UNIVERSITAT DE BARCELONA

Implementació d'eines bioinformàtiques per la millora en el diagnòstic del càncer hereditari

Elisabet Munté Roca

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (**www.tdx.cat**) i a través del Dipòsit Digital de la UB (**diposit.ub.edu**) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (**www.tdx.cat**) y a través del Repositorio Digital de la UB (**diposit.ub.edu**) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (**www.tdx.cat**) service and by the UB Digital Repository (**diposit.ub.edu**) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Implementació d'eines bioinformàtiques per la millora en el diagnòstic del càncer hereditari

Tesi doctoral sotmesa per
Elisabet Munté Roca
per a l'obtenció del títol de doctorat
a la Universitat de Barcelona (UB)

Aquest treball ha estat desenvolupat sota la supervisió conjunta de la
Dra. Conxi Lázaro García i la Dra. Lídia Feliubadaló Elorza
a l'Institut de Recerca Biomèdica de Bellvitge (IDIBELL)
i a l'Institut Català d'Oncologia (ICO)

Programa de Doctorat de Biomedicina
Universitat de Barcelona (UB)
Tutor**: Dr. Francesc Viñals Canals**

Elisabet Munté Roca
Barcelona, 2024

AGRAÏMENTS

Sabeu la típica pregunta: *"On t'imagines d'aquí a deu anys?"* Us asseguro que el meu jo de fa una dècada mai hauria pensat que estaria aquí, escrivint els agraïments de la meva tesi doctoral. En aquell moment, vaig triar la carrera de Genètica, entre moltes altres opcions (extremadament diferents) que també em cridaven. I no, no puc dir que fos amor a primera vista. Els primers anys van ser complicats, el laboratori no feia per mi, i vaig arribar a pensar en abandonar. Però a tercer vaig descobrir un món que em va captivar: l'assessorament genètic. Per diversos motius, en aquell moment no va ser possible seguir aquest camí, però, a vegades, quan una porta no s'acaba d'obrir del tot, en sorgeixen d'altres que no havíem previst. Així és com vaig descobrir una nova passió: la bioinformàtica.

Ara, mentre escric aquestes línies, només puc pensar en totes les persones que s'han anat creuant en aquests últims anys i que, amb el seu granet de sorra, han fet que el camí fos aquest i, a més, que fos molt més amè, divertit i enriquidor.

Primer de tot, aquesta tesi no hauria estat possible sense el gran esforç de les meves dues directores: la Conxi i la Nònia.

Conxi, gràcies per donar-me l'oportunitat de formar part del teu grup i apostar per mi des del principi. Encara recordo el dia què em vas citar al teu despatx i em vas preguntar si mai havia considerat fer una tesi doctoral, que si em sumava a la idea de demanar una beca juntes! Va ser un punt d'inflexió que ho va canviar tot, i per això t'estaré sempre agraïda. M'has donat confiança i espai per créixer, tant com a investigadora com a persona. Gràcies per valorar-me, escoltar-me i guiar-me.

Nònia, no hi ha prou paraules per agrair-te tot el que has fet per mi. Ets, sense dubte, una de les persones més brillants que conec. És impressionant com el teu cervell connecta idees i va sempre un pas més enllà. Però, sobretot, mai tens un no per resposta, et desvius per ajudar els que t'envolten, vaja, tens un cor molt gran. Has estat una mentora excepcional i una font d'inspiració constant. Gràcies per les infinites hores que m'has dedicat; sense tu, aquesta tesi no hauria estat possible.

També vull fer un agraïment especial a altres persones que m'han guiat o amb qui he treballat colze a colze en els diferents articles d'aquesta tesi.

Marta, a més de ser una font inesgotable de coneixement i guiar-me en el projecte de *PMS2*, tens el radar de "mami" sempre activat per detectar quan estic amb les energies baixes i animar-me. Gràcies per la teva proximitat, pels consells, tant científics com personals, i per estar sempre al meu costat.

El projecte de *PMS2* tampoc hauria estat el mateix sense l'ajuda del **Jesús**, la persona amb més curiositat per descobrir coses peculiars**.** Gràcies pel teu toc d'humor i sarcasme, em dona la vida cada cop que et vinc a visitar! No desisteixis, que el Nobel arribarà! Jeje

José Marcos, el meu déu bioinformàtic. Gràcies per la teva paciència infinita i la capacitat d'organització i d'ensenyar. Treballar amb tu ha estat un autèntic plaer i una experiència de la qual he après moltíssim. Has fet que el "currazo" de les CNVs, l'article amb més figures suplementàries de la història, fos amè i divertit. No cal que et desitgi sort com a líder, perquè has demostrat de sobres que tens fusta i que ho petaràs. Estic supercontenta de continuar treballant amb tu.

Carla, em sento afortunada de poder compartir co-first amb una amiga de fa tant temps. Des del primer dia de la carrera... mare meva, no ha plogut ni res! Hem crescut plegades i, saps què més hem fet plegades? Surfejar per la BIOS del Linux com si fos casa nostra! Un doctorat en *how to* ens haurien de donar. Gràcies per ser tan pesada i per molestar-me tant, perquè em dona la vida. Visca les cosines Roca!

Paula, vull dir-te que estic molt contenta d'haver pogut treballar amb tu, sobretot aquest últim any. Rematar la tesi amb el projecte dels split reads al teu costat ha estat *lo más*! Fas fàcils les coses difícils, ets una supercompanya i una peça clau de l'equip!

Jose Luis Mosquera, quan vaig aterrar al laboratori amb poca experiència en bioinformàtica, sempre vas estar disposat a dedicar-me temps i a formar-me per poder afrontar el projecte del *vaRHC*. Gràcies per la teva paciència, suport i per ser una part tan important al principi d'aquesta aventura!

I would like to express my gratitude to **Alex Martin-Geary**, **Nicky Whiffin**, and the entire **Computational Rare Disease Genomics group** for warmly welcoming me during my week-long visit. Their comprehensive introduction to 5' UTRs—a subject previously unfamiliar to me—was both enjoyable and highly beneficial. The experience was truly enriching, and I am thankful for the opportunity to learn from such a dedicated team.

Però sabeu, perquè una cosa surti bé cal treball en equip, i tenim la sort que el càncer hereditari és una gran família amb experts de moltes branques.

Vull agrair a tot l'equip del **Servei de Diagnòstic Molecular**, amb els quals he tingut la sort de treballar conjuntament. A la **Sara**, per mostrar-me molt d'afecte i per garantir una excel·lent qualitat en el diagnòstic. A tots els tècnics especialistes: **Eva, Maribel, Mireia, Gardenia, Xavi** i **Dani**. M'alegreu sempre que us vinc a veure! També vull donar les gràcies a les **tècniques de laboratori** i al **personal administratiu**, perquè res d'això seria possible sense la vostra feina.

Al magnífic grup d'assessorament genètic, sense el qual, probablement no hauria pres aquest camí. Gràcies, **Joan Brunet,** per donar-me l'oportunitat d'entrar al grup el 2017! **Ares**, tu vas ser qui va acollir el pollet acabat de sortir de la carrera, i em vas ensenyar tot el que sabies amb paciència i amor. El vincle que m'uneix a tu és profund i sincer, i difícil de posar en paraules, perquè tot se li queda curt.

Èlia, admiro com lluites pel que sents; t'has convertit en una amiga per mi, i et trobem molt a faltar per aquí. **Mònica**, gràcies per trobar sempre el positiu i lluitar pels teus. **Sílvia**, la teva energia és contagiosa i encomanadissa, poses de bon humor a qualsevol que està a prop teu! **Àlex**, si mai et faltés feina com a oncòleg, podries guanyar-te la vida com a humorista, com he rigut amb tu! Moltes gràcies per ajudar-me professionalment i personalment! **Núria**, ets una de les persones més pràctiques que he conegut i amb tu sento que comparteixo moltes afinitats, gràcies per tots els consells. **Mati**, la persona más sincera, que las suelta sin miramientos y a la vez con el corazón tan y tan grande, gracias por preocuparte siempre por mi. ¡Querida **Carmen**, eres crac como pocas! Te deseo mucha suerte con la aventura del PhD. **Olga,** al·lucino amb la teva capacitat de gestionar-ho tot i amb la teva col·lecció infinita d'ulleres, fashion com tu poques! Gràcies per fer fàcil la feina de tothom. També vull agrair **l'equip de Girona i Badalona**!

Pel que fa al món de la recerca, també tinc molt a agrair. Gràcies, **Gabi**, per donar-me sempre un feedback tan valuós quan t'ho he demanat. **Bárbara**, son muchos donettes comidos a tu lado jaja me ha encantado compartir laboratorio contigo; has hecho que me sienta una más de tu grupo y has sido un ejemplo para mí. **Laura**, gracias por la enorme pasión y energía que transmites; ¡siempre eres el alma de la fiesta!

Jack, tenir-te just darrere ha estat un autèntic plaer. La teva manera d'entendre la vida i la ciència és admirable. M'has donat un dels records més macos que m'enduc de la tesi, la cançó del confinament!!!!

Fati, ha sigut molt guai tenir-te de companya al lab. De tu és molt fàcil aprendre, tens tantes però que tantes virtuts! Gràcies per dedicar-me temps i per preocupar-te per mi.

Agraïr també a la resta de post docs, lab i project managers que tenim: **Cris, Paula, Mariona, Mireia Morell, Pati i Guillem**. Senzillament gràcies per tot!

I em deixo pel final, el gran estament, el grup de **Precarios**, amb qui he compartit tants i tants moments. **Clara, Edgar, Julen, Lluís, Manuel, Pili, Sara, Arnau, Oscar, Sophie, José Luis, Marina, Míriam, Àngels, Júlia, Isa, Dimitria, Diego, estudiants de màster i grau i persones que ja he anomenat abans**. Heu fet que aquesta aventura sigui molt més divertia i amb chi chi chi. No m'imagino el PhD sense vosaltres i les vostres farres!

Fora del grup de càncer hereditari també he comptat amb grans aliats, l'equip del PADO. Gràcies en especial **David, Ania i Maria** per sempre estar disposats a donar-me un cop de mà i a fer un consell de savis bioinformàtics.

