tumor.[25] Algunhas destas mutacións oncoxénicas poden inducir alteracións estructurais da conformación da cromatina a grande escala e condicionar a expresión anormal de oncoxenes e xenes supresores tumorais a longa distancia.[26] Ademais, a distribución deste tipo de mutacións ao longo do xenoma non é aleatoria. *Hornshøj e cols.* (2018) identificaron un enriquecimiento significativo en mutacións no xenoma do cancro que afectan a secuencias consenso CCCT de unión ao factor *"CCCT-binding factor"* (CTCF, polas súas siglas en inglés).[24] De xeito similar, *Liu e cols.* (2019) identificaron 21 rexións illantes ricas en CTCF que se atopaban recurrentemente mutadas no xenoma do cancro, e demostraron elegantemente que algunhas destas mutacións inducen a proliferación tumoral.[27]

Até o de agora, o tratamento da LLC retrásase ata a progresión da enfermidade (fallo da médula ósea, organomegalias, desenvolvemento de sintomatoloxía xeral ou transformación a linfoma de alto grado), así como no caso infrecuente do desenvolvemento de citopenias autoinmunes refractarias.[28,29] Pola contra, debido ao desenvolvemento de novos tratamentos dirixidos contra a tirosín cinasa de Bruton (Ibrutinib[30]), contra a fosfatidilinositol-3-cinasa (idelalisib[31]) ou contra a proteína antiapoptótica BCL2 (venetoclax[32]), é cada vez más tentadora a idea de tratar a algúns individuos dende momentos máis precoces, cando a masa tumoral sexa menor e os doentes posúan unha mellor situación física. Polo tanto, a mellora na caracterización biolóxica da LLC e a detección de novos biomarcadores xenómicos premitirá crear terapias máis personalizadas para esta patoloxía.

1.2 Obxectivos

Os principais objectivos desta tese son os seguintes:

1. Detectar novas mutacións causantes de cancro no xenoma da LLC e analizar a súa importancia pronóstica.
2. Identificar variantes xerminais comúns que condicionen o pronóstico dos doentes con LLC tratados con inmunoquimioterapia.
3. Mellorar a caracterización das áreas do xenoma non codificantes de proteínas, particularmente en rexións con actividade regulatoria (epixenética) coñecida.
4. Identificar novas aberracións estructurais no xenoma dos doentes con LLC, así como a súa asociación co estado mutacional de *IGHV* e a súa importancia pronóstica.
5. Identificar novos patróns de expresión xénica con valor pronóstica en LLC.

1.3 Resultados e discusión

Nesta tese analizamos información xenómica masiva derivada da análise de centos de doentes con LLC incluidos en varios dos proxectos de acceso público de maior impacto no campo da xenómica desta patoloxía. Os nosos resultados amosan a existencia dunha serie de novas mutacións recurrentes e/ou infrecuentes como inductoras a nivel somático deste tipo de leucemia, algunhas das cales tamén albergan información pronóstica naqueles doentes tratados con reximes consistentes en inmunoquimioterapia. Ademas, tamén observamos a existencia de variantes a nivel xerminal asociadas coa evolución clínica da LLC. Finalmente, aportamos evidencia que indica que a información molecular deste tipo de tumor pódese empregar para predecir os resultados en vida real dos doentes independentemente de outros biomarcadores comúnmente empregados nesta patoloxía tales como as alteracións citoxenéticas e o estado de mutación de *IGHV*. Polo tanto, esta tese cubre diferentes aspectos sobre a xenómica do cancro:

dende o descubrimento de novas mutacións con potencial oncoxénico ata o descubrimento de novos biomarcadores que poderán ser empregados para a creación de novos modelos predictivos de evolución tumoral. Polo tanto, cremos que esta capacidade de trasladar a heteroxeneidade molecular da LLC en información predictiva e útil a nivel clínico anticipa unha nova era na medicina personalizada desta desorde linfoproliferativa.

Ao longo da discusión destes resultados, centraremos inicialmente os nosos esforzos na identificación de novos eventos somáticos no xenoma da LLC, así como nas súas correlacións cos subgrupos citoxenómicos comúnmente usados na práctica actual e coa supervivencia dos doentes. Posteriormente, focalizaremos a nosa pescuda na asociación de certas variantes en línea xerminal co pronóstico dos doentes afectados de LLC. Finalmente, presentaremos e discutiremos a utilidade dos algoritmos de intelixencia artificial para a mellora da estratificación pronóstica da LLC en base á análise de perfís moleculares.

1.3.1 Novas mutacións somáticas no xenoma da LLC

A análise de centos de xenomas de LLC contribuíu a esclarecer os mecanismos xenómicos detrás desta patoloxía.[5,33] *Puente e cols.* analizaron unha cohorte de 506 casos de orixe española, e puideron identificar 36 xenes recurrentemente mutados, así como 23 xenes adicionais con mutación diferencial entre doentes con e sen mutación de *IGHV*.[33] Por outra banda, *Landau e cols.* analizaron unha cohorte de 538 doentes e identificaron unha lista de 44 xenes recurrentemente mutados.[5] Comparando estes resultados, tan só 28 xenes foron comunmente identificados por ambos estudos como recurrentemente mutados no xenoma da LLC. Polo tanto, os resultados discordantes ben se puideron deber a diferencias nas características basais dos doentes, erros na técnica de secuenciación, mutacións con prevalencia diferente nas distintas poboacións analizadas ou a diferenzas metodolóxicas na análise dos datos.

A detección de mutacións por tecnoloxías de secuenciación masiva supón unha importante limitación no desenvolvemento desta técnica. Difereznas na clonalidade, pureza de mostra, cobertura de secuenciación e calidade conlevan dificultades para a maioría dos algoritmos de detección de mutacións, que intentan resolver de diversas maneiras. Aqueles métodos coa maior sensibilidade acompáñanse frecuentemente dunha menor precisión, o cal conleva a diferenzas notables entre diferentes métodos de identificación de mutacións.[34] Polo tanto, é altamente probable que moitas mutacións patoxénicas pasasen desapercibidas a grandes proxectos de secuenciación por este motivo, o cal implica a probable existencia de numerosas mutacións oncoxénicas pendentes de identificar. Neste traballo, realizamos unha análise complementaria da bases de datos de LLC do *International Cancer Genome Consortium* (ICGC) que incluíu 49 doentes con linfocitose B monoclonal e 390 doentes con LLC sen tratamiento previo.[35] A detección de mutacións realizouse por dous métodos complementarios, e posteriormente empregáronse diferentes metodoloxías para diferenciar aquelas mutacións patoxénicas recurrentes ou infrecuentes das mutacións pasaxeiras sen significación patolóxica.

Un total de 28.350 mutacións foron detectadas nos 439 doentes analizados, das cales 12.057 afectaron rexións codificantes de proteína. Entre estas, 8.965 foron non silentes e 3.095 resultaron silentes. A ampla maioría das mutacións non silentes foron de tipo missense (7.558 eventos). Detectáronse 66 xenes enriquecidos en mutacións oncoxénicas, dos cales 32 foran previamente identificados por *Puente e cols*.[33] Entre os novos xenes, os máis frecuentemente mutados foron *DTX1*, *LPHN3*, *LRP1B*, *LTB* e *WDFY3*. Ademas, os xenes *BIRC6*, *DOCK1*, *KMT2C*/*MLL3*, *PTPRB* e *PTPRT* víronse afectados por unha mutación silente que predictivamente crea un novo punto de splicing críptico. Finalmente, observamos mutacións en

FREM1 en 4 casos, ademais de 2 mutacións sinónimas que curiosamente afectaron o mesmo nucleótido.

A maioría dos novos xenes causantes de LLC propostos teñen un rol ben definido na carcinoxénese, tal como é o caso de *EPHA7*,[36] *MYCBP2*[37] e *PTPRM*.[38] Outros foron previamente vinculados coa oncoxénese, como o regulador da autofaxia *WDFY3*,[39] o xen regulador da vía de Notch *DTX1*,[40] e os xenes das latrofilinas *LPHN2* e *LPHN3*.[41] De xeito semellante, as mutacións en *CARD11* e *SI* foron previamente descritas en LLC,[42] mentres que os xenes *BIRC6* e *KMT2C*/*MLL3* son parálogos dos xenes *driver* de LLC *BIRC3* and *KMT2D*. Así mesmo, detectamos mutacións patoxénicas e infrecuentes en 60 xenes, con tendencia a afectar a coñecidos *drivers* de cancro (*EGFR*, *ERBB4*, *MAP2K1*, *NF1*, *NFKB1*, *NOTCH3* e *SRSF1*), incluído xenes vinculados coa linfomaxénese tales como *BAX*, *BCOR*, *BCR*, *BTG2*, *DIS3*, *IKZF3*, *KRAS*, *PPM1D*, *PTPN11*, *SETD1B*, *TLR2* e *TRAF3*. Nesta longa lista tamén se aprecian xenes de vías asociadas á regulación dos linfocitos (*CD19*, CD36 e *ALCAM*) e de vías de sinalización oncoxénicas tales como Notch (*NOTCH3*, *DMXL2* e *SBNO1*), WNT/β-catenina (*DACT1*); polimerización do ADN (*POLE*) e regulación epixenética (*KDM5A*, *HIST1H1D*, *PHF1* e mutacións únicas en *HIST1H2BC* e *HIST1H2BG*). Ademas, mutacións illadas e patoxénicas en oncoxenes e xenes supresores tumorais importantes tamén se apreciaron entre os resultados máis significativos, tales como *EP300*, *KIT*, *MELK* e *PTEN*.

1.3.2 Mutacións en rexións non codificantes de proteína no xenoma da LLC

Esforzos recentes por *Puente e cols.* permitiron o descubrimento de 24 rexións non codificantes de proteína recurrentemente mutadas no xenoma da LLC, algunhas das cales se asocian con cambios funcionais tales como a expresión dos xenes *NOTCH1* e *PAX5*.[33] Porén, a escaseza de anotacións funcionais das rexións non codificantes e a dificultade para clasificar as mesas como patoxénicas dificultan unha mellor comprensión deste tipo de mutacións no xenoma do cancro, as cales probablemente albergan múltiples elementos reguladores pendentes de ser descubertos. Neste sentido, analizamos os datos de secuenciación de xenoma completo (WGS, polas súas siglas en inglés) usando un protocolo de mellores prácticas para a detección de mutacións. Posteriormente, identificamos sinais de selección positiva nas rexións reguladoras, e finalmente intentamos descubrir cales destas mutacións se asocian con expresión aberrante dos xenes adxacentes. Os nosos resultados indican a existencia de doceas de rexións regulatorias enriquecidas en mutacións nos *loci* dos xenes asociados a cancro e vías inmunes, algunhas das cales inducen unha alteración na expresión xénica a nivel local.

De entre estas rexións, 10 foron previamente descritas por *Puente e cols.*, en particular aquelas nos loci de *BACH2*, *BLC2*, *BCL6*, *BCL7A*, *BIRC3*, *S1PR2*, *PCDH15*, *ZCCHC7*/*PAX5* e *ZFP36L1*.[33] Numerosas novas rexiósn recurrentemente mutadas tamén foron atopadas, incluído puntos de unión a factores de transcripción, rexións hipersensibles ao ADNase, rexións 5' non traducidas, promotores, rexións potenciadoras da expresión e ARNs non codificantes de proteína. Estes eventos tenderon a acumularse preto de xenes vinculados con vías oncoxénicas. As zonas máis mutadas atopáronse nun punto de unión a SETB1 no primeiro intrón de *DAP3* (unha proteína de unión a GTP que participa na vía da apoptose[43]); e nunha rexión hipersensible a ADNase preto de *ING2*, un xen supresor tumoral ben caracterizado[44]). De entre todos os novos achados, tan só 3 xenes foran previamente definidos como dianas de hipermutación somática aberrante en linfomas B (*LTB*, *MALAT1* and *ST6GAL1*).[45] Ademais, algunhas destas rexións mutadas asociáronse a cambios na expresión xénica, como no caso de *PHF2*, *S1PR2* e *RPL39L*. Estes 3 xenes están involucrados na regulación de importantes procesos oncoxénicos. *PHF2*

codifica unha histona demetilase con actividade supresora tumoral.[46] *S1PR2* regula a vía de TGF-β e exerce de supresor tumoral en linfomas B.[47] Finalmente, *RPL39L* está involucrado na perpetuación das células nai cancerosas e na resposta á hipoxia.[48] Estes resultados son concordantes con outros datos que indican que as mutacións non codificantes de proteína poden actuar como *drivers* de linfomaxénese.[49-51]

1.3.3 Importancia das mutacións somáticas sobre o pronóstico da LLC

Identificamos fenómenos de hipermutación somática nunha rexión de 1.543 pares de bases localizada no flanco e rexión non traducida 5' do xen *IGKC* (172 doentes, 40% da poboación global). Estas mutacións atopáronse asociadas de forma moi intensa tanto co tempo ao primeiro tratamento (p-valor 7.23×10^{-11}, HR 0.21–0.44) independentemente do estado mutacional de *IGHV*, o sexo dos doentes e o estadio clínico ao diagnóstico (p-valor 6.3×10^{-3}, q-value 3.7×10^{-2}, HR 0.39–0.86). De xeito semellante, atopamos unha asociación significativa cunha maior supervivencia global (p-valor 2.81×10^{-4}), que non resultou ser independente doutras covariables (p-valor 0.54). Esta rexión inclúe varias secuencias codificantes de fragmentos da inmunoglobulina lixeira kappa (en concreto *IGKJ1*, *IGKJ2*, *IGKJ3*, *IGKJ4* e *IGJK5*). A pesares de que estes xenes estaban mutados en 126 casos, a maioría dos mesmos (96%) albergaron mutacións concurrentes na rexión adxacente non codificante.

1.3.4 Análise de rutas biolóxicas enriquecidas en mutacións e o seu impacto pronóstico en doentes con LLC

Identificamos 62 rutas biolóxicas enriquecidas significativamente en mutacións non-sinónimas (p-valor axustado por Bonferroni < 0.1). As rutas máis significativamente mutadas foron as de *"Retinoblastoma"*, *"TP53"*, *"ATM"*, *"Sinalización apoptótica en resposta ao dano do ADN"*, *"Ruta de TP53 asociada a hipoxia"* e *"Ruta G1"*. De xeito importante, 4 rutas biolóxicas significativas non albergan ningún xen driver de alta frecuencia, en particular *"Ruta de CDK5"*, *"Ruta apoptose inducida por fragmentación do ADN"*, *"Cascada mediada por FRS2"* e *"Cascada de RAF MAP quinase"*. Ademais, observamos que as mutacións en *"Ruta efectora de TP53"* se asociaron de forma intensa co tempo ao primeiro tratamento dos doentes (p-value 3.80×10^{-5}), e que isto resultou independiente dos casos con mutación de *TP53* (p-value 5.3×10^{-4}). Estas mutacións tamén estiveron asociadas con mejor supervivencia glboal (p-value 2.81×10^{-4}), se ben non de forma independente do estado mutacional de *IGHV* (p-value 0.54). Estes resultados suxiren que a disrupción da vía de *TP53*, e non só deste xen, xoga un rol activo na patoxénese da LLC.

1.3.5 Novas alteracións estructurais focais no xenoma da LLC

Detectamos unha serie de novas CNAs e perdas de heterozigosidade no xenoma da LLC por medio do uso de datos de secuenciación de exomas, o cal nos permite detectar alteracións de menor tamaño que as detectadas mediante as técnicas de arrays. Os nosos resultados non só reproducen o perfil citoxenético coñecido da LLC, senon que tamén revelan a existencia de novos eventos afectando a xens de rutas clave na oncoxénese, nalgúns casos con implicacións pronósticas asociadas.

En particular, encontramos 54 CNAs de pequeño tamaño no xenoma da LLC, con tendencia a afectar xenes importantes en rutas oncoxénicas. Por exemplo, cinco xenes recurrentemente amplificados (*HFM1*, *ANAPC10*, *TAF5*, *COBRA1* e *SYCE3*) e un xen delecionado (*FAM175A*/*ABRAXAS*) están vinculados coa transcripción, replicación e reparación do ADN. Os

xenes amplificados *PHF21A*, *PCGF6* e *SUZ12* codifican reguladores epixenéticos con actividade represora sobre a expresión xénica,[52-54] mentres que os xenes delecionados *AMD1*, *TP53INP1*, *ULK1* e *NFATC1* xogan diversos roles na oncoxéntese.[55-58] Ademais, detectamos a presenza de fenómenos de perdas de hetercigosidade con número de copias neutro nos loci de *ATM*, *NOTCH1*, *TP53*, *ARID1A*, *ASXL1*, *CREBBP* e *PI4KB*/*IL6R* loci, así como na rexión telomérica de 11p. Dado que só unha parte destas alberga mutacións nos xenes driver correspondentes, estos achádegos suxiren que outros mecanismos de patoxeneicidade están pendentes de ser identificados.

Ademais, a análise da interralación entre as alteracións estructurais e o estado mutacional de *IGHV* permitíunos observar que os doentes con trisomía do cromosoma 12 e *IGHV* mutado forman un grupo pronóstico adverso tanto en termos de supervivencia libre de tratamiento como en supervivencia global. Outros achádegos pronósticos importantes foron a asociación das delecións en *SETD2*, 11q22.3 e 14q32.33 con curto tempo ao primeiro tratamiento, e as ganancias de *IRF4* coa menor supervivencia global.