Ara bé, una part de l'èxit del meu PhD no ha passat dins les parets de l'IDIBELL, sinó que es deu al teixit fora de la feina. Em sento molt afortunada de la família i els amics que tinc.

D'una banda al meu compi de pis i amic, **Marc.** Conviure amb tu és molt fàcil, i hem compartit tantíssims moments d'estrès i de complicitat. Tu ja vas aconseguir les més que merescudes opos, i ara jo el PhD! *Olé nosaltres!* Gràcies per ser sempre tan comprensiu i un amic de veritat.

A la **Sònia**, per aparèixer de cop l'any passat per un dolor d'esquena i convertir-te en una bestie. M'encanta com de diferents i iguals som a la vegada. No m'imagino ara un dia a dia sense tu. Sempre tot per la causa! Ah i veieu la portada? Aquesta obra d'art és tota seva! Eternament agraïda.

Al **Martí,** gràcies per introduir-me en l'apassionant món de la Bioinfo i ajudar-me sempre que ho he necessitat! Comptar amb el teu recolzament ha estat clau per mi, sempre completament fascinada d'on pots arribar. No tens sostre.

A les de sempre **Quesi, Ire, Llore i Maria**. Perquè cadascuna ha pres camins molt diferents però ens seguim tenint allà. També als meus gym sisters, **Cuc, Ivan** i **Jansà,** amb qui he compartit tant últimament.

A totes les **nenes que he portat a vòlei durant aquests anys**. Ha estat estressant en alguns moments combinar les dues coses, però la vostra il·lusió és contagiosa i heu aconseguit encomanar-me l'energia fins i tot en aquells dies més complicats del PhD. També a les **sèniors**, per tants partits i tercers temps compartits.

I a tots els **Besties Mataró**, que vaig conèixer fa dos estius per casualitat i que s'han convertit en un pilar per mi. Que bonic és el vòlei, com uneix...

Pol, gràcies pel teu amor i suport. La teva voluntat per ajudar-me sempre, per cuidar-me en aquests moments d'estrès, disposat a ser *chico pa todo*: ara et cuino, ara t'arreglo el pc, ara et faig riure, ara juguem una estona a vòlei així et distreus... Tinc molta sort que siguis el meu company d'aventures, ets tot un pilar per mi.

Per últim, vull agrair als meus pares. **Mama,** et desvius per mi i em cuides amb un amor incondicional. Gràcies per totes les vegades que m'has ajudat quan jo no arribava. **Papa**, ets un tresor; admiro com sempre em fas reflexionar, la teva calma i paciència. Ets el meu model a seguir. Tinc una sort immensa de tenir-vos. Aquest títol també és vostre, perquè m'heu educat amb amor, esforç i valors, i heu treballat dur perquè jo pugui ser qui soc avui.

ÍNDEX

AGRAÏMENTS 3
ÍNDEX 9
ABREVIATURES 13
RESUM 19
INTRODUCCIÓ 23
1. El Genoma Humà 25
1.1. Estructura gènica 25
1.2. Variació genètica 32
2. Càncer 35
2.1. Bases genètiques del càncer 35
2.2. Predisposició genètica al càncer 37
3. Tècniques per detectar variants germinals 39
3.1. Seqüenciació Sanger 39
3.2. NGS 39
3.3 Seqüenciació de tercera generació 42
3.4 MLPA 43
3.5. LR-PCR 44
4. Anàlisi de dades d'NGS 45
4.1. Anàlisi primària 45
4.2. Anàlisi secundària 46
4.3. Anàlisi terciària 51
5. Funcionament del Programa de Càncer Hereditari a l'ICO - IDIBELL 52
6. Reptes en la identificació de variants utilitzant NGS de lectures curtes 56
7. Classificació de variants 60
7.1. Sistemes de classificació de variants 60
7.2. Implementació de la classificació de variants en laboratoris clínics 67
HIPÒTESIS 73
OBJECTIUS 77
INFORME DE LES DIRECTORES 81
ARTICLES 85
Article publicat 1: *Open-Source Bioinformatic Pipeline to Improve PMS2 Genetic Testing Using Short-Read NGS Data* 89
Article publicat 2: *Detection of germline CNVs from gene panel data: benchmarking the state of the art* 107
Article publicat 3: *vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines* 129
Resultat no publicat 1: *Optimizing GRIDSS for clinical use: a targeted NGS filtering strategy for germline structural variant detection* 155

Resultat no publicat 2: *Identifying potential pathogenic variants in 5'UTR regions within a hereditary cancer cohort* 167
RESUM DE RESULTATS 185
DISCUSSIÓ 193
Millora del rendiment diagnòstic a partir de l'optimització de dades de panell 195
Automatització de la classificació de variants 210
CONCLUSIONS 217
REFERÈNCIES 221
ANNEX 235
Taules suplementàries 237
Publicacions addicionals 239

ABREVIATURES

3'UTR	3' untranslated region	regions no traduïdes a 3'
5'UTR	5' untranslated region	regions no traduïdes a 5'
A	standalone	autònom
ACMG	American College of Medical Genetics and Genomics	Col·legi Americà de Genètica i Genòmica Mèdica
AMP	Association of Molecular Pathology	Associació de Patologia Molecular
BAM	binary alignment map	mapa binari d'alineament
BED	browse extensible data	dades extensibles de navegació
CanVIG-UK	Cancer Variant Interpretation Group UK	grup d'interpretació de variants del Regne Unit
CHIP	clonal hematopoiesis of indeterminate potential	hematopoesis clonal de potencial indeterminat
CHM	complete hydatidiform mole	mola hidatiforme completa
CMMRD	constitutional missmatch repair deficiency	deficiència constitucional de la reparació de desajustos
CNV	copy number variant	variant que afecta el nombre de còpies
cPAS	combinatorial probe-anchor synthesis	Síntesi combinatòria d'ancoratge de sondes fluorescents
CRAM	compressed reference-oriented alignment map	mapa comprimit d'alineament orientat per referència
CSS	circular Consensus Sequencing	seqüenciació de consens circular
ddNTP	dideoxynucleoside triphosphate	Didesoxiribonucleòtid trifosfat
dNTP	deoxynucleoside triphosphate	Desoxiribonucleòtids trifosfat
DPE	downstream promoter element	element promotor corrent avall
ET	transposable element	transposó o element transposable
GATK	the Genome Anlaysis Toolkit	kit d'anàlisi del genoma
GWAS	genome-wide association study	estudi d'associació de genoma complet
HC	hereditary cancer	càncer hereditari
HiFi	high fidelity	alta fiabilitat
HPRC	human pangenome reference consortium	Consorci de Referència del Pangenoma Humà
hTP53r	heritable *TP53*-related cancer syndrome	síndrome hereditària de càncer relacionada amb *TP53*
I2HCP	ICO-IMPPC hereditary cancer panel	panell de càncer hereditari ICO-IMPPC

ICO	Catalan Institute of Oncology	Institut Català d'Oncologia
IHQ	immunohistochemistry	immunohistoquímica
Inr	initiator	iniciador
IRES	Internal ribosome entry sites	llocs interns d'entrada del ribosoma
LINE	long interspersed nuclear element	Element nuclear intervingut llarg
LR-PCR	long range polymerase chain reaction	reacció en cadena de la polimerasa de llarg abast
LS	Lynch syndrome	síndrome de Lynch
LTR	long terminal repeat	repetició terminal llarga
M	moderate	moderat
MAF	minor allele frequency	freqüència de l'al·lel minoritari
MANE	Matched Annotation from the NCBI and EMBL-EBI	Anotació concordant del NCBI i de l'EMBL-EBI
MAVE	Multiplexed assay of variant effect	assajos massius de mutagènesi funcional
MLPA	multiplex ligation probe amplification	amplificació múltiple de sondes lligades
MMR	mismatch repair	reparació de desajustos
MTE	motif ten element	element de motiu deu
ncRNA	non-coding RNA	RNA no codificant
NGS	next generation sequencing	seqüenciació de segona generació
NIH	National Institute of Health	Institut Nacional de Salut
NMD	non-sense mediated decay	mecanisme de degradació mediat per codons sense sentit
oORF	overlapping open reading frame	marc de lectura obert solapant
P	supporting	de suport
pb	base pair	parell de bases
PCR	polymerase chain reaction	reacció en cadena de la polimerasa
PEM	paired-end mapping	mapatge d'extrems aparellats
PRS	polygenic risk score	puntuació de risc poligènic
PSV	paralogous sequence variant	variant paràloga de seqüència
RD	read depth	profunditat de lectura
S	strong	fort
SAM	sequence alignment map	mapa d'alineament de seqüències

SDMCH	Hereditary Cancer Molecular Diagnostics Service	Servei de Diagnòstic Molecular de Càncer Hereditari
SINE	short interspersed nuclear element	element nuclear escampat curt
SMRT	Single Molecule Real-Time	seqüenciació en temps real de molècula única
SNP	single nucleotide polymorphism	polimorfisme d'un sol nucleòtid
SNV	single nucleotide variants	variants d'un sol nucleòtid
SR	split reads	lectures dividides
STR	short tandem repeats	repeticions curtes en tàndem
SV	structural variant	variant estructural
T2T	Telomere-to-Telomere	De Telòmer a telòmer
TGS	third generation sequencing	Seqüenciació de tercera generació
TSG	tumor-supressor gene	gen supressor de tumors
TSS	transcription start site	lloc d'inici de la transcripció
TSV	tab-separated values	valors separats per tabulació
uAUG	upstream AUG	AUG corrent amunt
uORF	upstream open reading frame	marc de lectura obert corrent amunt
VAF	variant allele frequency	freqüència al·lèlica de la variant
VCF	variant call format	format d'identificació de variants
VS	very strong	molt fort
VSD	Variant of unknown significance	Variant de significat clínic desconegut
WES	whole exome sequencing	seqüenciació de l'exoma sencer
WGS	whole genome sequencing	seqüenciació de genoma sencer

RESUM

RESUM

Tot i que el càncer té sempre un origen genètic, només entre un 5% i un 10% dels casos són deguts a variants genètiques constitucionals que es poden transmetre a la descendència. Aquests casos s'anomenen càncer hereditari donat que presentar aquestes variants incrementa la possibilitat de desenvolupar càncer, heretant-se la predisposició o susceptibilitat genètica al càncer. D'altra banda, la transició de l'estudi de gens individuals a l'ús de tècniques de seqüenciació massiva de segona generació (NGS) ha revolucionat el camp de la genètica clínica, augmentat el nombre de variants identificades que cal classificar aplicant guies complexes de consens internacional. Aquest pas és sovint un coll d'ampolla en els laboratoris de diagnòstic genètic. Aquests estudis exhaustius han incrementat la proporció de casos on s'identifica una variant deletèria, tot i així encara existeix una proporció significativa en els que no s'ha determinat la causa genètica. Aquest fet es pot atribuir a diversos factors, com ara que la causa genètica estigui en gens no estudiats, que aquesta es localitzi en zones no analitzades en gens coneguts o que hagi passat desapercebuda per les limitacions inherents a la tecnologia NGS de lectures curtes.

Aquesta tesi doctoral, estructurada en cinc articles principals, aborda alguns d'aquests reptes en el camp del diagnòstic genètic en càncer hereditari i ha consistit en la implementació de noves metodologies així com el desenvolupant d'eines bioinformàtiques per maximitzar el rendiment diagnòstic. En concret: 1) S'ha desenvolupat PMS2_vaR, un *pipeline* per optimitzar l'anàlisi mutacional del gen *PMS2*, afectat per la presència del pseudogèn *PMS2CL*. PMS2_vaR ha augmentat la sensibilitat en la detecció de variants patogèniques i ha permès limitar l'ús de les LR-PCR únicament a la confirmació dels casos positius. 2) S'ha elaborat una comparativa exhaustiva de dotze eines per a la detecció de variants que afecten el nombre de còpies, destacant l'ús de GATK-gCNV com una eina prometedora en contextos diagnòstics. 3) S'han dissenyat filtres específics per prioritzar les variants detectades per GRIDSS, un algoritme capaç d'identificar variants estructurals de mida mitjana i variants equilibrades, que actualment passen desapercebudes en la rutina diagnòstica. Aquesta aproximació ha permès augmentar el rendiment diagnòstic un 0,6%, amb la identificació de vuit variants rellevants en gens d'alt i moderat risc, cinc de les quals eren insercions de transposons en la regió codificant. 4) S'han explorat les regions 5'UTR, no analitzades habitualment en un context de diagnòstic, aplicant filtres descrits per experts per prioritzar variants potencialment rellevants amb impacte en la regulació gènica i la traducció proteica. 5) S'ha dissenyat vaRHC, un paquet d'R que semi-automatitza la classificació de variants seguint les guies ACMG i les guies específiques de gens de ClinGen. Aquesta eina ajuda a agilitzar el procés de classificació, reduint un dels principals colls d'ampolla en les unitats de diagnòstic molecular.

Aquest treball destaca la importància de desenvolupar eines adaptades a les limitacions tecnològiques actuals i posa les bases per a un diagnòstic genètic més precís i eficient, amb l'objectiu de resoldre el màxim nombre de famílies amb predisposició genètica al càncer.

PARAULES CLAU:

Càncer Hereditari; NGS; Diagnòstic; Variants estructurals; Classificació de variants.

INTRODUCCIÓ

1. El Genoma Humà

El descobriment de l'estructura del DNA per Watson i Crick el 1953 va marcar un punt d'inflexió en el camp de la biologia i la genètica. En aquesta molècula, composta per uns 3.200 milions de bases, rau la clau per comprendre la transmissió de l'herència i l'origen de moltes malalties. Cal destacar, però, que aquest avenç no hauria estat possible sense els estudis de difracció de raigs X proposats per Wilkins i realitzats per Rosalind Franklin, qui morí a l'edat de trenta-set anys sense rebre mai el reconeixement merescut.

El 1990 es va iniciar el projecte Genoma Humà, una col·laboració internacional amb l'objectiu de desxifrar el DNA d'un grup d'individus. Es va completar el 2003, amb el 99% de l'eucromatina seqüenciada (International Human Genome Sequencing Consortium*, 2004). Aquest projecte va revolucionar la investigació biomèdica, ja que va proporcionar informació sobre la distribució de gens (20.000-25.000) i transposons, d'entre molts altres aspectes.

Des de llavors, s'han publicat diverses versions del genoma humà, destacant la GRCh37 (2009), que va millorar la cobertura i va corregir errors, i la GRCh38 (2013) que va millorar la precisió de la seqüència i va incorporar regions seqüenciades addicionals, especialment en àrees pericentromèriques i subtelomèriques.

Finalment, el 2022 es va completar la primera seqüència completa de telòmer a telòmer (T2T-CHM13), de tots els autosomes i cromosomes sexuals (Nurk *et al.*, 2022).

Tot i aquest avenç, les versions GRCh37 i GRCh38 continuen sent àmpliament utilitzades, ja que moltes eines bioinformàtiques estan optimitzades per a elles.

1.1. Estructura gènica

El genoma humà està estructurat en gens i seqüències relacionades amb gens separats per DNA intergènic (Figura 1, Taula 1).

Taula 1: Nombre gens descrits en l'última versió (v.46) de GENCODE

Element	**Nombre**	
Gens que codifiquen per proteïna	19411	
Gens transcrits a RNA com a producte final	27875	20310 lncRNA
		7565 sncRNA
Pseudogens	14716	10657 pseudogens processats
		3564 pseudogens no processats
		258 pseudogens unitaris

Extret de GENCODE (https://www.gencodegenes.org/human/stats_46.html, data d'últim accés: 26 d'agost 2024). LncRNA: RNA no codificant llarg (de l'anglès *long non-coding RNA*), sncRNA: RNA no codificant curt (de l'anglès *short non-coding RNA)*.

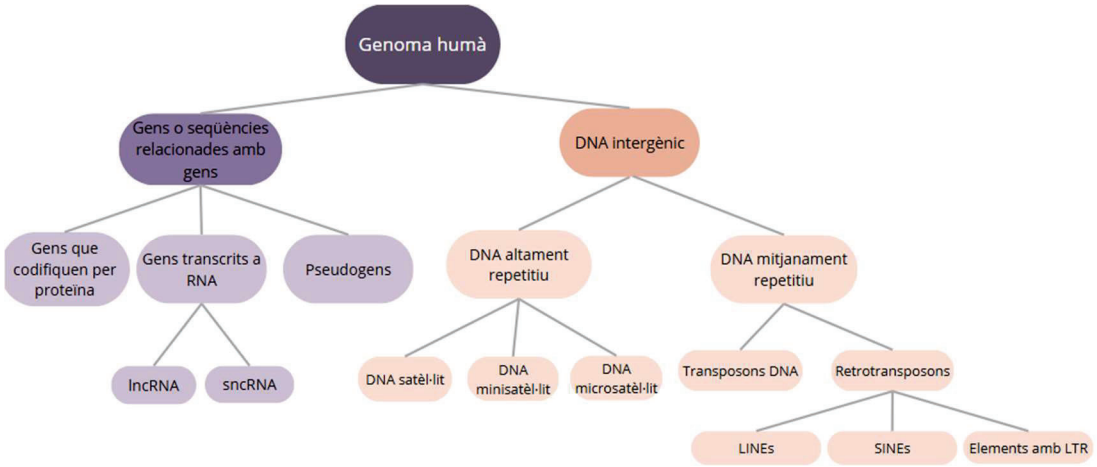


Figura 1: Principals elements del genoma humà. LncRNA: RNA no codificant llarg (de l'anglès *long non-coding RNA*), sncRNA: RNA no codificant curt (de l'anglès *short non-coding RNA)*, LINE: elements nuclear escampat llarg (de l'anglès long interspersed nuclear element), SINE: element nuclear escampat curt (de l'anglès short interspersed nuclear element), LTR: repeticions terminals llargues (de l'anglès long terminal repeats).

Gens que codifiquen per proteïna

Els gens que codifiquen per proteïna en eucariotes són estructures complexes, compostes per diferents regions que es troben altament regulades a diferents nivells (Figura 2). Generalment són gens monocistrònics, és a dir, codifiquen per una única proteïna, i la majoria es troben inactius en les cèl·lules diferenciades. Els gens estan estructurats en seccions contigües anomenades exons i introns, precedides per una regió anomenada promotor.

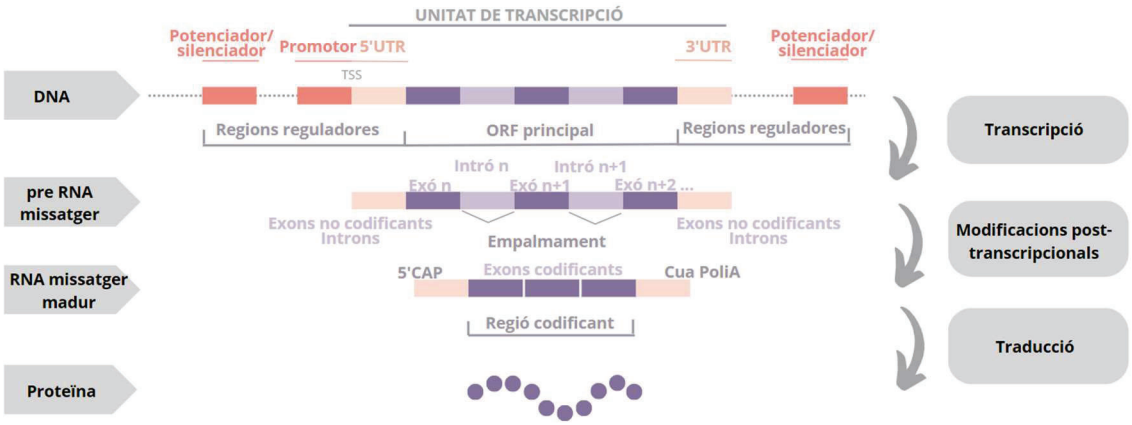


Figura 2: Principals parts estructurals d'un gen que codifica per proteïna i biosíntesi de les macromolècules. Els potenciadors/silenciadors i promotors regulen la transcripció del DNA a pre-RNA missatger. Aquest pateix una sèrie de modificacions post-transcripcionals, les quals eliminen els introns i afegeixen el 5'-CAP i la cua Poli-A. Finalment, les regions 5' i 3' UTR regulen la traducció del RNA missatger madur a proteïna final. Aquesta figura representa una simplificació del procés, ja que s'han descrit mecanismes que poden modificar el flux de la informació genètica, com la retrotranscripció (Baltimore, 1970) o mecanismes independents d'alguns d'aquests passos Adaptat de Shafee i Lowe (2017) i de Pakay et al., (2023).