1.3.6 Variantes xerminais asociadas coa evolución clínica da LLC

A predisposición herdada ao desenvolvemento de LLC foi estudada por diversas análises a nivel de xenoma completo, onde se atoparon ducias de variantes comúns en xenes como *BCL2*, *EOMES*, *CASP10* e *POT1* asociadas co risco de padecer esta patoloxía.[14-16] Nesta tese, abordamos a asociación de variantes xenéticas comúns co tempo ao primeiro tratamiento e a supervivencia global dos doentes con LLC incluidas na cohorte do ICGC. Os nosos resultados máis reseñables suxiren unha modulación da agresividade da LLC por variantes comúns. O achádego máis importante neste sentido foi a asociación de rs7620924 preto do xen *PPP4R2* cun tempo máis curto ao tratamento. Esta proteína participa na regulación da supervivencia celular e a reparación do ADN nas células de leucemia aguda,[59] e de feito o complexo no que participa (proteina-fosfatase 4, PPP4) resulta esencial para a reparación do ADN e a maduración axeitada dos linfocitos B ao seu paso polo centro xerminal.[60-62] Outros polimorfismos nos xenes *MAP3K4*, *PEX26* and *TTLL12* tamén se viron asociados de forma significativa co tempo ao primeiro tratamento. Ademais, unha análise que integra toda a variabilidade a nivel de xenes completos permitíu detectar como dita heteroxeneidade xerminal se asocia coa supervivencia dos doentes. Os xenes máis relevantes nesta análise resultaron ser *RIP3K* (regulardor da morte celular[63]), *NIFK* (codificante dunha proteína pro-proliferativa[64]), *SIK1* (supresor tumoral[65,66]), *ZCCHC7* (adxacente ao xen con expresión específica en linfocitos *PAX5*[67,68]), *CLUAP1* (regulador de citoesqueleto e de crecemento tumoral[69-71]) e *GAMT* (enzima metabólica que axuda a satisfacer as necesidades enerxéticas das células cancerosas[72]).

A principal limitación desta análise e a carencia dunha cohorte independente para a validación dos resultados. No obstante, os nosos resultados son tanto estatística como biolóxicamente plausibles, polo que está xustificado levar a cabo novos experimentos que estuden os mecanismos polos que estas variantes exercen efecto sobre o pronóstico da LLC.

1.3.7 Predicción do tempo ao primeiro tratamento en base a perfís de expresión xénica

As alteracións xenómicas no cancro inducen cambios na expresión xénica que poden ser medidos a través de tecnoloxía de secuenciación masiva do ARN (RNAseq pola súas siglas en inglés). A definición de patróns de expresión xénica con implicacións clínicas é unha estratexia que permite pechar o círculo entre a investigación básica e a práctica clínica. Neste traballo,

identificamos un patrón de expersión xénica capaz de predicir o tempo ao tratamento dos doentes con LLC cun grupo relativamente pequeno de xenes.

En particular, usáronse datos de RNAseq de doentes incluídos na cohorte CLLE-ICGC. Usando un algoritmo de Modelado Mixto Gaussiano - Maximización da Esperanza (GMM-EM polas súas siglas en inglés), fomos capaces de estratificar os doentes de LLC en dous grupos en base á expresión de 290 xenes, e observamos que este patrón era predictivo de tempo ao primeiro tratamento independentemente do estado mutacional de *IGHV*. En particular, identificamos un grupo de doentes con LLC e *IGHV* mutado cun patrón transcriptómico de baixo risco que só precisou tratamiento nun 25% de casos ao longo da evolución da enfermidade. Dous grupos adicionais (un formado por doentes con *IGHV* mutado e un perfil transcriptómico de algo risco e outro formado por doentes *IGHV* non mutado e perfil transcriptómico de baixo risco) amosaron unha evolución similar de risco intermedio, e finalmente un grupo composto por doentes con *IGHV* non mutado e un perfil transcriptómico de alto risco amosou ter a maior necesidade de tratamento ao longo da evolución da patoloxía. No mesmo traballo, presentamos un novo algoritmo de intelixencia artificar capaz de predecir con grande precisión a necesidade de tratamento durantes os 5 primeiros anos dende o diagnóstico. Estas ferramentas poderán ser usadas no futuro para identificar aqueles doentes con LLC de alto risco de progresión que poidan beneficiarse dunha intervención terapéutica máis precoz.

1.3.8 Identificación de doentes con LLC en algo risco de mortalidade precoz en base a subgrupos moleculares

De forma análoga ao apartado anterior, razoamos que as tecnoloxías de aprendizaxe automatizado sobre datos moleculares tamén poderían axudar a indentificar doentes con LLC en alto risco de morte precoz. Usando o algoritmo GMM-EM puidemos detectar que a expresión de 3 xenes se asociada significativamente coa supervivencia dos doentes (p-valor axustado por Benjamini–Hochberg <0.05). Estes xenes foron *SCGB2A1*, *KLF4* e *PPP1R14B*. A clusterización multivariante destes 3 xenes asociouse de forma marcada coa supervivencia global (p-value 4.31×10^{-6}, hazard ratio 4.86, regresión de cox). O clúster de doentes de mal pronóstico supuxo un 4.22% da cohorte de descubrimento. Este impacto pronóstico puido ser validado nunha cohorte independente (p-valor 5.7×10^{-6}, hazard ratio 10.79, regresión de cox), representando un 5.60% de dita cohorte. Utilizando os datos de mutacións dispoñibles na cohorte de descubrimento (*Puente e cols.*[33]) puidemos comprobar que dita capacidade predictiva resultou independente das principais variables clínicas (idade, estadio Binet) e moleculares (mutación de *IGHV*, deleción de 17p ou mutación de *TP53*, deleción de 11q ou mutación de *ATM*, mutación de *NOTCH1*, mutación de *SF3B1* e mutación de *BIRC3*). En conclusión, reportamos un patrón de expresión xénica baseado en 3 tránscritos que é capaz de identificar un 5% de doentes con LLC con supervivencia curta. O impacto pronóstico deste perfil resultou independente dos principais marcadores citoxenómicos de mal pronóstico. O pequeno tamaño do patrón facilitará a súa aplicación en futuros estudos sobre estratificación pronóstica e resposta a novos fármacos no eido da LLC.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.4 Conclusións

As principais conclusións desta tese son:

1. A hipermutación somática nos xenes "junction" adxacentes a *IGKC* asóciase de forma intensa coa supervivencia global e o tempo ao primeiro tratamento. A combinación destas mutacións co estado mutacional de *IGHV* crea 4 subgrupos de doentes con taxas de progresión marcadamente diferentes.
2. 32 novos xenes recurrentemente mutados foron identificados no xenoma da LLC.
3. A presenza de mutacións infrecuentes e predictivamente patoxénicas en importantes xenes driver de cancro é un fenómeno frecuente no xenoma da LLC.
4. As mutacións en xenes da *"Vía efectora de TP53"* atópanse significativamente asociadas co tempo ao primeiro tratamento independentemente da presenza de mutacións en *TP53*.
5. 54 CNAs focais e recurrentes foron identificadas nos xenomas da LLC, con afectación frecuente de xenes relacionados coa mitose, a regulación do ciclo celular, a reparación e replicación do ADN e outras vías oncoxénicas.
6. A trisomía do cromosoma 12 asóciase con menor tempo ao primeiro tratamento e supervivencia global entre doentes con *IGHV* mutado.
7. Certas variantes xerminais en *MAP3K4* e *PPP4R2* atopáronse significativamente asociadas co tempo ao primeiro tratamento.
8. As mutacións non codificantes de proteína afectaron a expresión do oncoxén *RPL39L* e dos xenes supresores tumorais *PHF2* e *S1PR2*.
9. Un grupo de 290 tránscritos permite separar dous groupos de doentes con LLC que amosan unha marcada diferenza na probabilidade de precisar tratamento ao longo da evolución da patoloxía. Este patrón resultou independente do estado mutacional de *IGHV*.
10. Os clasificadores basados en aprendizaxe automatizado sobre datos de expresión xénica poden ser empregados para predecir con alta precisión qué doentes con LLC precisarán tratamento nos primeiros 5 anos dende o diagnóstico.
11. Un perfil de expresión xénica baseado en 3 tránscritos permite identificar un grupo próximo ao 5% de doentes con supervivencia global curta independientemente dos principais factores clínicos e moleculares de mal pronóstico.

1.5 Referencias

1. Zhao Y, Wang Y, Ma S. Racial differences in four leukemia subtypes: comprehensive descriptive epidemiology. Sci Rep. (2018) 8:548. doi: 10.1038/s41598-017-19081-4
2. Dores GM, Anderson WF, Curtis RE, Landgren O, Ostroumova E, Bluhm EC, et al. Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology. Br J Haematol. (2007) 139:809–19. doi: 10.1111/j.1365-2141.2007.0 6856.x
3. Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., et al. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. N. Engl. J. Med. 343, 1910–1916. doi: 10.1056/NEJM200012283432602
4. Pfeifer, D., Pantic, M., Skatulla, I., Rawluk, J., Kreutz, C., Martens, U. M., et al. (2007). Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. Blood 109, 1202–1210. doi: 10.1182/ blood-2006-07-034256
5. Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., et al. (2015). Mutations driving CLL and their evolution in progression and relapse. Nature 526, 525–530. doi: 10.1038/nature15395

6. Nadeu, F., Clot, G., Delgado, J., Martín-García, D., Baumann, T., Salaverria, I., et al. (2018). Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. Leukemia 32 (3), 645–653. doi: 10.1038/leu.2017.291
7. Raponi, S., Del Giudice, I., Marinelli, M., Wang, J., Cafforio, L., Ilari, C., et al. (2018). Genetic landscape of ultra-stable chronic lymphocytic leukemia patients. Ann. Oncol. 29 (4), 966–972. doi: 10.1093/annonc/mdy021
8. Gruber, M., Bozic, I., Leshchiner, I., Livitz, D., Stevenson, K., Rassenti, L., et al. (2019). Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature 570 (7762), 474–479. doi: 10.1038/s41586-019-1252-x
9. Edelmann, J., Tausch, E., Landau, D. A., Robrecht, S., Bahlo, J., Fischer, K., et al., (2017). Frequent evolution of copy number alterations in CLL following firstline treatment with FC(R) is enriched with TP53 alterations: results from the CLL8 trial. Leukemia 31, 734–738. doi: 10.1038/leu.2016.317
10. Yu, L., Kim, H. T., Kasar, S., Benien, P., Du, W., Hoang, K., et al. (2017). Survival of Del17p CLL depends on genomic complexity and somatic mutation. Clin. Cancer Res. 23, 735–745. doi: 10.1158/1078-0432.CCR-16-0594
11. Hernández-Sánchez, M., Rodríguez-Vicente, A. E., González-Gascón Y Marín, I., Quijada-Álamo, M., Hernández-Sánchez, J. M., Martín-Izquierdo, M., et al. (2019). DNA damage response-related alterations define the genetic background of patients with chronic lymphocytic leukemia and chromosomal gains. Exp. Hematol. 72, 9–13. doi: 10.1016/j.exphem.2019.02.003
12. Ljungström, V., Cortese, D., Young, E., Pandzic, T., Mansouri, L., Plevova, K., et al. (2016). Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. Blood 127 (8), 1007–1016. doi: 10.1182/blood-2015-10-674572
13. Leeksma, A. C., Taylor, J., Wu, B., Gardner, J. R., He, J., Nahas, M., et al. (2019). Clonal diversity predicts adverse outcome in chronic lymphocytic leukemia. Leukemia 33 (2), 390–402. doi: 10.1038/s41375-018-0215-9
14. Speedy HE, Di Bernardo MC, Sava GP, et al. A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. Nat Genet. 2014;46(1):56–60. https://doi.org/10.1038/ng.2843 Epub 2013 Dec 1.
15. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet. 2013;45(8):868–76. https://doi.org/10.1038/ng.2652 Epub 2013 Jun 16.
16. Berndt SI, Camp NJ, Skibola CF, et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Nat Commun. 2016;7:10933. https://doi.org/10.1038/ncomms10933.
17. Baecklund F, Foo JN, Bracci P, et al. A comprehensive evaluation of the role of genetic variation in follicular lymphoma survival. BMC Med Genet. 2014; 15:113. https://doi.org/10.1186/s12881-014-0113-6.
18. Ghesquieres H, Slager SL, Jardin F, et al. Genome-wide association study of event-free survival in diffuse large B-cell lymphoma treated with Immunochemotherapy. J Clin Oncol. 2015;33(33):3930–7. https://doi.org/10.1200/JCO.2014.60.2573 Epub 2015 Oct 12.

19. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 1160–1165, https://doi.org/10.1038/ng.3101 (2014).
20. Diederichs, S. et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non‐coding RNA and synonymous mutations. EMBO Mol. Med. 8, 442–457, https://doi.org/10.15252/emmm.201506055 (2016).
21. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. Nat. Rev. Genet. 11, 559–571, https://doi.org/10.1038/nrg2814 (2010).
22. Mansour, M. R. et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Sci. 346, 1373–1377,
23. Palamarchuk, A. et al. 13q14 deletions in CLL involve cooperating tumor suppressors. Blood 115, 3916–3922, https://doi.org/10.1182/blood-2009-10-249367 (2010).
24. Hornshøj, H. et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom. Med. 3, 1, https://doi.org/10.1038/s41525-017-0040-5 (2018).
25. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. bioRxiv 237313, https://doi.org/10.1101/237313.
26. Wadi, L. et al. Candidate cancer driver mutations in superenhancers and long-range chromatin interaction networks. bioRxiv 236802, https://doi.org/10.1101/236802
27. Liu, E. M. et al. Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell Syst. 8, 446–455, https://doi.org/10.1016/j.cels.2019.04.001 (2019).
28. Mozas P, Rivas-Delgado A, Baumann T, Villamor N, Ortiz-Maldonado V, Aymerich M, et al. Analysis of criteria for treatment initiation in patients with progressive chronic lymphocytic leukemia. Blood Cancer J. (2018) 8:10. doi: 10.1038/s41408-017-0044-5
29. Eichhorst B, Robak T, Montserrat E, Ghia P, Hillmen P, Hallek M, et al. Chronic lymphocytic leukaemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. (2015) 26 (Suppl. 5):v78–84. doi: 10.1093/annonc/mdv303
30. Burger JA, Tedeschi A, Barr PM, Robak T, Owen C, Ghia P, et al. Ibrutinib as Initial therapy for patients with chronic lymphocytic leukemia. N Engl J Med. (2015) 373:2425–37. doi: 10.1056/NEJMoa1509388
31. Brown JR, Byrd JC, Coutre SE, Benson DM, Flinn IW, Wagner-Johnston ND, et al. Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110δ, for relapsed/refractory chronic lymphocytic leukemia. Blood (2014) 123:3390–7. doi: 10.1182/blood-2013-11-535047
32. Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, et al. Roberts AW, Davids MS, Pagel JM, et al. Targeting BCL2 with venetoclax in relapsed chronic lymphocytic leukemia. N Engl J Med. (2016) 374:311–22. doi: 10.1056/NEJMoa1513257
33. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 526, 519–524 (2015).
34. Sandmann, S. et al. Evaluating variant calling tools for non-matched nextgeneration sequencing data. Sci. Rep. 7, 43169 (2017)

35. Ramsay, A. J. et al. Next-generation sequencing reveals the secrets of the chronic lymphocytic leukemia genome. Clin. Transl. Oncol. 15, 3–8 (2013).
36. Oricchio, E. et al. The Eph-receptor A7 is a soluble tumor suppressor for follicular lymphoma. Cell 147, 554–564 (2011)
37. Ge, Z. et al. Clinical significance of high c-MYC and low MYCBP2 expression and their association with Ikaros dysfunction in adult acute lymphoblastic leukemia. Oncotarget. 6, 42300–42311 (2015).
38. Sun, P. H., Ye, L., Mason, M. D. & Jiang, W. G. Protein tyrosine phosphatase μ (PTP μ or PTPRM), a negative regulator of proliferation and invasion of breast cancer cells, is associated with disease prognosis. PLoS One. 7, e50183 (2012)
39. Park, H. Y. et al. Whole-exome and transcriptome sequencing of refractory diffuse large B-cell lymphoma. Oncotarget. 7, 86433–86445 (2016).
40. Meriranta, L. et al. Deltex-1 mutations predict poor survival in diffuse large Bcell lymphoma. Haematologica 102, e195–e198 (2017).
41. Kan, Z. et al. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature. 466, 869–873 (2010)
42. Rodríguez, D. et al. Functional analysis of sucrase-isomaltase mutations from chronic lymphocytic leukemia patients. Hum. Mol. Genet. 22, 2273–2282 (2013).
43. Wazir, U. et al. The role of death-associated protein 3 in apoptosis, anoikis and human cancer. Cancer Cell Int. 15, 39, https://doi.org/10.1186/s12935-015-0187-z (2015).
44. Guérillon, C., Larrieu, D. & Pedeux, R. ING1 and ING2: multifaceted tumor suppressor genes. Cell Mol. Life Sci. 70, 3753–3772, https://doi.org/10.1007/s00018-013-1270-z (2013).
45. Khodabakhshi, A. H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget 3, 1308–1319 (2012).
46. Lee, K. H. et al. PHF2 histone demethylase acts as a tumor suppressor in association with p53 in cancer. Oncogene 34, 2897–2909, https://doi.org/10.1038/onc.2014.219 (2015).
47. Stelling, A. et al. The tumor suppressive TGF-β/SMAD1/S1PR2 signaling axis is recurrently inactivated in diffuse large B-cell lymphoma. Blood 131, 2235–2246, https://doi.org/10.1182/blood-2017-10-810630 (2018).
48. Dave, B. et al. Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. Proc Natl Acad Sci USA 111, 8838–8843, https://doi.org/10.1073/pnas.1320769111.
49. Batmanov, K., Wang, W., Bjørås, M., Delabie, J. & Wang, J. Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and BCL2 in follicular lymphoma. Sci. Rep. 7, 7040, https://doi.org/10.1038/s41598-017-07226-4 (2017).
50. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. Nat. Commun. 9, 4001, https://doi.org/10.1038/s41467-018-06354-3 (2018).
51. Mathelier, A. et al. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome Biol. 23(16), 84, https://doi.org/10.1186/s13059-015-0648-7 (2015).
52. Iwase, S., Shono, N., Honda, A., Nakanishi, T., Kashiwabara, S., Takahashi, S., et al. (2006). A component of BRAF–HDAC complex, BHC80, is required for neonatal survival in mice. FEBS Lett. 580 (13), 3129–3135. doi: 10.1016/j. Febslet.2006.04.065