Promotors

Són regions generalment curtes de DNA que comencen corrent amunt (5') del lloc d'inici de la transcripció (TSS, de l'anglès *transcription start site*) i es poden estendre fins una mica més enllà d'aquest. Engloben diferents estructures reguladores (Figura 3) que permeten el reconeixement per

part de la polimerasa II d'RNA, així com d'altres factors de transcripció imprescindibles per iniciar la transcripció i assegurar-ne la freqüència adequada, considerant el gen i el teixit.

Figura 3: Principals elements que poden formar part de l'estructura d'un promotor basal en humans: la caixa BRE, localitzada a unes 35pb corrent amunt del TSS, i la caixa TATA, situada entre unes 25-30bp corrent amunt del TSS, ambdues recluten factors de transcripció (Chen et al., 2021); l'iniciador (Inr), l'element de motiu deu (MTE, de l'anglès motif ten element), que es localitza entre 18 i 27 pb corrent avall del TSS i promou la transcripció (Yan Lim et al., 1999) i l'element promotor corrent avall (DPE, de l'anglès downstream promoter element), també essencial per al reconeixement d'algunes subunitats de factors de transcripció. Figura adaptada de Roy (2005).

Exons

Els exons són seqüències que es transcriuen a RNA missatger i que, un cop fet l'empalmament (*splicing, en anglès)* es mantenen en el transcrit final. Poden ser codificants i, per tant, ser traduïts a proteïna, o no codificants, i desenvolupar altres funcions reguladores necessàries per al gen.

Introns

Els introns són seqüències que es transcriuen a pre-RNA missatger, però són eliminades durant la maduració d'aquest gràcies al procés d'empalmament. La maquinària d'empalmament és capaç de reconèixer aquestes regions, perquè la majoria dels introns tenen una seqüència de consens en 5' o lloc donador i una a 3' o lloc acceptor (Mount, 1982).

A més, gràcies a l'empalmament alternatiu es poden formar diferents RNA missatgers amb diferents continguts d'exons. Així, una regió intrònica per una proteïna pot ser exònica per una altra isoforma.

Altres regions

Hi ha altres elements que encavalquen amb els anteriors, amb funcions generalment reguladores, definits segons la seva activitat i/o localització.

Potenciadors

Són estructures que augmenten l'eficiència de la transcripció fomentant el reclutament de factors de transcripció. No es troben en un punt fix dins del gen, sinó que es poden localitzar tan corrent amunt com avall com inclús al mig del gen.

Silenciadors

Són regions on s'uneixen factors que permeten que s'estableixi una compactació de la cromatina, inhibint l'expressió del gen.

Aïlladors

Per contra, els aïlladors impedeixen l'efecte dels potenciadors o silenciadors sobre el promotor. Inhibeixen la propagació de les modificacions de la cromatina provocant-ne el bloqueig, no la inactivació.

5'UTR

Les regions no traduïdes a 5' (5'UTR, de l'anglès *5' untranslated regions*) són essencials per a la regulació de l'expressió gènica. Aquestes regions, de mida variable entre gens, s'estenen des del lloc d'inici de la transcripció fins al codó d'inici de la traducció (Pesole *et al.*, 2001). Formen part de l'RNA missatger perquè es transcriuen, però no de la proteïna, ja que no es tradueixen.

Contenen estructures secundàries i terciàries d'RNA així com altres elements que participen en la regulació post-transcripcional. Intervenen en l'estabilitat de l'RNA, la seva localització dins la cèl·lula i l'eficiència de la traducció (Mignone *et al.*, 2002; Leppek, Das and Barna, 2018).

Per exemple, una modificació post-transcripcional clau del pre-RNA missatger en la regió 5'UTR consisteix en l'addició d'una guanina metilada a l'extrem 5'. Aquest procés és catalitzat per la guanil-transferasa i acompanyat de metilacions addicionals . Aquesta guanina s'anomena cap o caputxó i és essencial per protegir la degradació del m-RNA i per poder iniciar la traducció.

Taula 2: Descripció estadística de les 5'UTR anotades als 18.764 transcrits MANE *select*

Mida	Rang	1-3.561pb
	Mitjana	202 pb
	Mediana	136 pb
uAUG	Rang (per gene)	0-64
	Gens amb almenys 1 uAUG	42,5%
Introns	Rang	0-11
	Gens amb 1 o més introns	37,7%

El Projecte MANE(de l'anglès Matched Annotation from the NCBI and EMBL-EBI): unifica l'anotació de gens humans entre NCBI i EMBL-EBI. Per cada locus codificant, es selecciona un trasncrit representatiu (MANE Select) que serveix com estàndard universal. Traduït de Wieder et al. (2024). uAUG: codó d'inici aigües amunt del ORF principal.

S'ha vist que les 5'UTR poden tenir marcs de lectura oberts previs al codó d'inici de la traducció principal (uORF, de l'anglès *upstream open reading frame*). La mida dels uORFs, així com la quantitat d'exons i introns que contenen varia molt entre gens (Taula 2) (Wieder et al., 2024). En alguns casos, els uORFs poden ser reconeguts pels ribosomes i traduïts a polipèptids, que poden interferir amb la traducció de l'ORF principal. Així doncs, poden esdevenir reguladors en *cis* teixit-específics, podent arribant a reduir significativament l'expressió de la proteïna principal, en alguns casos fins al 80% (Calvo, Pagliarini and Mootha, 2009).

Els uORFs poden iniciar la traducció a partir del codó canònic AUG, però també s'ha vist que, sota certes condicions, poden començar la traducció amb una eficiència considerable a partir de codons no canònics. Per exemple, el codó CUG pot arribar a assolir fins al 50% de l'eficiència del codó canònic en un context de Kozak òptim, mentre que els codons ACG i GUG poden assolir fins a un 40% i 20% d'eficiència, respectivament. A més, tot i que amb menys eficiència, també s'ha vist que els codons AUU, AUA, AUC i UUG presenten nivells detectables de traducció (De Arce, Noderer and Wang, 2018; Andreev et al., 2022; Chothani et al., 2022).

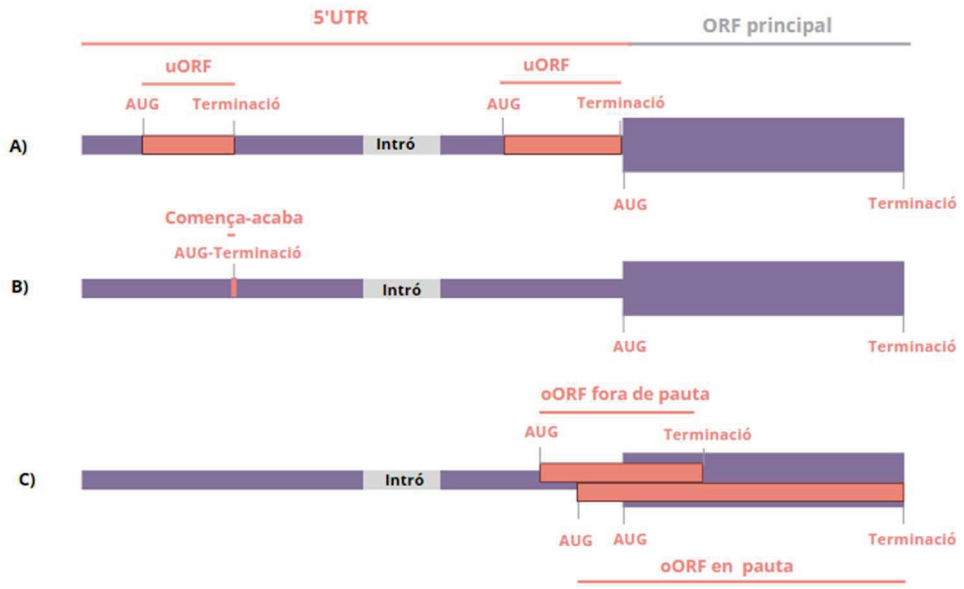


Figura 4: Resum dels diferents tipus d'ORF que es poden donar a 5'UTR. Representació del pre-RNA missatger. A) uORFs que no es superposen amb l'ORF principal. Poden estar llunyans o més propers a l'AUG de l'ORF principal, inclús arribant a tocar-se, però no encavalcant-se. Contenen codons entre l'AUG i el codó terminació. B) uORF de comença-acaba, tampoc encavalca l'ORF principal i no conté codons entre el codó d'inici i el de terminació C) uORFs encavalcats amb l'ORF principal (oORFs), ja sigui en pauta o fora de pauta. Adaptat de Wieder et al. (2024).

Els uORFs poden estar superposats o no amb l'ORF principal i poden incloure un nombre variable de codons (Figura 4, A i C) o inclús no contenir-ne, és a dir, l'AUG va immediatament seguit d'un codó terminació (Figura 4, B). Quan els uORFs superposen l'ORF principal esdevenen ORFs encavalcats (oORFs, de l'anglès *overlapping open reading frames*) (Figura 4, C). Aquests els podem trobar dins de la mateixa pauta de lectura que l'ORF principal, estenent-se fins al seu codó stop, o bé poden estar fora de pauta i acabar amb un codó terminació diferent del de l'ORF principal.

La traducció dels uORFs ve condicionada en gran part pel context local de la seqüència. És especialment rellevant la seqüència de Kozak, que és una seqüència de consens òptima al voltant del codó d'inici, crítica per l'eficiència de la iniciació de la traducció (GCCRCC**AUG**GG, on la R és una purina (A/G)) (Kozak, 1986, 1987a, 1987b). S'ha demostrat que dins d'aquesta seqüència, la presència d'una adenina o guanina a la posició -3 i una guanina a la posició +4 respecte a la A de l'AUG incrementen significativament la probabilitat de reconeixement del codó d'inici per part de la maquinària de traducció (Kozak, 1986, 1997). Estudis posteriors han demostrat que també les posicions -2, -4 i +5 influencien i han avaluat experimentalment totes les combinacions possibles (NNNNNNAUGNN, on la N és qualsevol nucleòtid) per determinar-ne l'eficiència de la traducció (Noderer *et al.*, 2014a). Si un primer AUG es troba en un context desfavorable segons la seqüència de Kozak, el ribosoma pot no reconèixer-lo com a punt d'inici i continuarà avançant per l'RNA missatger fins a trobar un AUG en un entorn més favorable. Aquest procés, conegut com a *leaky scanning*, permet que la traducció comenci en un ORF situat aigües avall, on la seqüència de Kozak és més favorable (Figura 5, A) (Ao-Kondo *et al.*, 2011).

D'altres vegades, malgrat que el ribosoma reconeix un codó AUG aigües amunt (uAUG, de l'anglès *upstream AUG*), la subunitat petita es manté unida a l'RNA missatger, provocant la re-iniciació de la traducció a l'AUG del gen (Ao-Kondo *et al.*, 2011). Aquest mecanisme, anomenat re-iniciació, s'ha vist més probable com més lluny es localitza el codó de terminació de l'uORF de l'ORF principal i també quan la longitud de l'uORF és curta (Figura 5, B) (Morris and Geballe, 2000).

La professora Whiffin, referent en l'estudi de les regions UTR, va liderar el desenvolupament d'UTRannotator, una eina que prediu l'efecte de les variants en les 5'UTR (Zhang *et al.*, 2021). Aquesta eina, incorporada a Variant Effect Predictor (VEP) d'Ensembl (McLaren et al., 2016), no només identifica si una variant crea o destrueix un uORF, sinó que també determina quin subtipus de uORF es crea o es destrueix i la força de la seqüència de Kozak al voltant. A més, per a les variants que destrueixen un uORF avalua si hi ha evidència experimental prèvia que indiqui que aquest uORF era traduït.

Posteriorment, la professora Whiffin va crear el grup de Genòmica Computacional de Malalties Rares, centrat en l'estudi d'aquestes regions UTR. El grup ha desenvolupat una pàgina web (https://vutr.rarediseasegenomics.org/) que ofereix una interfície intuïtiva per visualitzar l'impacte de les variants trobades a les bases de dades gnomAD (v.3.1.2) i ClinVar en els transcrits *MANE select*. Aquesta eina permet als usuaris seleccionar i visualitzar l'arquitectura de les regions 5'UTR, introduir variants en aquestes regions i analitzar-ne les conseqüències, ja que té el UTRannotator integrat.

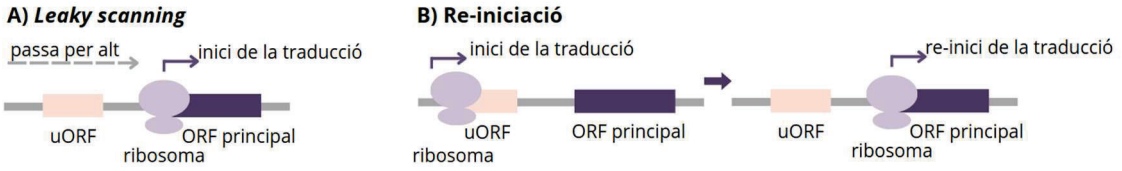


Figura 5: Models no convencionals d'escaneig del ribosoma en l'RNA missatger d'eucariotes. A*) Leaky scanning*: el ribosoma passa per alt l'uORF i inicia la traducció a l'ORF principal; B) Re-iniciació: el ribosoma inicia la traducció a l'uORF però després es torna a reiniciar a l'ORF principal. Adaptat d' Ao-Kondo et al. (2011).

A més, a les 5'UTR també es poden trobar llocs interns d'entrada del ribosoma (IRES, de l'anglès *internal ribosome entry sites*). Aquestes regions de l'RNA missatger, encara força desconegudes, actuen en *cis* i són capaces de reclutar la subunitat ribosòmica 40S, especialment rellevant per la traducció de gens policistrònics. La capacitat de promoure la traducció de manera independent del caputxó permet la traducció dels ORFs situats aigües avall (Yang and Wang, 2019).

3'UTR

Les regions no traduïdes a 3' (3'UTR, de l'anglès *3' unstraslated regions*) s'estenen des del codó de terminació fins al lloc final de la transcripció (Pesole *et al.*, 2001). Igual que les 5'UTR, són regions que es transcriuen però que no es tradueixen a proteïna i juguen també un paper important en la regulació post-transcripcional, condicionant la localització, estabilitat i traducció de l'RNA missatger.

Un exemple d'aquesta funció reguladora és la poliadenilació de l'extrem 3', un procés en el qual s'afegeix una cua poli-A, d'uns 250 residus, al pre-RNA missatger. Aquesta cua poli-A facilita la terminació de la transcripció, augmenta l'estabilitat de la molècula, ja que la protegeix de la degradació, facilita la seva exportació del nucli al citoplasma i també promou la traducció (Perales and Bentley, 2009; Weill *et al.*, 2012).

Gens transcrits a RNA com a producte final

Aquest grup està compost per gens que desenvolupen una funció dins la cèl·lula, però que no són codificants (ncRNAs, de l'anglès *non-coding RNA*), és a dir, es transcriuen a RNA, però no es tradueixen a proteïna. Es pensa que poden tenir un paper clau en molts processos biològics com la regulació de l'expressió d'altres gens, l'empalmament, la modulació de l'estructura de la cromatina, la regulació a nivell epigenètic, la reparació del DNA, i fins i tot poden ser rellevants en el desenvolupament tumoral (Amaral *et al.*, 2023; Poliseno, Lanza and Pandolfi, 2024).

Pseudogens

Els pseudogens provenen de la duplicació de gens que acumulen mutacions i es tornen inactius. Comparteixen alta homologia amb algun gen del genoma tot i que hi presenten alguna diferència. Per exemple, hi ha pseudogens que no tenen la regió promotora, mentre que d'altres tenen un codó de terminació al mig, fragments de gens, introns o regions no traduïdes (Poliseno *et al.*, 2024).

Tanmateix, estudis recents han revelat que els pseudogens poden tenir un paper important en la regulació gènica a diferents nivells. Aquestes funcions inclouen la influència en el desenvolupament i la progressió de malalties, especialment el càncer, demostrant que no són "escombraries" genètiques com antigament es creia (Chen *et al.*, 2020).

El genoma humà conté poc més de 14.700 pseudogens, dels quals un 10% es transcriuen (Taula 1). Els pseudogens es classifiquen en tres tipus: processats, derivats de retrotransposició, sense introns i ubicats en cromosomes diferents del gen d'origen; no processats, provinents de la duplicació de gens, amb introns, sovint amb regions reguladores i ubicats al mateix cromosoma que el gen parental; i unitaris, que resulten de l'acumulació d'alteracions en un gen ancestral sense còpia al genoma.

DNA intergènic

El DNA intergènic representa aproximadament unes 2.000 MB del genoma (62,5%) i comprèn les regions de DNA que se situen entre els gens. Conté elements reguladors i DNA repetitiu.

El DNA repetitiu pot ser altament repetitiu o mitjanament repetitiu. El primer inclou DNA satèl·lit, localitzat en zones centromèriques; DNA minisatèl·lit, localitzat en regions subtelomèriques, telomèriques i altament variables; i DNA microsatèl·lit, distribuït al llarg dels cromosomes. Aquests tipus de DNA es diferencien segons la longitud de les seqüències repetitives i el grau de repetició.

El DNA mitjanament repetitiu engloba els transposons o elements transposables (ET), fragments de DNA capaços de desplaçar-se i inserir-se en nous llocs del genoma. Aquests elements constitueixen aproximadament el 45% del genoma humà (Figura 6). Tenen un paper clau en la plasticitat del genoma, ja que són responsables de la recombinació entre cromosomes (Chénais Biosse, 2022). Com que no tenen una diana específica, en alguns casos poden generar mutacions que trunquen proteïnes, fet que pot predisposar a condicions genètiques.

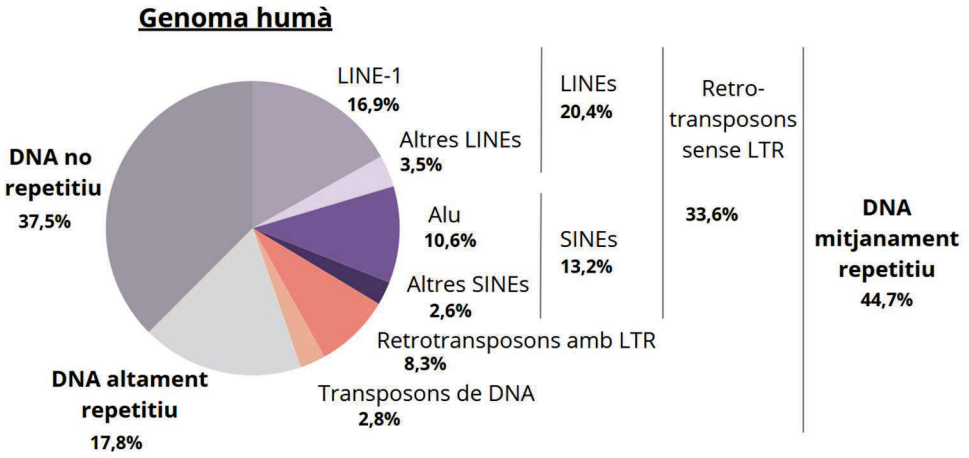


Figura 6: Proporció dels elements transposables en el genoma humà segregat per famílies i subfamílies. El percentatge és respecte tot el genoma. Adaptat de Chénais Biosse (2022) amb les dades de Lander et al. (2001).

En funció del mecanisme de transposició emprat, els ET poden classificar-se en retrotransposons (elements de classe I) o transposons de DNA (elements de classe 2) (Wicker *et al.*, 2007).

Retrotransposons

Els retrotransposons fan servir un mecanisme replicatiu, és a dir, augmenten el nombre de còpies en cada transposició. Inicialment, es transcriuen a RNA intermediari que serà després retrotranscrit per una transcriptasa reversa i s'inserirà a un nou lloc.