53. Vizán, P., Beringer, M., Ballaré, C., and Di Croce, L. (2015). Role of PRC2- associated factors in stem cells and disease. FEBS J. 282 (9), 1723–1735. doi: 10.1111/febs.13083
54. Zhao, W., Tong, H., Huang, Y., Yan, Y., Teng, H., Xia, Y., et al. (2017). Essential role for Polycomb group protein Pcgf6 in embryonic stem cell maintenance and a noncanonical Polycomb repressive complex 1 (PRC1) integrity. J. Biol. Chem. Feb 17292 (7), 2773–2784. doi: 10.1074/jbc.M116.763961
55. Scuoppo, C., Miething, C., Lindqvist, L., Reyes, J., Ruse, C., Appelmann, I., et al. (2012). A tumour suppressor network relying on the polyamine-hypusine axis. Nature 487 (7406), 244–248. doi: 10.1038/nature11126
56. Saadi, H., Seillier, M., and Carrier, A. (2015). The stress protein TP53INP1 plays a tumor suppressive role by regulating metabolic homeostasis. Biochimie 118, 44–50. doi: 10.1016/j.biochi.2015.07.024
57. Zachari, M., and Ganley, I. G. (2017). The mammalian ULK1 complex and autophagy initiation. Essays Biochem. 61 (6), 585–596. doi: 10.1042/EBC2017 0021
58. Märklin, M., Heitmann, J. S., Fuchs, A. R., Truckenmüller, F. M., Gutknecht, M., Bugl, S., et al. (2017). NFAT2 is a critical regulator of the anergic phenotype in chronic lymphocytic leukaemia. Nat. Commun. 8 (1), 755. doi: 10.1038/ s41467-017-00830-y
59. Herzig JK, Bullinger L, Tasdogan A, et al. Protein phosphatase 4 regulatory subunit 2 (PPP4R2) is recurrently deleted in acute myeloid leukemia and required for efficient DNA double strand break repair. Oncotarget. 2017;8(56): 95038–53. https://doi.org/10.18632/oncotarget21119 eCollection 2017 Nov 10
60. Liu J, Xu L, Zhong J, et al. Protein phosphatase PP4 is involved in NHEJmediated repair of DNA double-strand breaks. Cell Cycle. 2012;11(14):2643–9. https://doi.org/10.4161/cc.20957 Epub 2012 Jul 15.
61. Chen MY, Chen YP, Wu MS, et al. PP4 is essential for germinal center formation and class switch recombination in mice. PLoS One. 2014;9(9): e107505. https://doi.org/10.1371/journal.pone.0107505 eCollection 2014.
62. Su YW, Chen YP, Chen MY, et al. The serine/threonine phosphatase PP4 is required for pro-B cell development through its promotion of immunoglobulin VDJ recombination. PLoS One. 2013;8(7):e68804. https:// doi.org/10.1371/journal.pone.0068804 Print 2013.
63. Krysko O, Aaes TL, Kagan VE, et al. Necroptotic cell death in anti-cancer therapy. Immunol Rev. 2017;280(1):207–19. https://doi.org/10.1111/imr.12583.
64. Lin TC, Su CY, Wu PY, et al. The nucleolar protein NIFK promotes cancer progression via CK1α/β-catenin in metastasis and Ki-67-dependent cell proliferation. Elife. 2016;17:5. https://doi.org/10.7554/eLife.11288.
65. Selvik LK, Rao S, Steigedal TS, et al. Salt-inducible kinase 1 (SIK1) is induced by gastrin and inhibits migration of gastric adenocarcinoma cells. PLoS One. 2014;9(11):e112485. https://doi.org/10.1371/journal.pone.0112485 eCollection 2014.
66. Hong B, Zhang J, Yang W. Activation of the LKB1-SIK1 signaling pathway inhibits the TGF-β-mediated epithelial-mesenchymal transition and apoptosis resistance of ovarian carcinoma cells. Mol Med Rep. 2018;17(2): 2837–44. https://doi.org/10.3892/mmr.2017.8229 Epub 2017 Dec 8.
67. Núñez-Enríquez JC, Bárcenas-López DA, Hidalgo-Miranda A, et al. Gene expression profiling of acute lymphoblastic leukemia in children with very early relapse. Arch Med Res. 2016;47(8):644–55. https://doi.org/10.1016/j.arcmed.2016.12.005.

68. Bertrand P, Bastard C, Maingonnat C, et al. Mapping of MYC breakpoints in 8q24 rearrangements involving non-immunoglobulin partners in B-cell lymphomas. Leukemia. 2007;21(3):515–23 Epub 2007 Jan 18.
69. Beyer T, Bolz S, Junger K, et al. CRISPR/Cas9-mediated genomic editing of Cluap1/IFT38 reveals a new role in actin arrangement. Mol Cell Proteomics. 2018;17(7):1285–94. https://doi.org/10.1074/mcp.RA117. 000487 Epub 2018 Apr 3.
70. Ishikura H, Ikeda H, Abe H, et al. Identification of CLUAP1 as a human osteosarcoma tumor-associated antigen recognized by the humoral immune system. Int J Oncol. 2007;30(2):461–7.
71. Takahashi M, Lin YM, Nakamura Y. Isolation and characterization of a novel gene CLUAP1 whose expression is frequently upregulated in colon cancer. Oncogene. 2004;23(57):9289–94.
72. Yan YB. Creatine kinase in cell cycle regulation and cancer. Amino Acids. 2016;48(8):1775–84. https://doi.org/10.1007/s00726-016-2217-0 Epub 2016 Mar

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2 INTRODUCTION

Chronic lymphocytic leukemia (CLL) is a low-grade B-cell lymphoproliferative disease with an estimated yearly incidence in western countries of about 6.9 cases per 100,000 people[1] and remarkable variation between races. The incidence of CLL is higher in men than in women and it increases progressively from the age of 35 until the last decades of life.[2] CLL is characterized by its clinical and genetic heterogeneity. *Döhner et al.* (2000) described the widely used cytogenetic classification of CLL based on the most prevalent chromosomal aberrations in the CLL genome, that is, trisomy 12 and deletions in 13q14.2–14.3, 11q22.3, and 17p13.1.[3] Since that moment, genome-wide studies have revealed new recurrent genomic aberrations, such as trisomy 19, amplifications of 2p and 8q, and deletions of 8p, 6q21, 18p, and 20p.[4,5] Similarly, a wealth of genomic and epigenomic modulators of CLL's clinical aggressivity have been discovered,[6] such as point mutations in *NOTCH1*, *SF3B1*, *ATM*, *TP53*, and *POT1* and the absence of somatic hypermutation at the variable region of the immunoglobulin heavy chain (*IGHV*) locus. It has been observed that copy number aberrations (CNAs) in CLL genomes tend to be acquired early in disease evolution and usually remain stable, whereas the mutational heterogeneity progressively increases.[6] Indeed, mounting evidence indicates that the accumulation of these cytogenomic events modulates CLL proliferation and clinical aggressivity to a great extent,[7,8] acting as drivers of genomic complexity and clonal evolution[9-1] and accumulating in relapsed cases.[12,13] Additionally, germline genomic variation can also induce predisposition to the development of CLL, a fact which has been the focus of various genome wide association studies (GWAS) during the last years. In this regard, dozens of common variants at genes such as *BCL2*, *EOMES*, *CASP10* and *POT1* have been associated with significant risk of CLL development.[14-16] Similarly, GWAS studies in other lymphoproliferative disorders such as follicular lymphoma and diffuse large B cell lymphoma have found evidence for the association of germline variants with overall survival and progression-free survival.[17,18] Despite this evidence, most analysis about CLL clinical evolution have been limited almost exclusively to acquired somatic events.

A major part of mutations in the cancer genome occur in non-coding DNA regions, and their function is still beginning to be understood.[19] Non-coding DNA comprises approximately 98% of the human genome, but recent research has proven that most of these regions are either part of regulatory motifs or actively transcribed to RNA.[20,21] These mutations can induce functional genomic changes by altering the binding of transcription factors or by inducing high-order chromatin structural modifications.[20,22] For example, mutations in 5′ and 3′ untranslated regions (UTRs) may disturb RNA structural conformation, modify microRNA binding sites or disrupt polyadenylation signals.[20] In a similar fashion, mutations affecting non-protein coding genes such as microRNA and long intergenic RNA genes (lincRNAs) are known cancer driver events.[20,23] Different studies have evidenced that the expression of genes such as *BRCA1*, *CDH10*, *CCND1*, *MALAT1*, *PAX5*, *RB1*, *SDHD*, *TERT*, *TOX3*, and *TAL1* is influenced by non-coding DNA mutations in regulatory regions of the cancer genome.[19,24,25] The *Pancancer Analysis of Whole Genomes* (PCAWG) project has revealed the existence of common and tumor-specific recurrently mutated functional elements near known cancer drivers.[25] Some of these driver mutations can induce long-range changes in genome organization and trigger abnormal expression of distant oncogenes and tumor suppressors.[26] Furthermore, the sequence

distribution of these driver mutations is not random. *Hornshøj et al.* (2018) identified a significant enrichment in conserved CCCT-binding factor (CTCF) binding sites among recurrently mutated non-coding DNA regions with cancer specificity.[24] Similarly, *Liu et al.* (2019) identified 21 recurrently altered CTCF-rich insulator regions in the cancer genome, and elegantly demonstrated that some of these mutations drive tumor proliferation.[27]

Currently, CLL treatment is delayed until disease progression (bone marrow failure, organomegaly, general symptoms, or high-grade lymphoma transformation) and in the case of refractory autoimmune phenomena.[28,29] Nevertheless, with the advent of new targeted treatments such as Bruton's tyrosine kinase inhibitor ibrutinib,[30] phosphoinositol-3-kinase inhibitor idelalisib,[31] and BCL2 inhibitor venetoclax,[32] it is tempting to speculate that some individuals could benefit from early intervention immediately following diagnosis, when the tumoral mass is smaller and patients have a better physical condition. Therefore, an improvement in the characterization of CLL biology and in the detection of new biomarkers will facilitate the creation of more personalized therapeutic approaches to this disease.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3 OBJECTIVES

This main objectives of this thesis are:

1. To detect new mutational drivers (i.e., recurrently mutated genes) and prognostic germline variants in the genomes of CLL, and to analyze their prognostic relevance.
    a. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, Bello López JL. Identification of new putative driver mutations and predictors of disease evolution in chronic lymphocytic leukemia. Blood Cancer J. 2019 Sep 30;9(10):78. doi: 10.1038/s41408-019-0243-3. PMID: 31570692; PMCID: PMC6769000.
2. To identify common germline variants modulating the clinical evolution of CLL patients treated with immunochemotherapy.
    a. Mosquera Orgueira A, Antelo Rodríguez B, Alonso Vence N, Díaz Arias JÁ, Díaz Varela N, Pérez Encinas MM, Allegue Toscano C, Goiricelaya Seco EM, Carracedo Álvarez Á, Bello López JL. The association of germline variants with chronic lymphocytic leukemia outcome suggests the implication of novel genes and pathways in clinical evolution. BMC Cancer. 2019 May 29;19(1):515. doi: 10.1186/s12885-019-5628-y. PMID: 31142279; PMCID: PMC6542042.
3. To improve the characterization of the noncoding genome of CLL, particularly in regions with known regulatory (i.e. epigenetic) activity.
    a. Mosquera Orgueira A, Rodríguez Antelo B, Díaz Arias JÁ, Díaz Varela N, Alonso Vence N, González Pérez MS, Bello López JL. Novel Mutation Hotspots within Non-Coding Regulatory Regions of the Chronic Lymphocytic Leukemia Genome. Sci Rep. 2020 Feb 12;10(1):2407. doi: 10.1038/s41598-020-59243-5. PMID: 32051441; PMCID: PMC7015923.
4. To identify novel recurrent structural aberrations in the genomes of CLL patients, their association with *IGHV* status and their prognostic importance.
    a. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, González Pérez MS, Bello López JL. New Recurrent Structural Aberrations in the Genome of Chronic Lymphocytic Leukemia Based on Exome-Sequencing Data. Front Genet. 2019 Sep 20;10:854. doi: 10.3389/fgene.2019.00854. PMID: 31616467; PMCID: PMC6764480.
5. To identify new transcriptomic signatures in CLL patients with prognostic implications.
    a. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, Díaz Varela N, Bello López JL. A Three-Gene Expression Signature Identifies a Cluster of Patients with Short Survival in Chronic Lymphocytic Leukemia. J Oncol. 2019 Nov 7;2019:9453539. doi: 10.1155/2019/9453539. PMID: 31827514; PMCID: PMC6885206.
    b. Mosquera Orgueira A, Antelo Rodríguez B, Alonso Vence N, Bendaña López Á, Díaz Arias JÁ, Díaz Varela N, González Pérez MS, Pérez Encinas MM, Bello López JL. Time to Treatment Prediction in Chronic Lymphocytic Leukemia Based on New Transcriptional Patterns. Front Oncol. 2019 Feb 15;9:79. doi: 10.3389/fonc.2019.00079. PMID: 30828568; PMCID: PMC6384245.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4 METHODOLOGY

4.1 Data Source

The *International Cancer Genome Consortium* (ICGC) Data Access Committee granted us access to the CLL sequencing data [33] deposited in the *European Genome-Phenome Database* (EGA) under DACO-1040945. These data included matched tumor and germline next-generation sequencing data from 506 patients diagnosed with CLL in Spain.

4.2 Exome-seq data processing and mutation detection

Samples were processed by *Puente et al.* (2015) as described in their original paper.[34] Briefly, 3 μg of genomic DNA was used for paired-end sequencing library construction, followed by enrichment in exomic sequences using the *SureSelect Human All Exon 50Mb* kit (Agilent Technologies). Next, DNA was pulled down using magnetic beads with streptavidin, followed by 18 cycles of amplification. Sequencing was performed on an Illumina GAIIx or on a HiSeq2000 sequencer (2 × 76 bp). Exome-seq data were aligned to the reference genome (GRCh37.75) using *bwa*.[35] Duplicate read removal, sorting, and indexing were done using *samtools*.[36] Base quality score recalibration was made with *BamUtil*[37] using a logistic regression model.

Mutation detection was performed with two different methods: *VarsCan2*,[38] which uses a heuristic/statistical method for variant detection; and *Platypus*,[39] which implements a Bayesian approach and local realignment of reads for indel and complex mutation detection. Variant quality was recalibrated using a logistic model, and drivers were detected by integrating the results of methods based on mutation frequency (*MuSiC2*),[40] functional impact (*OncodriveFM*),[41] co-localization (*OncodriveClust* and *Mutation3D*),[42,43] and pathogenicity prediction (*VEST* and *CHASM*) (Supplementary Methods).[44]

4.3 CNA and CNN-LOH Detection and Filtering

We analyzed paired tumor-normal exome-sequencing data with *Control-FREEC* version 11.3 in order to identify somatic CNA and copy number neutral loss of heterozygosity (CNN-LOH) regions.[45] Control-FREEC uses aligned reads to construct and normalize a copy number profile and a B-allele frequency (BAF) profile. Then, it performs profile segmentation and infers genotype status for each segment using both copy number and allelic frequency information. Finally, genomic aberrations are identified and annotated. CNA calling was limited to regions covered in the *Agilent SureSelect Exome Capture 50Mb* version 4 kit.

CNN-LOH were called with p-values < 0.05 (Kolmogorov–Smirnov test) and an uncertainty upper threshold of 5%. Regions with low mappability according to UCSC 75-bp mappability tracks (score below 0.5) were filtered out. As expected, we observed regions that seemed prone to CNA erroneous detection. Thus, we decided to apply a hard filter and discard those CNAs significantly enriched in both amplifications and deletions, as well as those located near a telomeric or centromeric region. CNA events were detected using *GISTIC2.0*,[46] which was run with default parameters plus the arm peel-off filter. Focal recurrent CNAs were defined as those spanning less than 50% of a chromosome arm with a residual q-value < 0.05 and a wide peak size below 10 megabases. A 1 − log2 tumor/normal ratio above 0.3 was used to define amplifications, and a ratio below −0.3 was used to define deletions.

4.4 RNAseq Analysis

Two CLL RNA-seq batches from the CLLE-ICGC cohort were uploaded in two stages with the following accession codes: EGAD00001001443 and EGAD00001000258. The first cohort (EGAD00001001443) contains RNAseq data and from CLL-purified cells of 196 individuals along with clinical data. The cohort was composed of 169 CLL, 22 monoclonal B cell lymphocytosis (MBL), and five small lymphocytic lymphoma (SLL) samples. There were 132 *IGHV* mutated cases and 64 *IGHV* unmutated cases in 119 males and 77 females. By staging at diagnosis, there were 22 MBL cases, 151 Binet Stage A cases, 14 Binet Stage B cases, and 8 Binet C stage cases. The second cohort (EGAD00001000258) is composed of RNAseq data of CLL-purified cells from 98 individuals, of which 79 (55 males and 24 females) have publicly available phenotypic information. In this cohort there were 72 CLL, 4 SLL, and 3 MBL samples. 45 of the patients had mutated *IGHV* and 34 had unmutated *IGHV*. By staging at diagnosis, there were 3 MBL, 72 Binet Stage A, 3 Binet Stage B and 1 Binet Stage C cases.

Illumina adapters were removed using *cutadapt*,[47] and alignment to the human reference genome (GRCh37) was performed using *Hisat2*[48] with default specifications. We used the Hisat2-provided Hierarchical Graph FM index for GRCh37 with SNP and Ensembl transcript information. Bam files were sorted and indexed using *samtools*.[36] Bam files were processed in R[49] according to the RNAseq gene expression protocol developed by *Love et al.* (2015).[50] Briefly, bam files were read using *Rsamtools*,[51] followed by gene-level expression estimation using the *SummarizeOverlaps* function from the *GenomicAlignments* package.[52] Gene models in GTF format were downloaded from Ensembl (GRCh37.75 version).[53] A log2-transformation on normalized frames per kilobase counts was performed.

4.5 CNA correlation with gene expression

Focal CNAs were classified according to Gistic into low-range events (tumor/normal log2 ratio > 0.3 and <0.9 for amplifications and less than −0.9 and more than −0.3 for deletions) and higher-range events (tumor/normal log2 ratio > 0.9 for amplifications and less than −0.9 for deletions). Correlation between CNA status and gene expression was performed using Spearman's correlation. A minimum of 5 CNA events with matched transcriptomic data was set for analysis. Furthermore, immunoglobulin and T-cell receptor gene rearrangements were not included. P-values were adjusted for multiple testing using the BH method.

4.6 Mutation and CNA Survival Analysis

Variables associated with time to treatment and overall survival were analyzed using Cox regression as implemented in the survival R package.[54] Assessment of the proportional hazards assumption was performed using Schoenfeld's method. When appropriate, models were adjusted for relevant covariates such as age, sex and *IGHV* mutation status.

4.7 Non-coding mutation detection and enrichment analysis

130 tumor-normal matched CLL whole genomes were processed using the *bcbio-nextgen* pipeline, which provides best practices for analyzing high throughput sequencing data.[55] Low complexity regions, areas with abnormally high coverage, sequences with single nucleotide stretches >50 bp and loci with alternative or unplaced contigs in the reference genome were not analyzed. Some polymorphic regions are prone to be classified as highly mutated due to artifacts or biases in the sequencing process, and suspicious elements were manually removed from

downstream analysis. Single nucleotide and indel mutation detection was performed with *vardict*,[56] *varscan*,[38] *mutect2*[57] and *freebayes*[58] using default bcbio-nextgen parameters. Only variants with a minimum sequencing depth (DP) of 10 and a genotype quality (GQ) above 20 Phred in both tumor and normal samples were analyzed. A mutation was reported when detected by at least two different mutation callers. Mutations were annotated to the 1000G,[59] gnomAD[60] and ExAc[61] databases in order to filter likely germline variants. All mutations with a minimum allele frequency >0.001 in any population were discarded from the analysis.

Annotations corresponding to promoter regions, 5′UTR, 3′UTR and lincRNAs were retrieved from *Genecode* version 18.[62] DNAse hypersensitivity (DHS) regions and Transcription Factor Binding Sites (TFBS) tracks from the ENCODE project were obtained from *Lochovsky et al.*[63,64] Similarly, we used enhancer regions from the *GeneHancer* database[65] and analyzed those that were supported by two or more sources of evidence ("elite" enhancers). Regulatory regions within telomeric and centromeric positions were discarded.

Two different methods were used to identify areas with evidence of positive selection of mutations: *LARVA*[63] and *OncodriveFML*.[66] LARVA models the mutation counts of each target region as a β-binomial distribution in order to handle overdispersion. Furthermore, LARVA also includes replication timing information in order to estimate local mutation rate, and provides a β-binomial distribution adjusted for replication timing which is used to compute p-values. On the other hand, OncodriveFML is designed to analyze the pattern of somatic mutations across tumors in both coding and non-coding genomic regions. OncodriveFML uses functional predictions in order to identify signals of positive selection. OncodriveFML was run with CADD v1.3 scores and default parameters. TFBS tracks were not analyzed with OncodriveFML due to high computational demands. Regions were labeled as significantly mutated if the q-value was <0.05 with any of the two methods.

In order to analyze the association of gene expression with recurrent non-coding DNA mutations we used microarray gene expression data available for this cohort. Briefly, background correction, normalization and log2-transformation of microarray gene expression data was performed with the RMA algorithm.[67] In the case of genes targeted by multiple probes, the median expression was calculated. The Wilcoxon-Rank sum test was used to detect changes in gene expression between mutated and wild-type cases. Non-coding regulatory genomic regions cannot be directly ascribed to any gene, and they can affect the transcription of virtually any part of the genome. However, this study is underpowered to detect long-range interactions due to small sample size and the need for extreme p-values passing multiple-testing correction. Therefore, we centered our efforts on changes in expression of the nearest gene. We annotated the closest gene to each recurrently mutated non-coding genomic region as the nearest transcription start site to the middle position of the corresponding region. In the case of multiple overlapping regulatory regions, we selected the most significant one for downstream analysis. P-values were adjusted for multiple testing using the FDR method, with a significance threshold of 0.05.

4.8 Germline polymorphism detection and correlation with patient survival

Base quality score recalibration was made with *BamUtil* [12] using a logistic regression model. Variant detection and filtering *Platypus2* [39] was run on genotyping mode. All dbSNP variants[68] were used as input for genotyping. Heterozygous loci with variant allele frequency (VAF) < 35% or > 70% were also discarded. We used principal component analysis (PCA) to detect outliers in our study cohort. Similarly, identity-by-descent (IBD) was used to discard all individuals with a

degree of relatedness equivalent to third degree or higher. PCAs and IBD data were computed on a linkage disequilibrium (LD) pruned dataset (LD upper threshold of 0.2) using the Bioconductor[69] package *SNPRelate*.[70] Our final filtered dataset contained 426 cases. Among these, 253 were males and 173 were females. By *IGHV* status, there were 146 unmutated cases and 273 mutated cases; and by clinical staging, the data contained 47 monoclonal B-cell lymphocytosis (MBL), 332 Binet A, 37 Binet B and 8 Binet C cases. Information about clinical staging and *IGHV* mutation status was not available for 2 and 7 cases, respectively.

Cox regression and assumption of proportional hazards was performed with the survival R package.[54] Variables with p-value < 0.2 in a univariate model were selected as covariates for the GWAS. In the cases of time to first treatment these were "donor sex", "*IGHV* mutation status" and "Binet stage"; whilst in the case of overall survival we used "*IGHV* mutation status", "Binet stage" and "donor age at diagnosis" as covariates. Three association models were computed: an additive model, a dominant model and a recessive model. P-values were adjusted using the Benjamini-Hochberg (BH) method.

Due to the heterogeneity of exome-seq coverage and quality metrics, many variables had incomplete data. We included in the analysis variables with at least 25% call rate, a minimum of 10 events (progression or death). A minimum allele frequency of 1% was selected as the lowest threshold. Furthermore, we only analyzed polymorphisms where *Platypus* called at least 10 minor alleles (additive model) or genotypes (dominant and recessive models).

Inflation values were estimated with the R package *QQperm*.[71] Briefly, a random distribution of p-values was created by randomly permuting phenotype variables. Then, the association p-values are compared with the null. This method doesn't consider the null distribution to be distributed uniformly.

VEGAS2[72] was used to calculate LD-adjusted gene-level association of p-values for time to treatment and overall survival. Briefly, VEGAS2 takes GWAS p-values, and then uses a simulation-based approach using information from population variant reference panels to adjust for LD effects. We used the 1000 Genomes phase 3 data from the iberian population in Spain as our reference population, since all patients of this cohort were of Spanish origin.[73] Only variants falling within the 5′ and 3′ coordinates of RefSeq genes were included. P-values were adjusted for multiple testing using the BH method.

GSEA4GWAS version 1.1[74] was used for testing significant associations in pathways and biological process annotations. Our input pathways were *"Canonical Pathways"* and *"Gene Ontology Biological Process"*. Maximum distance was set to 20 Kb, and the major histocompatibility complex region was masked from the analysis. P-values were adjusted with the BH method.

4.9 Machine learning-based prediction of Time to First Treatment Prediction based on gene expression profiles

We analyzed gene expression association with CLL's time to first treatment using cox regression.[54] In this model we included the covariates donor sex and CLL stage (MBL, Binet Stage A, Binet Stage B, and Binet Stage C). Time to Treatment was calculated as the period between CLL diagnosis and the initiation of the first treatment for CLL. The day of last follow-up was used for right censoring the data of patients with incomplete follow-up. Clustering was performed using the *Mclust* package[75] with default parameters. Briefly, Mclust infers the likeliest data clusters based on Gaussian Mixture Modeling fitted by an Expectation-Maximization (GMM-EM) algorithm.

Those genes with significant association with time to first treatment in the study cohort (cox regression BH q-value below 0.05) were selected as our initial list of genes. Variable selection was performed by adding one new gene in p-value ascending order to the model (starting with the first two most significant genes until reaching the top at 2,198 genes [FDR<5%]) and computing the most likely clusters. For the sake of simplicity, we discarded the 25% least variable genes, the 50% least expressed genes and those with a high (>0.9) Spearman's rank correlation with any other gene in the input data. In the case of a highly correlated pair of genes, the one with the lowest p-value was discarded. In each iteration we forced Mclust to calculate the two most likely groups of samples in our data, and to select the best model according to the maximal Bayesian Information Criterion (BIC). Association with time to first treatment was calculated using cox regression, including *IGHV* mutation status as covariate in each iteration. P-value adjustment was performed with the Bonferroni method.

For *IGHV* status and need of treatment at 5 years prediction we ran boosted trees analysis using BigML applications with a 2,000 tree node threshold.[76] We chose 5 years due to the following reasons: (1) it is important to differ which patients will have progression in the first years since diagnosis; and (2) the number of cases progressing in earlier years was too small in order to train a good classificator. Varying percentages of learning rates were tested. The best model was selected based on receiver operating characteristic (ROC) curves, Precision-Recall curves, and Kolmogorov-Smirnov statistics.

4.10 Identification of CLL patients at high risk of early death

We used two public databases of gene expression data in CLL patients in order to create a training and a validation cohort. The training cohort was composed of transcriptomic data from 450 CLL cases enrolled in the ICGC (data accessible in the European Genome-phenome Archive, accession code EGAD00010000875). Samples were collected and analyzed by the aforementioned consortium before initiation of any treatment. Overall survival was calculated as time from CLL diagnosis to time of death from any cause. Transcriptomic data were measured with Affymetrix HG-u219 microarrays. The Robust Multichip Algorithm (RMA)[67] was used to preprocess, normalize, and log2-transform expression data. For genes targeted by multiple probes, the median value was extracted. For each gene, we determined its individual clusterization capacity. The Mclust[75] algorithm was used in order to detect the 2 most likely patient clusters according to the expression of each gene (Mclust function, parameter G=2). The association of each of these single-gene clusters with overall survival was calculated using cox regression. Thereafter, those genes whose clusterization was significantly associated with survival (q-value <0.05) were selected for multivariate clusterization using the same Mclust algorithm.

An independent cohort of 107 CLL samples was used for validation (accessible in the Gene Expression Omnibus, accession code GSE22762, array platform Affymetrix Human Genome U133 Plus 2.0 Array). This dataset was composed of samples from patients with newly diagnosed and preexisting CLL, a fraction of whom had been previously treated. Overall survival was calculated as the period of time from microarray analysis to death from any cause. Briefly, normalized gene expression estimates were extracted and median expression for multiprobe genes was calculated. Then, cluster prediction was performed with parameters estimated in the training cohort, and cox regression was used to verify the association of this clusterization with survival.

5 DISCUSSION

CLL is a heterogeneous disease from both the biological and the clinical perspectives. Due to its high frequency in western countries and the novelty of drug innovations in the field, the understanding of its molecular pathogenesis along with the identification of novel biomarkers is key not just to improve patient prognostication, but also to optimize treatment strategies with new molecular profiles.

In this thesis, we have analyzed genomic data arising from CLL patients included in several public databases in order to identify new genomic drivers and prognostic factors in CLL. Our results demonstrate the existence of a series of new recurrent or infrequent somatic drivers of CLL, some of which bear prognostic information among patients treated with immunochemotherapy. Furthermore, we also observed the existence of germline variants associated with CLL clinical outcomes. Finally, we provide evidence indicating that molecular data from CLL tumors can be used to predict disease outcomes independently of state-of-the-art cytogenetic and molecular biomarkers. Therefore, this thesis covers different facets about cancer genomics: from basic driver discovery to biomarker identification and, finally, the development of new models of disease evolution. We believe that this capacity to translate the molecular heterogeneity of CLL into clinically-useful predictive information anticipates a new era of personalized medicine in the treatment of this lymphoid disorder.

During the discussion of our results, we initially center our focus on the identification of novel somatic events in the CLL genome, along with their correlations with state-of-the-art cytogenomic subgroups and patient survival. Afterwards, we describe the significant association of some germline polymorphisms with disease outcomes. Eventually, we present and discuss the usefulness of machine learning applications to genomic data for improving CLL risk stratification.

5.1 Novel somatic mutations in the CLL genome

The analysis of hundreds of CLL exomes has shed new light on the genomic determinants of this disease.[5,34] *Puente et al.*[34] analyzed a CLL cohort with 506 cases, identifying 36 recurrently mutated genes and 23 additional genes which were differentially mutated between *IGHV-hypermutation* status subgroups. On the other side, *Landau et al.*[5] analyzed a cohort of 538 cases and reported a list of 44 mutational drivers. By comparing the results, 28 genes were reported as drivers by both studies. Thus, the non-overlapping set of drivers (47 genes) could either be due to differences in patient characteristics, false findings, unevenly distributed rare mutations or methodological differences.

Mutation detection from next-generation sequencing data represents a big bottleneck in the development of this technology. Differences in clonality, sample purity, sequencing coverage and quality constitue difficulties for most variant callers, which are addressed in different fashions. Those methods with the highest sensitivity are frequently accompanied by lower precision, which leads to remarkable differences in mutation detection between different variant callers.[77] *Hoffman et al.*[78] compared 10 variant caller algorithms on simulated tumor-normal data, reporting considerable differences in sensitivity and precision depending on coverage and variant allele frequency (VAF). Notably, they observed that the best solution was obtained after combining mutations detected by different variant callers. Concordantly, *Cai et al.*[79] analyzed a

set of cancer samples with four different algorithms and observed that only 20.7% of variants were detected by 2 or more callers. Therefore, it is highly likely that numerous variants in large sequencing projects could have passed unnoticed, implicating the existence of numerous drivers yet to be known.

In this work, we performed a complementary analysis on the CLLE-ICGC cohort.[33] The final analysis included 49 monoclonal B cell lymphocytosis and 390 treatment-naive CLL samples. Mutation detection was performed with two different methods, and different algorithms used to detect driver genes and infrequent pathogenic mutations (see Methods section).

A total of 28,350 mutations were detected in 439 treatment-naive patient samples, of which 12,057 affected protein-coding regions. There were 8,965 non-silent and 3,095 silent mutations. The large majority of the non-silent mutations were missense (7,558 events). Point mutations were the most frequent (21,180), followed by short deletions (3,240) and insertions (2,041). There were 1,888 multi-nucleotide mutations (involving 2 or more consecutive nucleotides). Sixty-six genes were detected as putative drivers, of which thirty-two had been previously described by *Puente et al.*[34]. Among the novel ones, the most frequently mutated were *DTX1*, *LPHN3*, *LRP1B*, *LTB*, and *WDFY3*. *LPHN2* and *SI* were mutated in six patients; *BIRC6*, *DOCK1*, *MLL3*, *PCDH15*, *PTPN13*, *PTPRM*, *RELN* and *TFEB* were mutated in five patients and the remaining putative drivers were mutated in four different cases. Furthermore, *WDFY3* harbored two additional silent mutations that are predicted to create new donor or acceptor cryptic sites. *BIRC6*, *DOCK1*, *KMT2C*/*MLL3*, *PTPRB*, and *PTPRT* were each affected by one silent mutation predicted to create a new cryptic splice site. Mutations in *IGLL5* were frequent and located in hotspots, but they were accompanied by a high rate of silent mutations. Finally, we observed that *FREM1* was targeted by four likely functional non-synonymous mutations and two additional silent mutations in the same position.