Dins d'aquest grup, podem diferenciar dos subgrups principals: els elements amb repeticions terminals llargues (LTR, de l'anglès *long terminal repea*ts) i els que no posseeixen aquestes regions terminals. Mentre que el primer grup no és molt freqüent en genoma humà (8,3%), el segon subgrup representa aproximadament un 33,6% (Figura 6) (Chénais Biosse, 2022).

Entre els elements sense LTR, destaquen els LINE i els SINE (de l'anglès *long interspersed nuclear elements* i *short interspersed nuclear elements*, respectivament). La principal diferència entre aquests dos és que els SINE són més curts perquè han perdut part de la seqüència i, per tant, no poden replicar-se de manera autònoma. Així doncs, per poder-se transposar depenen de les proteïnes produïdes pels elements LINE. Dins dels SINE, els elements *Alu* són especialment abundants en el genoma humà (10,6%).

Transposons de DNA

Els transposons de DNA utilitzen un mecanisme de transposició conservatiu, és a dir, el nombre de còpies d'aquests elements no augmenta. Per aquest motiu representen únicament un 2,8% del genoma humà. Es poden transposar perquè són capaços d'escindir-se i després reinserir-se a una altra regió, sense necessitat de transcriure's a RNA intermediari.

1.2. Variació genètica

Els humans comparteixen el 99,9% del DNA. Per petit que pugui semblar, aquest 0,1% és suficient per explicar les notables diferències individuals, com per exemple, l'aspecte físic, certs trets de comportament o la susceptibilitat a malalties. No obstant això, cal recordar que el fenotip observat és fruit de la interacció amb l'ambient i estils de vida de cada individu.

La variació genètica es deu a canvis tant en la composició com en l'estructura del nostre DNA, i engloba des de variants d'un sol nucleòtid (SNV, de l'anglès *single nucleotide variants*) fins a alteracions més grans, com ara variants que involucrin a cromosomes sencers (Figura 7).

En aquest sentit, el genoma de referència, encara que útil, és insuficient per representar adequadament la diversitat genètica global. Cada població té variants genètiques úniques que poden tenir un impacte significatiu sobre la salut i els trets biològics.

Per abordar aquesta problemàtica, el *Human Pangenome Reference Consortium* (HPRC) ha publicat recentment un esborrany d'un pangenoma humà a partir de 47 genomes que representen la diversitat genètica, i que permeten reduir el biaix a causa de la situació geogràfica (Liao *et al.*, 2023).

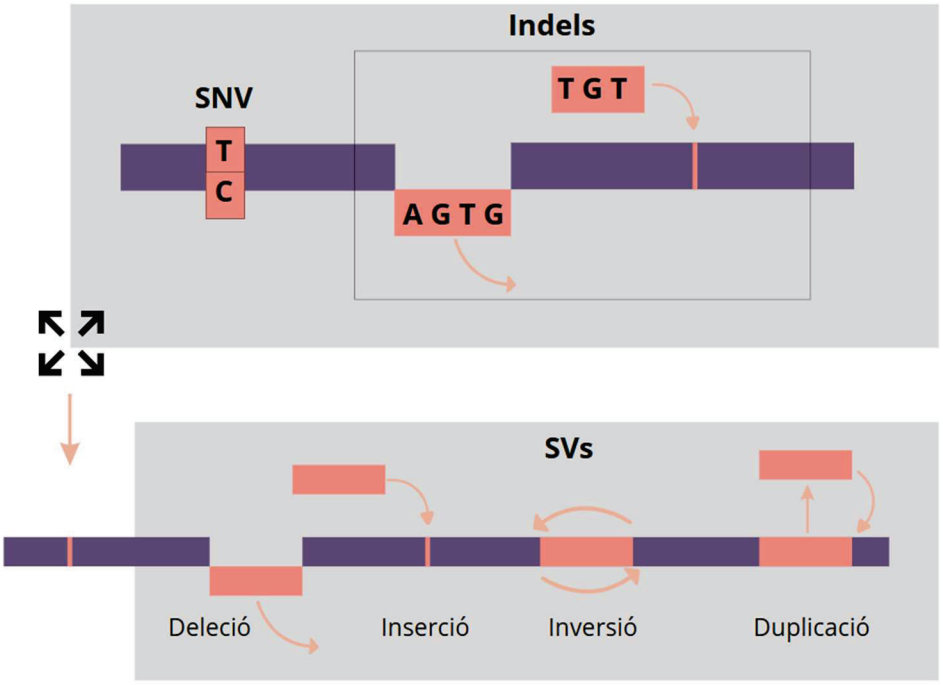


Figura 7: Diferents tipus de variants genètiques. En el requadre de dalt es mostren les variants d'un únic nucleòtid (SNVs) i les indels. En el de sota es veuen exemples de variants estructurals (SV). La transposició es considera un cas particular de les insercions que poden o no estar precedides de la deleció del fragment que s'inserta. Adaptat de Smith, Kawash and Grigoriev (2017).

SNV

Les SNV són variants genètiques molt comunes, presents en cada genoma humà amb una freqüència d'entre 4 i 5 milions (Eichler, 2019). Es basen en substitucions que afecten a un sol nucleòtid.

Les SNV poden ocórrer tant en regions codificants com no codificants del genoma. Dins de les regions codificants, les SNV poden tenir diferents efectes sobre la proteïna resultant. D'una banda, hi ha variacions sinònimes o silents, on a causa de la degeneració del codi genètic, donen lloc al mateix aminoàcid, fent que no solguin tenir impacte en la proteïna. D'altra banda, hi ha SNV no sinònimes (*missense*, en anglès), que provoquen un canvi d'aminoàcid. Finalment, les mutacions sense sentit (*non-sense*, en anglès) es donen quan la variant provoca l'aparició d'un codó de terminació (UAA, UAG o UGA) prematur generant una proteïna truncada. En funció d'on es localitza la mutació sense sentit es pot activar el mecanisme de degradació mediat per codons sense sentit (NMD, de l'anglès *non-sense mediated decay*). És un sistema que tenen les cèl·lules per identificar i destruir les molècules d'RNA missatger que contenen errors i així evitar la síntesi de proteïnes no funcionals o nocives.

A més, també hi ha variants en regions codificants o no codificants que poden condicionar el mecanisme d'empalmament. Aquestes variants poden provocar la no inclusió d'exons (*exon-skipping,* en anglès), la retenció d'introns (*intron-retention*, en anglès) o l'allargament o escurçament d'exons, que poden acabar codificant una proteïna anòmala.

Petites insercions i delecions (indels)

Les indels són delecions o insercions que involucren menys de 50 parells de bases. Es troben a una freqüència aproximada d'entre 700.000 i 800.000 per genoma (Eichler, 2019).

Aquestes variants es localitzen tant en regions codificants com no codificants del genoma. En les regions codificants, el seu efecte depèn de si la seva longitud és múltiple de tres, la qual cosa mantindria la pauta de lectura (*in-frame,* en anglès), o si no ho és, fet que l'alteraria (*frameshift,* en anglès). Aquestes últimes causen un canvi en la seqüència d'aminoàcids que pot provocar l'aparició d'un codó de terminació prematur, tenint un efecte equivalent a les mutacions *non-sense*.

Variants estructurals

Les variants estructurals (SV, de l'anglès *structural variant*) són regions del DNA que engloben des de 50 pb fins a vàries megabases i que mostren diferències respecte el genoma de referència. Aquestes diferències poden ser desequilibrades, afectant així la dosi (delecions, insercions o duplicacions) o equilibrades, afectant o bé l'orientació (inversions) o la localització (translocacions) (Escaramís et al., 2015).

Aquestes variants tenen un paper clau en la variació genètica entre individus. A més, s'ha observat patrons de SVs específics en diferents poblacions, que modulen l'evolució genòmica (Levy-Sakin *et al.*, 2019).

Són especialment rellevants les variants que afecten el nombre de còpies (CNV, de l'anglès *copy number variant*), ja que provoquen una situació de desequilibri, principalment causada per delecions i duplicacions. S'ha demostrat que les CNVs constitueixen una proporció significativa del genoma humà, estimada entre un 4,8 i un 9,5% (Zarrei *et al.*, 2015). A part de contribuir a la variabilitat genètica, poden influir en certs processos biològics, arribant a causar o predisposar a diverses malalties, d'entre elles, el càncer (Shlien and Malkin, 2009; Pinto *et al.*, 2010; Wheeler *et al.*, 2013; Shaikh, 2017).

Variants epigenètiques

Les variants epigenètiques són modificacions del material genètic que afecten l'expressió gènica sense alterar la seqüència de DNA. Engloben, entre d'altres, la metilació de citosines, modificacions de les cues de les histones que fan més o menys accessible la cromatina, miRNA i ncRNA. Cada cèl·lula té una signatura epigenètica única, que reflecteix tant el genotip com les influències ambientals, per exemple, l'alimentació, l'estrès o l'exposició a tòxics. Aquest tipus de variants no s'han tractat en aquesta tesi doctoral.

2. Càncer

2.1. Bases genètiques del càncer

El càncer és la segona causa de mort als països desenvolupats, només darrere de les malalties cardiovasculars. Tanmateix, no s'ha de considerar com una única patologia, sinó que engloba més de 200 malalties heterogènies, amb una alta variabilitat morfològica i pronòstica. Totes elles es basen en un procés seqüencial en el qual un conjunt de cèl·lules acumulen errors genètics durant les divisions cel·lulars. Aquests errors fan que escapin del control cel·lular i puguin proliferar de manera descontrolada, desplaçar la resta de cèl·lules sanes, envair els teixits veïns i acabar produint metàstasis mitjançant l'angiogènesi (Weinberg, 2013).

El càncer es considera un procés clonal perquè, durant la fase S del cicle cel·lular, cada cèl·lula duplica el material genètic i transmet els seus errors genètics a les cèl·lules filles (Figura 8, A). A més, aquestes cèl·lules se seleccionen favorablement perquè es divideixen més ràpidament en comparació a l'estat d'homeòstasi, on la divisió i la mort cel·lular estan clarament regulades. Aquest fet perpetua encara més notablement aquests errors genètics.