Most of the new proposed drivers play well-defined roles in carcinogenesis, such as *EPHA7*,[80] *MYCBP2*[81] and *PTPRM*.[82] Other putative drivers have been linked to oncogenesis before, such as the autophagy regulator *WDFY3*,[83] the Notch pathway gene *DTX1*,[84] the latrophilin genes *LPHN2* and *LPHN3*,[85] as well as *FREM1*, which encodes the MYD88 and NFkB pathways related-protein TILRR.[86] Similarly, driver mutations in *CARD11* and *SI* have been previously described in CLL,[87] and the genes *BIRC6* and *KMT2C*/*MLL3* are paralogs of the CLL drivers *BIRC3* and *KMT2D*. Low-frequency and likely pathogenic mutations in 60 genes were detected. This type of mutations affected known cancer drivers (*EGFR*, *ERBB4*, *MAP2K1*, *NF1*, *NFKB1*, *NOTCH3*, and *SRSF1*), including multiple drivers of lymphoproliferation such as *BAX*, *BCOR*, *BCR*, *BTG2*, *DIS3*, *IKZF3*, *KRAS*, *PPM1D*, *PTPN11*, *SETD1B*, *TLR2*, and *TRAF3*. The list also includes regulators of lymphocyte pathways (*CD19*, CD36, *ALCAM*) and of relevant cancer pathways such as the Notch pathway (*NOTCH3*, *DMXL2*, and *SBNO1*), WNT/β-catenin pathway (*DACT1*); DNA polymerization (*POLE*) and epigenetic regulation (*KDM5A*, *HIST1H1D*, *PHF1* and single mutations at *HIST1H2BC* and *HIST1H2BG*). Moreover, isolated missense mutations in relevant oncogenes and tumor suppressor genes such as *EP300*, *KIT*, *MELK*, and *PTEN* were among the most significant events.

In conclusion, we describe the existence of multiple new putative driver mutations in the CLL genome that enhance our knowledge about the biology of this disease. Some of our results will need further clarification in the future. Particularly, the frequency, functional and clinical implications of the new putative drivers needs to be carefully analyzed in new studies. Altogether, we anticipate that our results will have direct implications in the biological comprehension and in the development of new therapies for CLL patients.

5.2 Non-coding mutations in the CLL genome identified from WGS analysis

Recent efforts by *Puente et al.*[34] enabled the discovery of 24 recurrently mutated non-coding genomic regions in the CLL genome, some of which are associated with functional changes such as mutations in the 3′UTR of *NOTCH1* and in the *PAX5* super-enhancer. Nevertheless, both the sparsity of annotations in non-coding DNA regions and the difficult functional classification of non-coding DNA mutations hinder a better understanding of the non-coding cancer genome, which probably harbors multiple deregulated elements yet to discover. In this section, we analyzed whole genome sequencing (WGS) data using a best-practice mutation detection pipeline. Then, we identified signals of positive selection of mutations in regulatory regions. Finally, our last attempt was to analyze if any of these recurrent mutations in non-coding DNA regions were associated with abnormal expression of the putatively regulated gene. Our results point toward the existence of dozens of mutation-enriched regulatory regions near cancer and immune-related genes, some of which influence local gene expression.

We identified dozens of recurrently mutated regulatory regions in the CLL genome. Among these, 10 were previously reported by the original analysis performed by *Puente et al.*,[34] namely those near *BACH2*, *BLC2*, *BCL6*, *BCL7A*, *BIRC3*, *S1PR2*, *PCDH15*, *ZCCHC7*/*PAX5* and *ZFP36L1*. Numerous novel regions were also enriched in non-coding DNA mutations, including transcription factor binding sites, DNAse hypersensitivity regions, 5′UTR regions, promoters, enhancers and non-coding RNAs. These events were frequently found in the vicinity of genes previously vinculated with oncogenic pathways. Indeed, the most significantly mutated regions were a *SETB1* binding site within the first intron of *DAP3*, a GTP-binding protein that participates in the apoptosis pathway;[88] and a DNAse hypersensitivity region downstream to *ING2*, a well-characterized tumor suppressor.[89] Other highly mutated regulatory regions affected cancer-related genes such as *DACT2*, *ERG*, *HIPK230*, *ITIH5*, *LRP5*, *MAF1*, *MALAT1*, *PHF2*, *PDGFA*, *RBFOX3*, *ROR2*, *ST6GAL1* and *XRCC5*;[90-102] and others were detected near genes involved in immunity, such as *LTB* and *MADCAM1*.[103,104] Overall, only three of the novel genes (*LTB*, *MALAT1* and *ST6GAL1*) were previously defined as targets of somatic hypermutation in B cell lymphomas.[105] Finally, it is worthwhile to mention that recurrent and even highly significant enrichments were detected around barely characterized genes (e.g. *C21ORF89*/*LINC0334*) and intergenic regions.

The reported mutations can either be bystander or have functional implications related to their potential to modify gene expression or to induce high-order chromatin structural changes. Although limited by low sample size, we devised significant changes in the expression of *PHF2*, *S1PR2* and *RPL39L*. These three genes are involved in the regulation of important oncogenic processes. *PHF2* encodes a histone demethylase with tumor suppressor activity.[97] *S1PR2* participates in the TGF-β pathway and acts as a tumor suppressor of B cell lymphomas.[106] Finally, *RPL39L*[107] is involved in cancer stem cell self-renewal and hypoxia response. These results are concordant with other reports of non-coding regulatory mutations driving gene expression changes in B-cell lymphomas.[108-110]

5.3 Importance of somatic mutations in CLL prognosis

Somatic hypermutation events occurred in a 1,543 base pair region located in the 5′ flank and UTR region of *IGKC* (172 patients, 40% of the total population). These mutations were strongly associated with longer time to first treatment (p-value 7.23×10^{-11}, HR 0.21–0.44) and were independent of *IGHV* status, sex, and clinical stage at diagnosis (p-value 6.3×10^{-3}, q-value 3.7

× 10^{-2}, HR 0.39–0.86). Similarly, an association with longer overall survival was detected (p-value 2.81 × 10^{-4}), but not independently of other covariates (p-value 0.54). This region includes protein-coding sequences of some immunoglobulin genes (namely *IGKJ1*, *IGKJ2*, *IGKJ3*, *IGKJ4*, and *IGJK5*). Although these genes were mutated in 126 cases, most of them (96%) had concurrent mutations in the surrounding non-coding region. The functional implications of these mutations is a matter of speculation, but we hypothesize their function as surrogate markers of B cell maturation.

Non-synonymous mutations in 16 genes were significantly associated with time to first treatment. The list included known CLL drivers such as *ATM*, *SF3B1*, *BRAF*, *NOTCH1*, *BIRC3*, *IRF4* and *ZMYM3*, as well as other putative novel drivers such as *EPHA7* and *SI*. Mutations in *IGLV3-21*, *DOCK1*, and *EPHA7* were associated with time to treatment after covariate adjustment (q-value < 0.1). In order to assess the potential effect of silent mutations on time to treatment, we included them in the regression, revealing new significant associations in *IGHV1-69*, *IGKJ5*, *IGHV2-70*, and *FAT1*. Furthermore, silent mutations in *IGLV3-21* became even more significant. Only two *IGLV3-21* mutated cases co-expressed *IGHV3-21*, indicating an independent role of the IGHV3-21/IGLV3-21 stereotyped B cell receptor. This is in concordance with a recent report about the adverse prognosis of *IGLV3-21* expression in CLL.[111] Finally, mutations in *ASXL1*, *ATM*, *IGHV1-69*, *SPEN*, *SF3F1*, *PLCH1* and *POT1* were associated with overall survival (q-value <10%), but none of these was significant after covariate adjustment (q-value < 0.1).

The genes *IGLL5*, *LTB*, *ZFP36L1*, *LRP1B*, and *PCDH15* were significantly enriched in intronic mutations (q-value < 0.1). Mutations in *ZFP36L1* and *DAPK1* were independently associated with time to first treatment (adjusted q-value < 0.1), whereas those in *IGHV3-49* were independently associated with overall survival (adjusted q-value < 0.1).

5.4 Pathway-oriented analysis of the mutational landscape of CLL and impact on disease prognosis

A pathway-level inquiry detected 62 terms enriched in non-synonymous mutations (Bonferroni p-value < 0.1). The most significantly mutated pathways were *"RB pathway"*, *"TP53 pathway"*, *"ATM pathway"*, *"Apoptotic Signaling in Response to DNA Damage"*, *"TP53 Hypoxia pathway"* and the *"G1 pathway"*. Most of the significant associations with clinical evolution were influenced by the presence of frequent driver mutations within the pathway.

Importantly, four pathways enriched in mutations did not include any high-frequency CLL-driver gene: *"CDK5 pathway"*, *"Apoptosis-induced DNA fragmentation"*, *"FRS2 mediated cascade"*, and the *"RAF MAP Kinase cascade"*. We detected an interesting pattern in the *"TP53 downstream pathway"*, which affected ~10% of the patients. Mutations in this pathway were strongly and independently associated with shorter time to first treatment (p-value 3.80 × 10^{-5}), and removing *TP53* mutated cases from the analysis did not affect the association substantially (p-value 5.3 × 10^{-4}). These mutations were also significantly associated with lower overall survival (p-value 2.81 × 10^{-4}), but not independently of *IGHV* status (p-value 0.54). These results suggest that the disruption of the *TP53* pathway plays an active role in CLL.

USC

5.5 Novel focal structural aberrations in the CLL genome

CNAs and CNN-LOH are oncogenic mechanisms that induce gene-dosage effects, disrupt coding sequences, cause structural rearrangements, or potentiate epigenetic effects. Oncogenes are frequently affected by copy number gains, while tumor suppressor genes tend to be deleted.

Massive array-based techniques such as array comparative genomic hybridization and single-nucleotide polymorphism (SNP) arrays have enabled the analysis of structural aberrations on cancer genomes to an unprecedented resolution of 10–100 kb. With the development of massive sequencing technologies, large databases of cancer sequence data have been published. This motivated the development of a variety of CNA detection algorithms from exome-sequencing data, which have the additional benefit of detecting smaller CNA events at the expense of increased false discoveries and reduced sensitivity and specificity.[112] These methods are specifically designed to face particular issues, particularly those inherent to the sequencing protocol (such as biases induced by hybridization, GC content, and read mappability), due to cancer biology (ploidy estimation and subclonality)[113] and due to the presence or absence of matched controls.[114] In this analysis, we used previously published exome-seq data in order to detect small recurrent structural events involved in the pathogenesis of CLL. Our results not only reproduce the known cytogenetic aberrations in CLL but also support the existence of multiple recurrent focal CNAs and CNN-LOH affecting key oncogenic pathways, some of which are clearly associated with higher proliferative capacity, shorter survival, and altered gene expression. We conclude that focal CNAs may be more relevant than previously expected in the pathogenesis of CLL, and they merit further consideration for prognostic stratification.

In particular, we report the existence of 54 putatively recurrent focal CNAs in the CLL genome. Focal recurrent gains and losses tended to target genes that participate in oncogenic pathways. For example, five amplified genes (*HFM1*, *ANAPC10*, *TAF5*, *COBRA1* and *SYCE3*) and one deleted gene (*FAM175A*/*ABRAXAS*) are involved in DNA transcription, replication, and repair mechanisms. Both *COBRA1* and *FAM175A* physically interact with the tumor suppressor *BRCA1*,[115,116] whereas *ANAPC10* belongs to the anaphase-promoting complex/cyclosome family of proteins that control sister chromatid segregation and cytokinesis.[117] The amplified genes *PHF21A*, *PCGF6* and *SUZ12* encode epigenetic regulators with repressor activity.[118-120] Likewise, the deleted genes *AMD1*, *TP53INP1*, *ULK1* and *NFATC1* are also important in tumorigenesis. *AMD1* and *TP53INP1* participate in metabolic pathways, and both have tumor suppressor activity,[121,122] whereas *ULK1* plays a decisive role in autophagy initiation[123] and *NFATC1* maintains an anergic phenotype in CLL cells.[124] Finally, two cyclin-dependent kinase genes were recurrently deleted: *CDK6* and *CDK19*. Furthermore, new CNN-LOH events affecting CLL drivers were also detected, including novel events affecting the loci of *ATM*, *NOTCH1*, *TP53*, *ARID1A*, *ASXL1*, *CREBBP* and *PI4KB*/*IL6R* loci, as well as the telomeric region of 11p. Only part of these events had concurrent mutations in their corresponding driver genes, suggesting the existence of other mechanisms of pathogenicity. Finally, we could detect significant positive correlations between six recurrent CNAs and the expression of genes encoded in their respective loci, as well as correlations between 19 CNAs and the expression 926 protein-coding genes genome wide.

Detection of copy number changes based on exome-sequencing has been proven to be prone to false positives by some studies.[125] In this analysis, we have included patient matched control samples, we applied stringent thresholds in order to minimize false detections, and we recapitulated most of the cytogenetic findings of CLL in the expected frequency. Furthermore, we could correlate the presence of this structural aberration with changes in gene expression in a subgroup of patients, although we believe that this study may be underpowered to detect such associations.

In conclusion, our study presents proof-of-concept evidence for the existence of new focal recurrent CNAs and CNN-LOH in the genome of CLL, some of which influence clinical outcome. Furthermore, we observed that some of these novel events have significant correlations with gene expression changes. The results are concordant with the possible involvement of a set of oncogenes and tumor suppressors in the development of CLL. These results should be considered a "proof of concept," and their existence and functionality should be validated in the future.

5.6 Prognostic importance of genomic structural aberrations in the CLL genome

Recurrent broad cytogenetic aberrations characteristic of CLL were identified at the expected frequency, and interestingly, we detected a significant adverse time to event and overall survival effect of trisomy 12 among *IGHV*-mutated cases. The importance of *IGHV* mutation as a predictor of disease evolution within the group of patients with trisomy 12 was previously reported by others,[126,127] but to our knowledge, this is the first report about the prognostic importance of trisomy 12 among *IGHV*-mutated cases. Likewise, deletions in *SETD2*, 11q22.3, and 14q32.33 were associated with shorter time to first treatment, and gains of *IRF4* were independently associated with short survival.

5.7 Germline variants associated with CLL clinical evolution

Inherited predisposition to the development of CLL has been proved by the results of various genome wide association studies (GWAS). In this regard, dozens of common variants at genes such as *BCL2*, *EOMES*, *CASP10* and *POT1* have been associated with significant risk of developing CLL.[14-16] Similarly, GWAS studies in other lymphoproliferative disorders such as follicular lymphoma and diffuse large B cell lymphoma have found evidence for the association of germline variants with overall survival and progression-free survival.[17,18] Despite this evidence, analysis of CLL clinical evolution has been limited almost exclusively to acquired somatic events. In this thesis, we addressed for the first time to our knowledge the association of common genomic variants with time to treatment time to first treatment and overall survival of CLL patients participating in the Spanish CLLE-ICGC cohort.

Our results suggest the existence of germline variation modulating CLL's clinical aggressivity. The most remarkable finding was the strong and recessive association of rs7620924 near *PPP4R2* with short time to treatment. The implication of *PPP4R2* in the regulation of cell survival and DNA repair in hematopoietic and leukemia cells has been recently reported.[128] Indeed, different studies have identified *PPP4R2* as a modulator of protein phosphatase 4 (PPP4), which regulates DNA repair through non-homologous end joining.[129] Concordantly, the ablation of PPP4 activity in mice increases genomic instability and aborgates class switch recombination in B cells, leading to an abnormal immune response;[130] and its function also seems to be essential in V(D)J recombination during normal B cell maturation.[131]

Other polymorphisms associated with time to first treatment were located in *MAP3K4*, *PEX26* and *TTLL12*. *MAP3K4* participates in the TRAIL/MAP3K4/p38/HSP27/Akt pathway, thereby modulating processes such as autophagy and cell migration. Indeed, *MAP3K4* is affected by recurrent loss-of-function mutations in different types of cancers.[132–135] Conversely, less is known about the peroxisome-related gene *PEX26*;[136] and about *TTLL12*, which participates in chromosome stability and mitosis-related processes.[137] On the contrary, although we did not find any variant significantly associated with overall survival, we devised some variants with

suggestive associations. The two most significant ones were in the loci of *TTC32/WDR35* and *CLIP1*, the last of which is overexpressed in Reed-Sternberg cells of Hodgkin lymphoma.[138, 139]

In a similar fashion, we detected that germline variation at the gene was also associated with CLL evolution. The most relevant was the association of *RIP3K* with both time to treatment and overall survival. *RIP3K* encodes a protein that regulates necroptosis, a form of regulated cell death characterized by cell membrane permeabilization.[140] Other relevant genes associated with rapid progression were the pro-proliferative gene *NIFK*,[141] the tumor suppressor *SIK1*[142,143] and *ZCCHC7*, which is encoded near the B-cell specific PAX5 super-enhancer locus.[144,145] On the other hand, various genes were associated with overall survival, such as *CLUAP1*, which participates in tumor growth and cytoskeleton regulation;[146–148] and the enzyme *GAMT*, which converts S-adenosylmethionine to creatine in order to foster high energy demands.[149] Moreover, the BCL-2 interacting gene *BNIPL*[150,151] was suggestively associated with survival and deserves further characterization. In a similar fashion, the most remarkable pathway-level association with survival was that of the pentose phosphate metabolic pathway, which fuels cells with metabolites for nucleotide and lipid biosynthesis, and provides reducing power to promote cell survival under stressful conditions.[152] Significant results were observed for the GNα13 and Nitric Oxide pathways. Concordantly, recurrent inactivating mutations in the G-protein superfamily gene *GNA13* have been described in B cell lymphomas,[151–156] and the contribution of nitric oxide to apoptosis resistance in CLL cells has been addressed by various studies.[157,158]

The main limitation of this study is the lack of an independent cohort for validation of these findings. Furthermore, although inflation values were low, we assume that treatment heterogeneity could have an impact on overall survival associations. Nevertheless, the global results are not only statistically significant but also biologically plausible. Thus, we believe that they will motivate further studies in order to confirm the effect of these variants and to determine their mechanisms of action in lymphoproliferative disorders.

5.8 Prediction of time to first treatment based on machine-learning analysis of gene expression profiles

Currently, CLL treatment is delayed until disease progression (bone marrow failure, organomegaly, general symptoms, or high-grade lymphoma transformation) and in the case of refractory autoimmune phenomena.[28,29] Recent advances in CLL genomics have discovered new drivers of disease, many of which are associated with a different clinical evolution. Several recurrent copy number aberrations and gene mutations in the CLL genome are linked to rapid disease progression.[34] Additionally, *IGHV* mutation status, which is an indirect measure of the tumor lymphocytes' maturation stage,[159] is among the most important single predictive factors known to date.[160] *IGHV* unmutated patients show remarkably worse prognosis than *IGHV* mutated patients[160,161] and only a few other genomic factors have proven to be associated with clinical evolution independent of this variable. Lymphocyte maturation is such an important indicator that DNA methylation status has been used to classify CLL into three different groups that resemble different B cell maturation stages (naive B cell, intermediate, and memory B cell). This classification was shown to outperform *IGHV* status at predicting time to first treatment.[162]

Mutations, genomic aberrations, and DNA methylation patterns induce transcriptomic changes that can be measured using RNA sequencing (RNAseq), a technique that offers an opportunity to identify new biomarkers for disease progression and drug response prediction.[163-165] In fact, previous efforts to improve CLL risk stratification based on RNAseq data have demonstrated impressive results,[166] but the clinical application is difficult due to the

expense of extensive technical and bioinformatics efforts. Therefore, there is a need for smaller transcriptomics patterns correlated with disease evolution for medical use. Defining reproducible gene expression patterns with clinical implications is a strategy that can close the gap between research and clinical practice. In this thesis, we define gene expression patterns that can improve CLL risk stratification with a relatively small set of the transcriptome. These results may pave the way for the design of new treatment strategies involving early CLL treatment in high-risk patients before disease progression.

RNAseq from tumor lymphocytes arising from patients included in the CLLE-ICGC cohort was used to make predictions about time to first treatment of CLL patients. We used the Gaussian Mixture Modelling - Expectation Maximization (GMM-EM) algorithm to stratify patients in two clusters with remarkably different clinical behavior based on the expression of 290 genes, and we observed that this pattern was independent of *IGHV* mutation status. Interestingly, we identified a group of CLL patients with mutated *IGHV* and a low-risk transcriptomic profile that only need treatment in approximately 25% of the cases during disease evolution. Two additional groups (one composed of patients with mutated *IGHV* and a high-risk transcriptomic profile and the second composed of unmutated *IGHV* patients with a low-risk transcriptomic profile) have similar intermediate evolution, while a final group (composed of patients with unmutated *IGHV* and an adverse transcriptomic profile) has the highest probability of treatment need in the first years following diagnosis. These results are concordant with previous reports in the field. For example, *Yepes et al.*[167] reported a division of CLL cases in two groups based on microarray transcriptome characterization through unsupervised clustering analysis, which was validated in 4 independent cohorts. Similarly, *Friedman et al.*[168] described a 180 probe classifier based on microarray data that also divided two clusters of CLL patients independently of *IGHV* mutation status. Our findings are also similar to those published by *Ferreira et al.*[166], who described two gene expression clusters that show *IGHV* mutation-independent association with time to first treatment using an early release of the ICGC CLL cohort. Nevertheless, there are remarkable differences between our analysis and that of *Ferreira et al.*, *Yepes et al.* and *Friedman et al*. Firstly, our clusterization is based on a transcriptional pattern of a small subgroup of genes that facilitates its future applicability, whilst those of *Ferreira et al.* and *Yepes et al.* are based on whole transcriptome analysis. Secondly, our classifier is based on RNAseq data, a technology that has outperformed microarray analysis in most fields. With the use of RNAseq it will be possible to couple transcriptome clusterization with targeted gene mutation detection, stereotyped B cell receptor expression or *IGHV* hypermutation status analysis.

In the same work, we also present a novel artificial intelligence algorithm that can predict the need for therapy during the first 5 years following diagnosis with high precision and accuracy. This is in line with other ML applications to oncologic malignancies that are starting to change paradigms in patient risk stratification and drug response prediction. For example, *Aziz et al.*[169] recently reported the identification of a ML model that integrates clinical and genomic data from patients with myelodysplastic syndrome (MDS). This model outperformed all commonly used prediction models in the field of MDS. Similarly, *Yousefi et al.*[170] used bayesian-optimized deep learning for survival prediction in pan-cancer analysis, showing not only better performance than other state-of-the-art methods, but also improved predictability of cancer survival through transfer learning in different types of cancer genomic data. Thus, it is likely that ML-driven algorithms applied to genomic and transcriptomic data will be used in the near future for the identification of "smoldering" CLLs that may benefit from early intervention.

5.9 Identification of high-risk CLL patient subgroups based on molecular patterns

The emergence of artificial intelligence has brought new expectations to the field of medicine, particularly for disease diagnosis and prognostication. Classical models such as cox proportional hazard model and the log-rank test assume that patient outcome consists of a linear combination of covariates, and do not provide decision rules for prediction in the real-world.[171] On the contrary, machine learning (ML) is a field of artificial intelligence that performs outcome prediction based on complex interactions between multiple variables, making little assumptions about the relationship between the dependent and independent variables.[172]

The inherent continuous nature of gene expression supposes an opportunity to dissect heterogeneous tumor types into comprehensive molecular subclasses. Indeed, previous efforts have proven the usefulness of this approach in CLL prognostication. *Rodríguez et al.* reported a seven-gene signature correlated with *IGHV* mutation status that predicts time to treatment,[173] whereas *Herold et al.* reported an 8-gene prognostic signature that predicted overall survival, but the predictability of this pattern was not superior to that of the combination of conventional FISH and *IGHV* mutation status.[174] Thus, we reasoned that the application of machine learning methods to gene expression could be used to improve the identification of CLL patients at high-risk of death independently of other cytogenetic and mutational variables associated with adverse outcome.

Three transcripts were able to individually clusterize patients in two groups with significantly different survival in the study cohort (Benjamini–Hochberg q-value <0.05). These genes were *SCGB2A1*, *KLF4*, and *PPP1R14B*. A multivariate clusterization based on the three genes was markedly associated with overall survival (cox regression p-value 4.31×10^{-6}, hazard ratio 4.86, lower 95%, confidence interval 2.48, upper 95% confidence interval 9.53). The cluster of patients with adverse survival supposed 4.22% of the study cohort. The prognostic impact of this clusterization on survival was validated in an independent cohort (cox regression p-value 5.7×10^{-6}, hazard ratio 10.79, lower 95% confidence interval 3.86, upper 95% confidence interval 30.17). The cluster of patients with adverse survival represented 5.60% of the validation cohort. In order to assess the independence of our clusterization approach, we used data from *Puente et al.*[34] to analyze for potential confounders in the study cohort. The following covariates were included in the model: patient's age at diagnosis, Binet stage at diagnosis, *IGHV* mutation status, presence of *TP53* mutation or 17p deletion, *ATM* mutation or 11q deletion, *NOTCH1* mutation, *SF3B1* mutation, and *BIRC3* mutation. The association of the transcriptomic clusterization with survival remained significant independently of the effect of these adverse prognostic factors (cox regression p-value 8.95×10^{-3}, hazard ratio 2.76).

The function of the genes included in the classifier was assessed in the literature. Interestingly, the Kruppel transcription factor *KLF4* has both growth suppressive and antiapoptotic functions. *KLF4* can both trigger cell-cycle arrest by inducing TP53-mediated expression of *CDKN1A* and also block apoptosis by inhibiting TP53 activity and suppressing BAX expression.[175] Less is known about *SCGB2A1* and *PPP1R14B*. *SCGB2A1* encodes a gene of the secretoglobin family. *SCGB2A1* is highly expressed in some tumor types,[176] and it has been linked to adverse cancer prognosis in others.[177] *PPP1R14B* encodes a putative inhibitor of protein phosphatase 1, a pleiotropic enzyme that plays multiple functions in cellular growth, cell cycle regulation, and apoptosis.[178]

Patients from both cohorts were diagnosed and treated in the era of chemoimmunotherapy. A limitation of this analysis is that we ignore which treatment regimens (if any) were

administered to each patient. Nevertheless, the remarkable strong association of the reported clusterization with overall survival in both the training and validation cohorts suggests a treatment-independent mechanism. It will be important to study the impact of new targeted drugs such as tyrosine kinase inhibitors or BCL2 antagonists in the survival of these CLL cases.

In conclusion, this report identified a 3-gene expression signature that characterizes a group of circa 5% of CLL patients with short survival. The prognostic impact of this signature was independent of the main cytogenomic markers of adverse prognosis at least in the study cohort. Such a small signature might be useful for future studies about disease prognostication and drug response in CLL.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6 CONCLUSIONS

The main conclusions of this thesis are:

1. Somatic hypermutation in the junction genes adjacent to *IGKC* is strongly associated with time to treatment and overall survival. The combination of *IGKC* mutations and hypermutated *IGHV* status creates four disease subgroups with significantly different progression rates.
2. Thirty-two novel genes were identified as mutational drivers of CLL.
3. Infrequent and predictively pathogenic mutations in known cancer drivers is a frequent event in CLL patients.
4. Mutations in genes belonging to the *"TP53 downstream pathway"* are significantly associated with time to disease treatment independently of *TP53* mutations.
5. Fifty-four focal CNA were recurrently detected in the CLL genome which frequently affected genes involved in mitosis, cell-cycle regulation, DNA repair and replication, as well as other oncogenic pathways. Additionally, new CNN-LOH events were detected affecting CLL drivers such as *ARID1A*, *ATM*, *CREBBP* and *NOTCH1*.
6. CDK6 deletions were associated with shorter time to first treatment, whereas losses of *SETD2* gains were associated with shorter overall survival.
7. Trisomy 12 was significantly associated with shorter time to treatment and overall survival among *IGHV* mutated cases.
8. Germline variants in *MAP3K4* and *PPP4R2* were significantly associated with time to first treatment.
9. Non-coding mutations in epigenetic regulator loci were associated with the expression of the putative oncogene *RPL39L* and the tumor suppressor genes *PHF2* and *S1PR2*.
10. A group of 290 transcripts could identify two distinct CLL subgroups with significantly different time to first treatment in an *IGHV*-independent fashion.
11. Machine learning classifiers based on gene expression data can be used to reliably predict which CLL patients will need treatment within the first 5 years since diagnosis.
12. A 3-gene expression pattern identifies a group of circa 5% of CLL cases with short overall survival independently of the main high-risk cytogenetic and mutational factors.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7 REFERENCES

1. Zhao Y, Wang Y, Ma S. Racial differences in four leukemia subtypes: comprehensive descriptive epidemiology. Sci Rep. (2018) 8:548. doi: 10.1038/s41598-017-19081-4
2. Dores GM, Anderson WF, Curtis RE, Landgren O, Ostroumova E, Bluhm EC, et al. Chronic lymphocytic leukaemia and small lymphocytic lymphoma: overview of the descriptive epidemiology. Br J Haematol. (2007) 139:809–19. doi: 10.1111/j.1365-2141.2007.0 6856.x
3. Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., et al. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. N. Engl. J. Med. 343, 1910–1916. doi: 10.1056/NEJM200012283432602
4. Pfeifer, D., Pantic, M., Skatulla, I., Rawluk, J., Kreutz, C., Martens, U. M., et al. (2007). Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. Blood 109, 1202–1210. doi: 10.1182/ blood-2006-07-034256
5. Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., et al. (2015). Mutations driving CLL and their evolution in progression and relapse. Nature 526, 525–530. doi: 10.1038/nature15395
6. Nadeu, F., Clot, G., Delgado, J., Martín-García, D., Baumann, T., Salaverria, I., et al. (2018). Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. Leukemia 32 (3), 645–653. doi: 10.1038/leu.2017.291
7. Raponi, S., Del Giudice, I., Marinelli, M., Wang, J., Cafforio, L., Ilari, C., et al. (2018). Genetic landscape of ultra-stable chronic lymphocytic leukemia patients. Ann. Oncol. 29 (4), 966–972. doi: 10.1093/annonc/mdy021
8. Gruber, M., Bozic, I., Leshchiner, I., Livitz, D., Stevenson, K., Rassenti, L., et al. (2019). Growth dynamics in naturally progressing chronic lymphocytic leukaemia. Nature 570 (7762), 474–479. doi: 10.1038/s41586-019-1252-x
9. Edelmann, J., Tausch, E., Landau, D. A., Robrecht, S., Bahlo, J., Fischer, K., et al., (2017). Frequent evolution of copy number alterations in CLL following firstline treatment with FC(R) is enriched with TP53 alterations: results from the CLL8 trial. Leukemia 31, 734–738. doi: 10.1038/leu.2016.317
10. Yu, L., Kim, H. T., Kasar, S., Benien, P., Du, W., Hoang, K., et al. (2017). Survival of Del17p CLL depends on genomic complexity and somatic mutation. Clin. Cancer Res. 23, 735–745. doi: 10.1158/1078-0432.CCR-16-0594
11. Hernández-Sánchez, M., Rodríguez-Vicente, A. E., González-Gascón Y Marín, I., Quijada-Álamo, M., Hernández-Sánchez, J. M., Martín-Izquierdo, M., et al. (2019). DNA damage response-related alterations define the genetic background of patients with chronic lymphocytic leukemia and chromosomal gains. Exp. Hematol. 72, 9–13. doi: 10.1016/j.exphem.2019.02.003
12. Ljungström, V., Cortese, D., Young, E., Pandzic, T., Mansouri, L., Plevova, K., et al. (2016). Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. Blood 127 (8), 1007–1016. doi: 10.1182/blood-2015-10-674572

13. Leeksma, A. C., Taylor, J., Wu, B., Gardner, J. R., He, J., Nahas, M., et al. (2019). Clonal diversity predicts adverse outcome in chronic lymphocytic leukemia. Leukemia 33 (2), 390–402. doi: 10.1038/s41375-018-0215-9
14. Speedy HE, Di Bernardo MC, Sava GP, et al. A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. Nat Genet. 2014;46(1):56–60. https://doi.org/10.1038/ng.2843 Epub 2013 Dec 1.
15. Berndt SI, Skibola CF, Joseph V, et al. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet. 2013;45(8):868–76. https://doi.org/10.1038/ng.2652 Epub 2013 Jun 16.
16. Berndt SI, Camp NJ, Skibola CF, et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. Nat Commun. 2016;7:10933. https://doi.org/10.1038/ncomms10933.
17. Baecklund F, Foo JN, Bracci P, et al. A comprehensive evaluation of the role of genetic variation in follicular lymphoma survival. BMC Med Genet. 2014; 15:113. https://doi.org/10.1186/s12881-014-0113-6.
18. Ghesquieres H, Slager SL, Jardin F, et al. Genome-wide association study of event-free survival in diffuse large B-cell lymphoma treated with Immunochemotherapy. J Clin Oncol. 2015;33(33):3930–7. https://doi.org/10.1200/JCO.2014.60.2573 Epub 2015 Oct 12.
19. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 1160–1165, https://doi.org/10.1038/ng.3101 (2014).
20. Diederichs, S. et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non‐coding RNA and synonymous mutations. EMBO Mol. Med. 8, 442–457, https://doi.org/10.15252/emmm.201506055 (2016).
21. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. Nat. Rev. Genet. 11, 559–571, https://doi.org/10.1038/nrg2814 (2010).
22. Mansour, M. R. et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Sci. 346, 1373–1377,
23. Palamarchuk, A. et al. 13q14 deletions in CLL involve cooperating tumor suppressors. Blood 115, 3916–3922, https://doi.org/10.1182/blood-2009-10-249367 (2010).
24. Hornshøj, H. et al. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. NPJ Genom. Med. 3, 1, https://doi.org/10.1038/s41525-017-0040-5 (2018).
25. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. bioRxiv 237313, https://doi.org/10.1101/237313.
26. Wadi, L. et al. Candidate cancer driver mutations in superenhancers and long-range chromatin interaction networks. bioRxiv 236802, https://doi.org/10.1101/236802
27. Liu, E. M. et al. Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell Syst. 8, 446–455, https://doi.org/10.1016/j.cels.2019.04.001 (2019).