El desenvolupament d'un tumor maligne es coneix com a carcinogènesi. Hanahan i Weinberg van establir que totes les cèl·lules tumorals experimenten una sèrie de modificacions que els permeten adquirir sis capacitats fonamentals: autosuficiència en senyals de creixement, insensibilitat a les senyals inhibidores del creixement, evasió de l'apoptosi, potencial replicatiu il·limitat, angiogènesi i capacitat d'envair teixits veïns i produir metàstasi (Hanahan and Weinberg, 2000). Posteriorment, els mateixos autors van ampliar el model per incloure quatre propietats addicionals: reprogramació del metabolisme energètic, evasió de la destrucció immune, desbloqueig de la plasticitat fenotípica i evitació de la senescència cèl·lula, així com quatre característiques facilitadores: inestabilitat genòmica, inflamació promotora del tumor, reprogramació epigenètica no mutacional i variabilitat polimòrfica en microbiomes (Hanahan and Weinberg, 2011; Hanahan, 2022). Aquests catorze trets sintetitzen els atributs clau responsables del desenvolupament i la progressió de la majoria dels càncers humans (Figura 8, B).

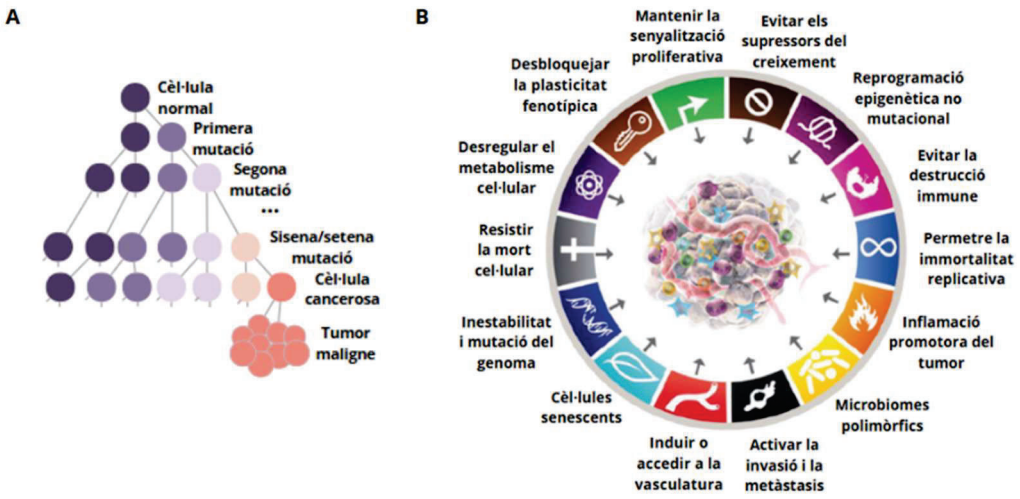


Figura 8: Procés i característiques de la carcinogènesi. A) Esquema que mostra com el càncer és un procés clonal en què les cèl·lules van adquirint diferents mutacions (unes 6-7) que les fan més proliferatives i que adquireixen canvis de morfologia fins a convertir-se en cèl·lules canceroses. B) Els catorze trets distintius del càncer en humans. Figura traduïda de (Hanahan, 2022).

Com que aquests mecanismes no estan regulats per un únic gen, són necessàries mutacions en diferents gens per acabar conferint un avantatge de creixement, fent que un clon s'expandeixi preferentment sobre les altres cèl·lules i s'acabi desenvolupant el càncer. Aquest fenomen es coneix com a cooperativitat oncogènica, on es requereixen almenys 6 o 7 mutacions somàtiques en gens associats al càncer perquè es desenvolupi un tumor maligne. Aquestes mutacions són anomenades mutacions conductores (de l'anglès *drivers*).

Aproximadament s'han identificat 300 gens *driver,* que es poden classificar en 3 grups: oncogens, gens supressors de tumors i gens reparadors del DNA (Bailey *et al.*, 2018).

Quan un proto-oncogen, gen que facilita el creixement i divisió cel·lular, adquireix una mutació de guany de funció, esdevé un oncogen, que té la capacitat d'estimular de manera descontrolada la proliferació cel·lular. Per tant, aquestes variants activen de manera constitutiva diferents rutes metabòliques que condueixen al procés de carcinogènesi. Les mutacions actitvadores en proto-oncogens tenen un efecte dominant, solen ser variants no sinònimes (missense) que es concentren en regions concretes del gen (*hotspots* en anglès).

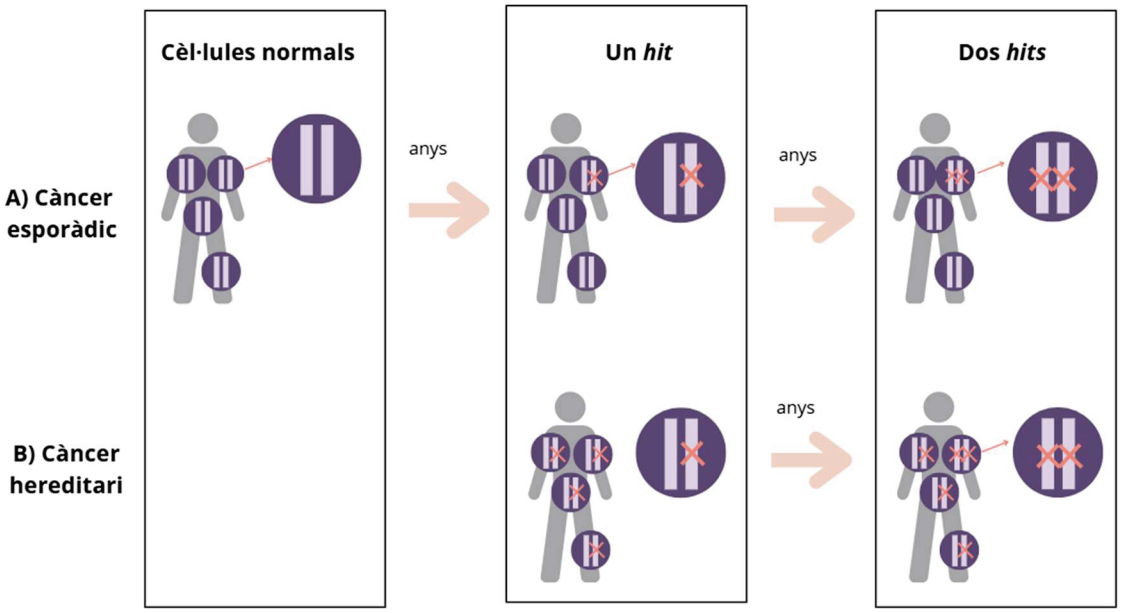


Figura 9: Representació del model dels dos *hits* de Knudson. Els cercles representen cèl·lules, els rectangles liles representen un gen (amb els dos al·lels). La creu simbolitza una variant patogènica, mentre que les fletxes de color cru representen el pas del temps. En el cas del càncer esporàdic, inicialment cap de les cèl·lules del cos presenta alteracions en el gen en qüestió. Amb el pas dels anys, diversos factors que afecten la integritat del DNA (ex: tabac, hormones, errors de replicació) poden danyar-lo i alterar un dels al·lels en una cèl·lula determinada. L'individu segueix sent sa perquè manté l'altra còpia intacta. Si amb el temps l'altra còpia d'un dels clons d'aquesta cèl·lula també adquireix un segon hit, el gen es desregula. Els individus amb predisposició genètica al càncer tenen una còpia alterada del gen en totes les cèl·lules del seu cos des del naixement, de manera que només necessiten un hit addicional per provocar la desregulació. Aquest model no aplica als gens que segueixen un patró d'herència autosòmica recessiva, els quals necessiten tenir les dues còpies alterades en totes les cèl·lules del seu cos des de naixement per conferir un risc de càncer augmentat.

Per altra banda, els gens supressors de tumors (TSG, de l'anglès *tumor-suppressor genes*) codifiquen proteïnes que tenen la funció de regular la divisió cel·lular. Les mutacions patogèniques en aquests gens tenen un efecte recessiu a nivell cel·lular, ja que es necessita una mutació de pèrdua de funció en cadascun dels al·lels per la inactiviació del TSG. Aquest model s'anomena hipòtesi de dos cops (en anglès *two-hits*) i va ser formulada per Knudson (1971) (Figura 9, A) al retinoblastoma . No obstant,

en alguns casos la pèrdua de funció de només un al·lel també comportaria una situació d'haploinsuficiència capaç de desembocar en el procés de carcinogènesi (Santarosa and Ashworth, 2004). A diferència dels oncogens, les alteracions dels TSG no es concentren en localitzacions concretes, sinó que es distribueixen al llarg del gen (Vogelstein *et al.*, 2013).

Per últim, els gens reparadors del DNA, coneguts popularment com gens guardians del genoma, són els responsables de mantenir l'estabilitat del mateix. Tal i com el nom indica, aquests gens tenen la funció de corregir errors genètics que es produeixen en el DNA de les cèl·lules. Per tant, malgrat ells no influeixen directament amb la proliferació cel·lular, són clau per evitar que prosperin altres alteracions genètiques que sí que ho fan.

El procés de carcinogènesi és complex i multifactorial, ja que està influenciat tant per factors intrínsecs del nostre DNA com per factors ambientals, és a dir, substàncies alienes que són capaces d'alterar el material genètic (Anand *et al.*, 2008). La combinació d'ambdós factors és el que facilitarà o dificultarà el procés.

2.2. Predisposició genètica al càncer

Malgrat que tots els càncers tenen un component genètic, no tots són hereditaris. De fet, el terme càncer hereditari (HC, de l'anglès *hereditary cancer*) pot portar a la confusió de que és el càncer en sí el que s'hereta, però en realitat, el que s'hereta és la predisposició a desenvolupar un càncer al llarg de la vida.

Només entre un 5 i un 10 % dels càncers i es deuen a variants germinals patogèniques heretades en gens que actuen en la reparació del DNA o gens supressors de tumors (J.E. and K., 2005; Jahn *et al.*, 2022).

La majoria dels casos de predisposició hereditària al càncer es donen en gens supressors de tumors que segueixen un patró d'herència autosòmica dominant, on els individus tenen constitucionalment un al·lel del gen alterat en totes les cèl·lules del seu cos. Així doncs, seguint la hipòtesi de Knudson (1971), només es requereix una alteració (*hit)* en alguna cèl·lula perquè el gen es desreguli (Figura 9, B). Per aquest motiu, els pacients amb predisposició hereditària al càncer solen caracteritzar-se per tenir:

- edats de diagnòstic més joves del que seria esperable;
- múltiples tumors primaris en el mateix òrgan;
- tumors bilaterals en òrgans aparellats;
- múltiples tumors en diferents òrgans en el mateix individu;
- l'acumulació de molts càncers del mateix òrgan o d'òrgans relacionats en diferents individus de la mateixa família ;
- l'aparició de tumors poc freqüents

Com que la línia germinal d'aquests pacients també presenta un al·lel alterat constitutivament, aquestes alteracions poden ser transmeses a la descendència. La majoria dels gens relacionats amb càncer hereditari tenen un patró d'herència autosòmic dominant, que representa un 50% de probabilitat d'heretar el risc en la següent generació.

Ara bé, en alguns casos, pot donar-se una condició de mosaïcisme, en què una mutació *de novo* apareix en una cèl·lula durant el desenvolupament de l'individu. Aquesta mutació es transmet només a les cèl·lules que deriven de la cèl·lula original, provocant que la variant es trobi únicament en alguns llinatges cel·lulars. Si la mutació afecta els teixits diana del gen alterat, pot generar un fenotip similar al germinal. La severitat d'aquestes condicions depèn sovint del moment en què s'ha produït la mutació i del percentatge de cèl·lules afectades, sent més greus quan la mutació es dona en estadis

inicials del desenvolupament. En aquests casos, la transmissió a la descendència només és possible si la línia germinal del progenitor conté la variant.

El fet que les persones amb predisposició hereditària al càncer desenvolupin o no la malaltia depèn, en gran part, de la penetrància del gen. La penetrància és la probabilitat que un individu amb una variant genètica específica manifesti el fenotip associat. Pot estar influenciada per factors com l'entorn, l'estil de vida i la presència d'altres variants genètiques. Es poden fer tres grups considerant la penetrància (Figura 10):

- **Al·lels d'alta penetrància**: són variants patogèniques poc comunes (freqüència de l'al·lel minoritari (MAF, de l'anglès *minor allele frequency*) <0.1-0.01% en la majoria de poblacions) o molt poc comunes (MAF <0.01%) que confereixen un risc relatiu de càncer almenys 4 vegades superior al de la població general. La majoria es troben en gens ben estudiats, pels quals solen existir guies clíniques de maneig dels pacients. En són exemples, la majoria de variants patogèniques dels gens *BRCA1* i *BRCA2,* que predisposen a la síndrome de càncer de mama i ovari hereditària (CMOH), dels gens reparadors *MLH1, MSH2* i *MSH6,* que predisposen a la síndrome de Lynch (LS, de l'anglès Lynch syndrome) i de *TP53,* que predisposen a la síndrome hereditària de càncer relacionada amb *TP53* (hTP53rc, de l'anglès *heritable TP53-related cancer syndrome*).

- **Al·lels de moderada penetrància**: són variants patogèniques rares que confereixen un risc relatiu de càncer d'entre 2 i 4 vegades superior al de la població general. Generalment es localitzen en gens relacionats recentment amb la susceptibilitat genètica al càncer, pels quals encara hi ha pocs estudis prospectius que en recolzin les guies clíniques. Per tant, el maneig dels pacients sovint representa un repte i s'acaba basant en la història familiar. Per exemple, la majoria de variants dels gens *CHEK2* i *ATM.*

- **Al·lels de baixa penetrància**: són variants sovint comunes que, considerades individualment, sumen un risc mínim de càncer. No obstant això, la combinació d'al·lels d'aquestes variants, conjuntament amb factors ambientals, pot conferir un risc considerable i agregació familiar. Per tal de quantificar el risc acumulat, aquestes variants es poden combinar en una puntuació de risc poligènic (PRS, de l'anglès *polygenic risk score*).

Cal precisar que una variant patogènica en un gen determinat pot conferir un alt risc de desenvolupar un tumor en un teixit, però, en canvi, risc moderat o baix en un altre teixit (Kamps *et al.*, 2017).

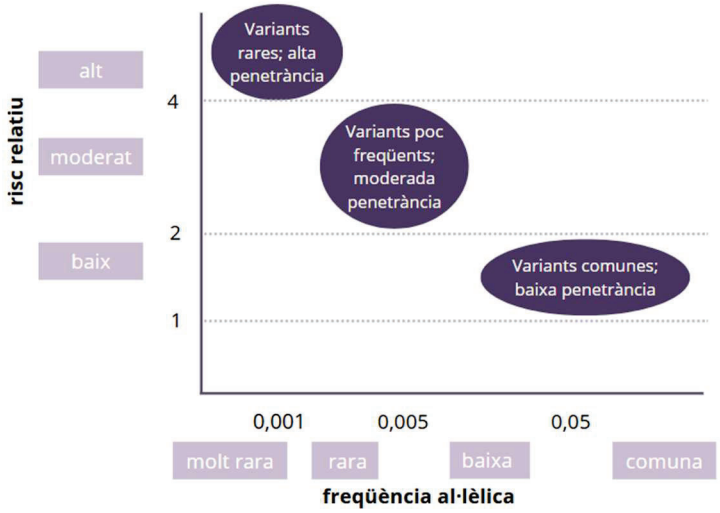


Figura 10: Distribució de les variants en funció de la freqüència al·lèlica (eix x) i el risc relatiu associat (eix y). Adaptada de Manolio et al. (2009).

3. Tècniques per detectar variants germinals

Les solucions bioinformàtiques proposades en aquesta tesi es desenvolupen a partir de les dades generades per la seqüenciació de segona generació (NGS, de l'anglès *Next Generation Sequencing*). A més, es comentaran breument altres tècniques rellevants, com la seqüenciació Sanger (precursora de l'NGS), la seqüenciació de tercera generació (TGS, de l'anglès *Third-Generation Sequencing*), l'MLPA i la LR-PCR, que sovint s'utilitzen per complementar i validar els resultats obtinguts amb l'NGS.

3.1. Seqüenciació Sanger

La seqüenciació Sanger , desenvolupada als anys 70 per Sanger i Coulson, va marcar un punt d'inflexió en el camp de la genètica. Aquest mètode va permetre la determinació ràpida i precisa de l'ordre dels nucleòtids en un fragment de DNA, consolidant-se com la tècnica de referència (Figura 11). Tot i que l'NGS ha substituït en gran part aquest mètode en projectes a gran escala, la seqüenciació Sanger continua sent àmpliament utilitzada per seqüenciar fragments curts o per comprovar mutacions concretes, ja que és una tècnica molt fiable.

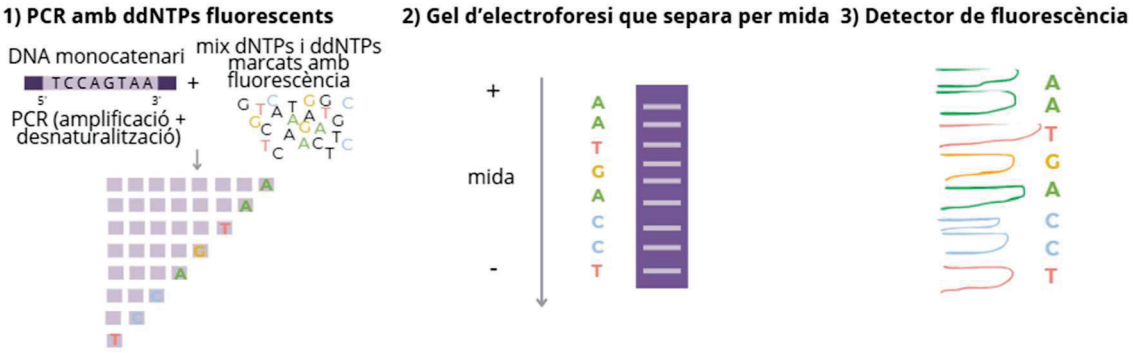


Figura 11: Representació gràfica de com funciona la seqüenciació Sanger. 1) El mètode requereix un encebador que s'aparella amb la seqüència diana i una barreja de desoxinucleòtids trifosfat (dNTPs, de l'anglès *deoxynucleotide triphosphates*) i de didesoxinucleòtids trifosfat (ddNTPs, de l'anglès *dideoxynucleotide triphosphates*). Els ddNTPs són crucials, ja que actuen com a terminadors de cadena gràcies a una modificació química que impedeix la incorporació de més nucleòtids. Actualment, aquests ddNTPs estan marcats amb fluorocroms de colors diferents per a cada base (A, T, C, G). Durant la reacció, la incorporació aleatòria de ddNTPs genera fragments de DNA de diferents mides, cadascun acabat amb un ddNTP. En mostres humanes, sovint es realitza una PCR prèvia per amplificar la regió d'interès. Aquesta etapa és necessària per assegurar que hi hagi suficient DNA per a la reacció de seqüenciació. 2) Un cop finalitzada la reacció de seqüenciació, els fragments resultants són separats per mida mitjançant electroforesi capil·lar. 3) A continuació passen a través d'un detector de fluorescència, que registra la llum emesa per cada fluorocrom dels ddNTPs. Aquest procediment permet determinar la seqüència exacta dels nucleòtids de la regió d'interès. Adaptada d' https://www.sigmaaldrich.com/ES/es/technical-documents/protocol/genomics/sequencing/sanger-sequencing.

3.2. NGS

La seqüència del genoma humà es va obtenir gràcies a la seqüenciació Sanger, però va requerir molt temps i va ser molt costosa. En total, van caldre vint anys i al voltant de 3 bilions de dòlars, ja que aquesta tècnica només permet seqüenciar un fragment de DNA a la vegada (Abdellah *et al.*, 2004).

Els avenços tecnològics i l'abaratiment dels costos de computació van ser clau pel desenvolupament de noves tècniques de seqüenciació, anomenades NGS, que permeteren analitzar