28. Mozas P, Rivas-Delgado A, Baumann T, Villamor N, Ortiz-Maldonado V, Aymerich M, et al. Analysis of criteria for treatment initiation in patients with progressive chronic lymphocytic leukemia. Blood Cancer J. (2018) 8:10. doi: 10.1038/s41408-017-0044-5
29. Eichhorst B, Robak T, Montserrat E, Ghia P, Hillmen P, Hallek M, et al. Chronic lymphocytic leukaemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. (2015) 26 (Suppl. 5):v78–84. doi: 10.1093/annonc/mdv303
30. Burger JA, Tedeschi A, Barr PM, Robak T, Owen C, Ghia P, et al. Ibrutinib as Initial therapy for patients with chronic lymphocytic leukemia. N Engl J Med. (2015) 373:2425–37. doi: 10.1056/NEJMoa1509388
31. Brown JR, Byrd JC, Coutre SE, Benson DM, Flinn IW, Wagner-Johnston ND, et al. Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110δ, for relapsed/refractory chronic lymphocytic leukemia. Blood (2014) 123:3390–7. doi: 10.1182/blood-2013-11-535047
32. Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, et al. Roberts AW, Davids MS, Pagel JM, et al. Targeting BCL2 with venetoclax in relapsed chronic lymphocytic leukemia. N Engl J Med. (2016) 374:311–22. doi: 10.1056/NEJMoa1513257
33. Ramsay, A. J. et al. Next-generation sequencing reveals the secrets of the chronic lymphocytic leukemia genome. Clin. Transl. Oncol. 15, 3–8 (2013).
34. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature. 526, 519–524 (2015).
35. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
36. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. 1000 Genome Project Data Processing Subgroup, et al. (2009). 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16), 2078–2079.
37. Breese, M. R., and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics 29, 494– 496. doi: 10.1093/bioinformatics/bts731
38. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics. 2013 Feb 15;29(4):494-6. doi: 10.1093/bioinformatics/bts731. Epub 2013 Jan 12. PMID: 23314324; PMCID: PMC3570212.
39. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF; WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014 Aug;46(8):912-918. doi: 10.1038/ng.3036. Epub 2014 Jul 13. PMID: 25017105; PMCID: PMC4753679.
40. Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589-1598. doi:10.1101/gr.134635.111
41. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012 Nov;40(21):e169. doi: 10.1093/nar/gks743. Epub 2012 Aug 16. PMID: 22904074; PMCID: PMC3505979.

42. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013 Sep 15;29(18):2238-44. doi: 10.1093/bioinformatics/btt395. Epub 2013 Jul 24. PMID: 23884480.
43. Meyer MJ, Lapcevic R, Romero AE, Yoon M, Das J, Beltrán JF, Mort M, Stenson PD, Cooper DN, Paccanaro A, Yu H. mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. Hum Mutat. 2016 May;37(5):447-56. doi: 10.1002/humu.22963. Epub 2016 Feb 18. PMID: 26841357; PMCID: PMC4833594.
44. Masica DL, Douville C, Tokheim C, Bhattacharya R, Kim R, Moad K, Ryan MC, Karchin R. CRAVAT 4: Cancer-Related Analysis of Variants Toolkit. Cancer Res. 2017 Nov 1;77(21):e35-e38. doi: 10.1158/0008-5472.CAN-17-0338. PMID: 29092935; PMCID: PMC5850945.
45. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012 Feb 1;28(3):423-5. doi: 10.1093/bioinformatics/btr670. Epub 2011 Dec 6. PMID: 22155870; PMCID: PMC3268243.
46. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41. Epub 2011 Apr 28. PMID: 21527027; PMCID: PMC3218867.
47. Martin, M. (2011).Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17 (1), 10–12. doi: 10.14806/ej.17.1.200
48. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. (2015) 12:357–60. doi: 10.1038/nmeth.3317
49. R Development Core Team R: A Language and Environment for Statistical Computing. Vienna, Austria: the R Foundation for Statistical Computing. ISBN: 3-900051-07-0. Available online at: http://www.R-project.org/ (2011)
50. Love MI, Anders S, Kim V, Huber W. RNA-seq Workflow: Gene-level Exploratory Analysis and Differential Expression. Bioconductor. Available online at: https://www.bioconductor.org/help/workflows/rnaseqGene/ (2017)
51. Morgan M, Pagès H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, Variant Call (BCF), and Tabix File Import. R package version 1.30.0, Available online at: http://bioconductor.org/packages/release/ bioc/html/Rsamtools.html (2017)
52. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R. et al. Software for computing and annotating genomic ranges. PLoS ComputBiol. (2013) 9:e1003118. doi: 10.1371/journal.pcbi.1003118
53. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. (2016) 44(D1):D710–6. doi: 10.1093/nar/gkv1157
54. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York, NY: Springer (2000) doi: 10.1007/978-1-475 7-3294-8
55. Valls-Guimera, R. Bcbio-nextgen: Automated, distributed, next-gen sequencing pipeline. EMBnet J. 17, 30, https://doi.org/10.14806/ej.17.B.286 (2012).

56. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44, e108, https://doi.org/10.1093/nar/gkw227 (2016).
57. do Valle, I. F. et al. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinforma. 17, 341, https://doi.org/10.1186/s12859-016-1190-7 (2016).
58. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv, 1207.3907 [q-bio.GN] (2012)
59. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526 68–74, https://doi.org/10.1038/nature15393 (2015).
60. Karczewski, K.J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210, https://doi.org/10.1101/531210.
61. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nat. 536, 285–291, https://doi.org/10.1038/nature19057 (2016).
62. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774, https://doi.org/10.1101/gr.135350.111 (2012).
63. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46, 794–801, https://doi.org/10.1093/nar/gkx1081 (2018).
64. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. 43, 8123–8134, https://doi.org/10.1093/nar/gkv803 (2015).
65. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford), https://doi.org/10.1093/database/bax028 (2017).
66. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016 Jun 16;17(1):128. doi: 10.1186/s13059-016-0994-0. PMID: 27311963; PMCID: PMC4910259.
67. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2013).
68. Database of single nucleotide polymorphisms (dbSNP). Bethesda (MD): National Center for biotechnology information, National Library of medicine. (dbSNP build ID: 150). Available from: http://www.ncbi.nlm.nih.gov/SNP/
69. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115–21. https://doi.org/10.1038/nmeth.3252.
70. Zheng X, Levine D, Shen J, Gogarten S, Laurie C, Weir B. A high performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28(24):3326–8. https://doi.org/10. 1093/bioinformatics/bts606.
71. Slave Petrovski, Quanli Wang. QQperm: permutation based QQ plot and inflation factor estimation. 2016. R package ersion 1.0.1. https://cran.rproject.org/web/packages/QQperm/index.html.
72. Mishra A, Macgregor S. VEGAS2: software for more flexible gene-based testing. Twin Res Hum Genet. 2015;18(1):86–91. https://doi.org/10.1017/thg. 2014.79 Epub 2014 Dec 18

73. Castaño-Vinyals G, Aragonés N, Pérez-Gómez B, et al. Population-based multicase-control study in common tumors in Spain (MCC-Spain): rationale and study design. Gac Sanit. 2015;29(4):308–15. https://doi.org/10.1016/j. gaceta.2014.12.003 Epub 2015 Jan 19.
74. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res. 2010;38(Web Server issue):W90–5. https://doi.org/10.1093/ nar/gkq324 Epub 2010 Apr 30.
75. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J. (2016) 8:289–317
76. BigML is Machine Learning made easy, viewed 15 February 2018, Available online at: https://bigml.com
77. Sandmann, S. et al. Evaluating variant calling tools for non-matched nextgeneration sequencing data. Sci. Rep. 7, 43169 (2017)
78. Hofmann, A. L. et al. Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. BMC Bioinformatics 18, 8 (2017).
79. Cai, L., Yuan, W., Zhang, Z., He, L. & Chou, K. C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci. Rep. 6, 36540 (2016).
80. Oricchio, E. et al. The Eph-receptor A7 is a soluble tumor suppressor for follicular lymphoma. Cell 147, 554–564 (2011)
81. Ge, Z. et al. Clinical significance of high c-MYC and low MYCBP2 expression and their association with Ikaros dysfunction in adult acute lymphoblastic leukemia. Oncotarget. 6, 42300–42311 (2015).
82. Sun, P. H., Ye, L., Mason, M. D. & Jiang, W. G. Protein tyrosine phosphatase μ (PTP μ or PTPRM), a negative regulator of proliferation and invasion of breast cancer cells, is associated with disease prognosis. PLoS One. 7, e50183 (2012)
83. Park, H. Y. et al. Whole-exome and transcriptome sequencing of refractory diffuse large B-cell lymphoma. Oncotarget. 7, 86433–86445 (2016).
84. Meriranta, L. et al. Deltex-1 mutations predict poor survival in diffuse large Bcell lymphoma. Haematologica 102, e195–e198 (2017).
85. Kan, Z. et al. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature. 466, 869–873 (2010)
86. Zhang, X., Pino, G. M., Shephard, F., Kiss-Toth, E. & Qwarnstrom, E. E. Distinct control of MyD88 adapter-dependent and Akt kinase-regulated responses by the interleukin (IL)-1RI co-receptor, TILRR. J. Biol. Chem. 287, 12348–12352 (2012)
87. Rodríguez, D. et al. Functional analysis of sucrase-isomaltase mutations from chronic lymphocytic leukemia patients. Hum. Mol. Genet. 22, 2273–2282 (2013).
88. Wazir, U. et al. The role of death-associated protein 3 in apoptosis, anoikis and human cancer. Cancer Cell Int. 15, 39, https://doi.org/10.1186/s12935-015-0187-z (2015).
89. Guérillon, C., Larrieu, D. & Pedeux, R. ING1 and ING2: multifaceted tumor suppressor genes. Cell Mol. Life Sci. 70, 3753–3772, https://doi.org/10.1007/s00018-013-1270-z (2013).
90. Li, J. et al. Methylation of DACT2 promotes breast cancer development by activating Wnt signaling. Sci. Rep. 7, 3325, https://doi.org/10.1038/s41598-017-03647-3 (2017).

91. Adamo, P. & Ladomery, M. R. The oncogene ERG: a key factor in prostate cancer. Oncogene 35, 403–414, https://doi.org/10.1038/onc.2015.109 (2016).
92. D'Orazi, G., Rinaldo, C. & Soddu, S. Updates on HIPK2: a resourceful oncosuppressor for clearing cancer. J. Exp. Clin. Cancer Res. 31, 63, https://doi.org/10.1186/1756-9966-31-63 (2012).
93. Rose, M. et al. ITIH5 induces a shift in TGF-β superfamily signaling involving Endoglin and reduces risk for breast cancer metastasis and tumor death. Mol. Carcinog. 57, 167–181, https://doi.org/10.1002/mc.22742 (2018).
94. Ren, D. N. et al. LRP5/6 directly bind to Frizzled and prevent Frizzled-regulated tumour metastasis. Nat. Commun. 6, 6906, https://doi.org/10.1038/ncomms7906 (2015).
95. Li, Y. et al. MAF1 suppresses AKT-mTOR signaling and liver cancer through activation of PTEN transcription. Hepatology 63, 1928–1942, https://doi.org/10.1002/hep.28507 (2016).
96. Liu, J., Peng, W. X., Mo, Y. Y. & Luo, D. MALAT1-mediated tumorigenesis. Front. Biosci. 22, 66–80 (2017).
97. Lee, K. H. et al. PHF2 histone demethylase acts as a tumor suppressor in association with p53 in cancer. Oncogene 34, 2897–2909, https://doi.org/10.1038/onc.2014.219 (2015).
98. Palomero, J. et al. SOX11 promotes tumor angiogenesis through transcriptional regulation of PDGFA in mantle cell lymphoma. Blood 124, 2235–2247, https://doi.org/10.1182/blood-2014-04-569566 (2014).
99. Liu, T. et al. RBFOX3 Promotes Tumor Growth and Progression via hTERT Signaling and Predicts a Poor Prognosis in Hepatocellular Carcinoma. Theranostics 7, 3138–3154, https://doi.org/10.7150/thno.19506 (2017).
100. Debebe, Z. & Rathmell, W. K. Ror2 as a therapeutic target in cancer. Pharmacol. Ther. 150, 143–148, https://doi.org/10.1016/j.pharmthera.2015.01.010 (2015).
101. Antony, P. et al. Epigenetic inactivation of ST6GAL1 in human bladder cancer. BMC Cancer 14, 901, https://doi.org/10.1186/1471-2407-14-901 (2014).
102. Zhang, Z. et al. XRCC5 cooperates with p300 to promote cyclooxygenase-2 expression and tumor growth in colon cancers. PLoS One 12, e0186900, https://doi.org/10.1371/journal.pone.0186900 (2017).
103. Nagy, B. et al. Lymphotoxin beta expression is high in chronic lymphocytic leukemia but low in small lymphocytic lymphoma: a quantitative real-time reverse transcriptase polymerase chain reaction analysis. Haematologica 88, 654–658 (2003).
104. Sakai, Y. & Kobayashi, M. Lymphocyte 'homing' and chronic inflammation. Pathol. Int. 65, 344–354, https://doi.org/10.1111/pin.12294 (2015).
105. Khodabakhshi, A. H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget 3, 1308–1319 (2012).
106. Stelling, A. et al. The tumor suppressive TGF-β/SMAD1/S1PR2 signaling axis is recurrently inactivated in diffuse large B-cell lymphoma. Blood 131, 2235–2246, https://doi.org/10.1182/blood-2017-10-810630 (2018).
107. Dave, B. et al. Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. Proc Natl Acad Sci USA 111, 8838–8843, https://doi.org/10.1073/pnas.1320769111.
108. Batmanov, K., Wang, W., Bjørås, M., Delabie, J. & Wang, J. Integrative whole-genome sequence analysis reveals roles of regulatory mutations in BCL6 and

BCL2 in follicular lymphoma. Sci. Rep. 7, 7040, https://doi.org/10.1038/s41598-017-07226-4 (2017).
109. Arthur, S. E. et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. Nat. Commun. 9, 4001, https://doi.org/10.1038/s41467-018-06354-3 (2018).
110. Mathelier, A. et al. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. Genome Biol. 23(16), 84, https://doi.org/10.1186/s13059-015-0648-7 (2015).
111. Stamatopoulos, B. et al. The light chain IgLV3-21 defines a new poor prognostic subgroup in chronic lymphocytic leukemia: results of a multicenter study. Clin. Cancer Res. https://doi.org/10.1158/1078-0432. CCR-18-0133. (2018)
112. Nam, J. Y., Kim, N. K., Kim, S. C., Joung, J. G., Xi, R., Lee, S., et al. (2016). Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics 17, 185–192. doi: 10.1093/bib/bbv055
113. Zare, F., Dow, M., Monteleone, N., Hosny, A., and Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics 18 (1), 286. doi: 10.1186/ s12859-017-1705-x
114. Kim, H. Y., Choi, J. W., Lee, J. Y., and Kong, G. (2017). Gene-based comparative analysis of tools for estimating copy number alterations using wholeexome sequencing data. Oncotarget 8 (16), 27277–27285. doi: 10.18632/ oncotarget.15932
115. Castillo, A., Paul, A., Sun, B., Huang, T. H., Wang, Y., Yazinski, S. A., et al. (2014). The BRCA1-interacting protein Abraxas is required for genomic stability and tumor suppression. Cell Rep. 8 (3), 807–817. doi: 10.1016/j.celrep.2014.06.050
116. Yun, H., Bedolla, R., Horning, A., Li, R., Chiang, H. C., Huang, T. H., et al. (2018). BRCA1 interacting protein COBRA1 facilitates adaptation to castrateresistant growth conditions. Int. J. Mol. Sci. 19 (7), Pii: E2104. doi: 10.3390/ ijms19072104
117. Chang, L. F., Zhang, Z., Yang, J., McLaughlin, S. H., and Barford, D. (2014). Molecular architecture and mechanism of the anaphase-promoting complex. Nature 513 (7518), 388–393. doi: 10.1038/nature13543
118. Iwase, S., Shono, N., Honda, A., Nakanishi, T., Kashiwabara, S., Takahashi, S., et al. (2006). A component of BRAF–HDAC complex, BHC80, is required for neonatal survival in mice. FEBS Lett. 580 (13), 3129–3135. doi: 10.1016/j. Febslet.2006.04.065
119. Vizán, P., Beringer, M., Ballaré, C., and Di Croce, L. (2015). Role of PRC2-associated factors in stem cells and disease. FEBS J. 282 (9), 1723–1735. doi: 10.1111/febs.13083
120. Zhao, W., Tong, H., Huang, Y., Yan, Y., Teng, H., Xia, Y., et al. (2017). Essential role for Polycomb group protein Pcgf6 in embryonic stem cell maintenance and a noncanonical Polycomb repressive complex 1 (PRC1) integrity. J. Biol. Chem. Feb 17292 (7), 2773–2784. doi: 10.1074/jbc.M116.763961
121. Scuoppo, C., Miething, C., Lindqvist, L., Reyes, J., Ruse, C., Appelmann, I., et al. (2012). A tumour suppressor network relying on the polyamine-hypusine axis. Nature 487 (7406), 244–248. doi: 10.1038/nature11126
122. Saadi, H., Seillier, M., and Carrier, A. (2015). The stress protein TP53INP1 plays a tumor suppressive role by regulating metabolic homeostasis. Biochimie 118, 44–50. doi: 10.1016/j.biochi.2015.07.024

123. Zachari, M., and Ganley, I. G. (2017). The mammalian ULK1 complex and autophagy initiation. Essays Biochem. 61 (6), 585–596. doi: 10.1042/EBC2017 0021
124. Märklin, M., Heitmann, J. S., Fuchs, A. R., Truckenmüller, F. M., Gutknecht, M., Bugl, S., et al. (2017). NFAT2 is a critical regulator of the anergic phenotype in chronic lymphocytic leukaemia. Nat. Commun. 8 (1), 755. doi: 10.1038/ s41467-017-00830-y
125. Rieber, N., Bohnert, R., Ziehm, U., and Jansen, G. (2017). Reliability of algorithmic somatic copy number alteration detection from targeted capture data. Bioinformatics 33 (18), 2791–2798. doi: 10.1093/bioinformatics/btx284
126. Bulian, P., Bomben, R., Bo, M. D., Zucchetto, A., Rossi, F. M., Degan, M., et al. (2017). Mutational status of IGHV is the most reliable prognostic marker in trisomy 12 chronic lymphocytic leukemia. Haematologica 102 (11), e443–e446. doi: 10.3324/haematol.2017.170340
127. Roos-Weil, D., Nguyen-Khac, F., Chevret, S., Touzeau, C., Roux, C., Lejeune, J., et al. (2018). Mutational and cytogenetic analyses of 188 CLL patients with trisomy 12: a retrospective study from the French Innovative Leukemia Organization (FILO) working group. Genes Chromosomes Cancer 57 (11), 533–540. doi: 10.1002/gcc.22650
128. Herzig JK, Bullinger L, Tasdogan A, et al. Protein phosphatase 4 regulatory subunit 2 (PPP4R2) is recurrently deleted in acute myeloid leukemia and required for efficient DNA double strand break repair. Oncotarget. 2017;8(56): 95038–53. https://doi.org/10.18632/oncotarget21119 eCollection 2017 Nov 10
129. Liu J, Xu L, Zhong J, et al. Protein phosphatase PP4 is involved in NHEJmediated repair of DNA double-strand breaks. Cell Cycle. 2012;11(14):2643–9. https://doi.org/10.4161/cc.20957 Epub 2012 Jul 15.
130. Chen MY, Chen YP, Wu MS, et al. PP4 is essential for germinal center formation and class switch recombination in mice. PLoS One. 2014;9(9): e107505. https://doi.org/10.1371/journal.pone.0107505 eCollection 2014.
131. Su YW, Chen YP, Chen MY, et al. The serine/threonine phosphatase PP4 is required for pro-B cell development through its promotion of immunoglobulin VDJ recombination. PLoS One. 2013;8(7):e68804. https:// doi.org/10.1371/journal.pone.0068804 Print 2013.
132. Kanchi KL, Johnson KJ, Lu C, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun. 2014;5:3156. https://doi. org/10.1038/ncomms4156.
133. Yang LX, Gao Q, Shi JY, et al. Mitogen-activated protein kinase kinase kinase 4 deficiency in intrahepatic cholangiocarcinoma leads to invasive growth and epithelial-mesenchymal transition. Hepatology. 2015;62(6):1804–16. https://doi.org/10.1002/hep.28149 Epub 2015 Oct 17.
134. Kim J, Kang D, Sun BK, et al. TRAIL/MEKK4/p38/HSP27/Akt survival network is biphasically modulated by the Src/CIN85/c-Cbl complex. Cell Signal. 2013; 25(1):372–9. https://doi.org/10.1016/j.cellsig.2012.10.010 Epub 2012 Oct 23.
135. Sollome JJ, Thavathiru E, Camenisch TD, Vaillancourt RR. HER2/HER3 regulates extracellular acidification and cell migration through MTK1 (MEKK4). Cell Signal. 2014;26(1):70–82. https://doi.org/10.1016/j.cellsig.2013. 08.043 Epub 2013 Sep 12.
136. Keil E, Höcker R, Schuster M, et al. Phosphorylation of Atg5 by the Gadd45βMEKK4-p38 pathway inhibits autophagy. Cell Death Differ. 2013;20(2):321–https://doi.org/10.1038/cdd.2012.129 Epub 2012 Oct 12. 33. Weller S,

Cajigas I, Morrell J, et al. Alternative splicing suggests extended function of PEX26 in peroxisome biogenesis. Am J Hum Genet. 2005;76(6): 987–1007 Epub 2005 Apr 27.

137. Brants J, Semenchenko K, Wasylyk C, et al. Tubulin tyrosine ligase like 12, a TTLL family member with SET- and TTL-like domains and roles in histone and tubulin modifications and mitosis. PLoS One. 2012;7(12):e51258. https://doi.org/10.1371/journal.pone.0051258. Epub 2012 Dec 12.

138. Perez F, Diamantopoulos GS, Stalder R, Kreis TE. CLIP-170 highlights growing microtubule ends in vivo. Cell. 1999;96(4):517–27.

139. Sahin U, Neumann F, Tureci O, et al. Hodgkin and reed-Sternberg cellassociated autoantigen CLIP-170/restin is a marker for dendritic cells and is involved in the trafficking of macropinosomes to the cytoskeleton, supporting a function-based concept of Hodgkin and reed-Sternberg cells. Blood. 2002;100(12):4139–45.

140. Krysko O, Aaes TL, Kagan VE, et al. Necroptotic cell death in anti-cancer therapy. Immunol Rev. 2017;280(1):207–19. https://doi.org/10.1111/imr.12583.

141. Lin TC, Su CY, Wu PY, et al. The nucleolar protein NIFK promotes cancer progression via CK1α/β-catenin in metastasis and Ki-67-dependent cell proliferation. Elife. 2016;17:5. https://doi.org/10.7554/eLife.11288.

142. Selvik LK, Rao S, Steigedal TS, et al. Salt-inducible kinase 1 (SIK1) is induced by gastrin and inhibits migration of gastric adenocarcinoma cells. PLoS One. 2014;9(11):e112485. https://doi.org/10.1371/journal.pone.0112485 eCollection 2014.

143. Hong B, Zhang J, Yang W. Activation of the LKB1-SIK1 signaling pathway inhibits the TGF-β-mediated epithelial-mesenchymal transition and apoptosis resistance of ovarian carcinoma cells. Mol Med Rep. 2018;17(2): 2837–44. https://doi.org/10.3892/mmr.2017.8229 Epub 2017 Dec 8.

144. Núñez-Enríquez JC, Bárcenas-López DA, Hidalgo-Miranda A, et al. Gene expression profiling of acute lymphoblastic leukemia in children with very early relapse. Arch Med Res. 2016;47(8):644–55. https://doi.org/10.1016/j.arcmed.2016.12.005.

145. Bertrand P, Bastard C, Maingonnat C, et al. Mapping of MYC breakpoints in 8q24 rearrangements involving non-immunoglobulin partners in B-cell lymphomas. Leukemia. 2007;21(3):515–23 Epub 2007 Jan 18.

146. Beyer T, Bolz S, Junger K, et al. CRISPR/Cas9-mediated genomic editing of Cluap1/IFT38 reveals a new role in actin arrangement. Mol Cell Proteomics. 2018;17(7):1285–94. https://doi.org/10.1074/mcp.RA117. 000487 Epub 2018 Apr 3.

147. Ishikura H, Ikeda H, Abe H, et al. Identification of CLUAP1 as a human osteosarcoma tumor-associated antigen recognized by the humoral immune system. Int J Oncol. 2007;30(2):461–7.

148. Takahashi M, Lin YM, Nakamura Y. Isolation and characterization of a novel gene CLUAP1 whose expression is frequently upregulated in colon cancer. Oncogene. 2004;23(57):9289–94.

149. Yan YB. Creatine kinase in cell cycle regulation and cancer. Amino Acids. 2016;48(8):1775–84. https://doi.org/10.1007/s00726-016-2217-0 Epub 2016 Mar 28.

150. Qin W, Hu J, Guo M, et al. BNIPL-2, a novel homologue of BNIP-2, interacts with Bcl-2 and Cdc42GAP in apoptosis. Biochem Biophys Res Commun. 2003;308(2):379–85.

151. Xie L, Qin W, Li J, et al. BNIPL-2 promotes the invasion and metastasis of human hepatocellular carcinoma cells. Oncol Rep. 2007;17(3):605–10.

152. Patra KC, Hay N. The pentose phosphate pathway and cancer. Trends Biochem Sci. 2014;39(8):347–54. https://doi.org/10.1016/j.tibs.2014.06.005 Epub 2014 Jul 15.

153. Love C, Sun Z, Jima D, et al. The genetic landscape of mutations in Burkitt lymphoma. Nat Genet. 2012;44(12):1321–5. https://doi.org/10.1038/ng.2468 Epub 2012 Nov 11.

154. Muppidi JR, Schmitz R, Green JA, et al. Loss of signalling via Gα13 in germinal Centre B-cell-derived lymphoma. Nature. 2014;516(7530):254–8. https://doi.org/10.1038/nature13765 Epub 2014 Sep 28.

155. Healy JA, Nugent A, Rempel RE, et al. GNA13 loss in germinal center B cells leads to impaired apoptosis and promotes lymphoma in vivo. Blood. 2016;127(22): 2723–31. https://doi.org/10.1182/blood-2015-07-659938 Epub 2016 Mar 17.

156. Morin RD, Mungall K, Pleasance E, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood. 2013;122(7):1256–65. https://doi.org/10.1182/blood-2013-02-483727 Epub 2013 May 22.

157. Kolb JP, Roman V, Mentz F, et al. Contribution of nitric oxide to the apoptotic process in human B cell chronic lymphocytic leukaemia. Leuk Lymphoma. 2001;40(3–4):243–57.

158. Zhao H, Dugas N, Mathiot C, et al. B-cell chronic lymphocytic leukemia cells express a functional inducible nitric oxide synthase displaying antiapoptotic activity. Blood. 1998;92(3):1031–43.

159. Oakes CC, Seifert M, Assenov Y, Gu L, Przekopowitz M, Ruppert AS, et al. DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat Genet. (2016) 48:253–64. doi: 10.1038/ng.3488

160. Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, Allen SL, et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood (1999) 94:1840–7.

161. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood (1999) 94:1848–54.

162. Queirós AC, Villamor N, Clot G, Martinez-Trillos A, Kulis M, Navarro A, et al. A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. Leukemia (2015) 29:598–605. doi: 10.1038/leu.2014.252

163. Maag JLV, Fisher OM, Levert-Mignon A, Kaczorowski DC, Thomas ML, Hussey DJ, et al. Novel aberrations uncovered in barrett's esophagus and esophageal adenocarcinoma using whole transcriptome sequencing. Mol Cancer Res. (2017) 15:1558–69. doi: 10.1158/1541-7786.MCR-17-0332

164. Wang Q, Gan H, Chen C, Sun Y, Chen J, Xu M, et al. Identification and validation of a 44-gene expression signature for the classification of renal cell carcinomas. J Exp Clin Cancer Res. (2017) 36:176. doi: 10.1186/s13046-017-0651-9

165. Zhang YH, Huang T, Chen L, Xu Y, Hu Y, Hu LD, et al. Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets. Oncotarget (2017) 8:87494–511. doi: 10.18632/oncotarget.20903

166. Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and

clinical subdivision in chronic lymphocytic leukemia. Genome Res. (2014) 24:212–26. doi: 10.1101/gr.152132.112

167. Yepes S, Torres MM, Andrade RE. Clustering of expression data in chronic lymphocytic leukemia reveals new molecular subdivisions. PLoS ONE (2015) 10:e0137132. doi: 10.1371/journal.pone.0137132

168. Friedman DR, Weinberg JB, Barry WT, Goodman BK, Volkheimer AD, Bond KM, et al. A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. Clin Cancer Res. (2009) 15:6947– 55. doi: 10.1158/1078-0432.CCR-09-1132

169. Nazha A, Komrokji RS, Barnard J, Al-Issa, K., Padron, E., Madanat, Y. F, et al. A personalized prediction model to risk stratify patients with myelodysplastic syndromes (MDS). Blood (2017) 130(Suppl 1):160.

170. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep. (2017) 7:11707. doi: 10.1038/s41598-017-11817-6

171. Bender R. Introduction to the use of regression models in epidemiology. Methods Mol Biol. 2009;471:179–95. https://doi.org/10.1007/978-1-59745-416-2_9 PubMed PMID: 19109780.

172. Cafri G, Li L, Paxton EW, Fan JJ. Predicting risk for adverse health events using random forest. J Appl Stat. 2018;45(12):2279–94. https://doi.org/10.1080/02664763.2017.1414166.

173. A. Rodríguez, R. Villuendas, L. Yañez et al., "Molecular heterogeneity in chronic lymphocytic leukemia is dependent on BCR signaling: clinical correlation," Leukemia, vol. 21, no. 9, pp. 1984–1991, 2007

174. T. Herold, V. Jurinovic, K. H. Metzeler et al., "An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia," Leukemia, vol. 25, no. 10, pp. 1639–1645, 2011.

175. A. M. Ghaleb and V. W. Yang, "Kr¨uppel-like factor 4 (KLF4): what we currently know," Gene, vol. 611, pp. 27–37, 2017

176. S. Bellone, R. Tassi, M. Betti et al., "Mammaglobin B (SCGB2A1) is a novel tumour antigen highly differentially expressed in all major histological types of ovarian cancer: implications for ovarian cancer immunotherapy," British Journal of Cancer, vol. 109, no. 2, pp. 462–471, 2013.

177. L. Chen, D. Lu, K. Sun et al., "Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis," Gene, vol. 692, pp. 119–125, 2019.

178. J. Figueiredo, O. da Cruz e Silva, and M. Fardilha, "Protein phosphatase 1 and its complexes in carcinogenesis," Current Cancer Drug Targets, vol. 14, no. 1, pp. 2–29, 2014.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8 APPENDIX 1

8.1 Conflicts of interest statement

The author of this thesis has received honoraria from Jansen and AstraZeneca (advisory boards and seminars). The author also has been granted with research funds by Roche.

8.2 Ethics statement

The results of this thesis is based on pre-existing data based on anonymous data, which doesn't require a previous evaluation & approval by any ethics committee.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

9 APPENDIX 2

This thesis is based on a compendium of scientific papers published in peer-reviewed journals. The scientific director (in representation of all co-authors) has approved the publication of this thesis. PhD co-authors have explicitly provided written consent for the use of these papers for the present work. The author of this thesis (Adrián Mosquera Orgueira) has been responsible for the development of the research ideas, experimental procedures, data interpretation and scientific writing of all the papers included in this compendium. Co-authors have collaborated in data interpretation and manuscript revision.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10 APPENDIX 3

The present thesis is a compendium of the following published manuscripts:

1. Mosquera Orgueira A, Antelo Rodríguez B, Alonso Vence N, Bendaña López Á, Díaz Arias JÁ, Díaz Varela N, González Pérez MS, Pérez Encinas MM, Bello López JL. Time to Treatment Prediction in Chronic Lymphocytic Leukemia Based on New Transcriptional Patterns. Front Oncol. 2019 Feb 15;9:79. doi: 10.3389/fonc.2019.00079. PMID: 30828568; PMCID: PMC6384245.

    - 2-year Thomson-Reuters Impact Factor 2019-2020: 4.848
    - Quartiles: Cancer Research (Q1); Oncology (Q1)
    - Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
    - Copyright © 2019 Mosquera Orgueira, Antelo Rodríguez, Alonso Vence, Bendaña López, Díaz Arias, Díaz Varela, González Pérez, Pérez Encinas and Bello López. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

2. Mosquera Orgueira A, Rodríguez Antelo B, Díaz Arias JÁ, Díaz Varela N, Alonso Vence N, González Pérez MS, Bello López JL. Novel Mutation Hotspots within Non-Coding Regulatory Regions of the Chronic Lymphocytic Leukemia Genome. Sci Rep. 2020 Feb 12;10(1):2407. doi: 10.1038/s41598-020-59243-5. PMID: 32051441; PMCID: PMC7015923.

    - 2-year Thomson-Reuters Impact Factor 2019-2020: 3.998
    - Quartiles: Multidisciplinary (Q1)
    - Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
    - This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated

otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

3. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, Díaz Varela N, Bello López JL. A Three-Gene Expression Signature Identifies a Cluster of Patients with Short Survival in Chronic Lymphocytic Leukemia. J Oncol. 2019 Nov 7;2019:9453539. doi: 10.1155/2019/9453539. PMID: 31827514; PMCID: PMC6885206.

- 2-year Thomson-Reuters Impact Factor 2019-2020: 2.206
- Quartiles: Oncology (Q2)
- Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
- Copyright © 2019 Adrián Mosquera Orgueira et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

4. Mosquera Orgueira A, Antelo Rodríguez B, Alonso Vence N, Díaz Arias JÁ, Díaz Varela N, Pérez Encinas MM, Allegue Toscano C, Goiricelaya Seco EM, Carracedo Álvarez Á, Bello López JL. The association of germline variants with chronic lymphocytic leukemia outcome suggests the implication of novel genes and pathways in clinical evolution. BMC Cancer. 2019 May 29;19(1):515. doi: 10.1186/s12885-019-5628-y. PMID: 31142279; PMCID: PMC6542042.

- 2-year Thomson-Reuters Impact Factor 2019-2020: 3.150
- Quartiles: Oncology (Q1); Cancer Research (Q2); Genetics (Q2)
- Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
- This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

5. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, González Pérez MS, Bello López JL. New Recurrent Structural Aberrations in the Genome of Chronic Lymphocytic Leukemia Based on Exome-Sequencing Data. Front Genet. 2019 Sep 20;10:854. doi: 10.3389/fgene.2019.00854. PMID: 31616467; PMCID: PMC6764480.

- 2-year Thomson-Reuters Impact Factor 2019-2020: 3.258

- Quartiles: Genetics (Q1); Genetics (clinical) (Q1); Molecular Medicine (Q1)
- Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
- Copyright © 2019 Mosquera Orgueira, Antelo Rodríguez, Díaz Arias, González Pérez, and Bello López. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

6. Mosquera Orgueira A, Antelo Rodríguez B, Díaz Arias JÁ, Bello López JL. Identification of new putative driver mutations and predictors of disease evolution in chronic lymphocytic leukemia. Blood Cancer J. 2019 Sep 30;9(10):78. doi: 10.1038/s41408-019-0243-3. PMID: 31570692; PMCID: PMC6769000.

- 2-year Thomson-Reuters Impact Factor 2019-2020: 8.023
- Quartiles: Hematology (Q1); Oncology (Q1)
- Contribution of the PhD student: study design, methodology selection, data analysis, interpretation of results, manuscript writing.
- This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA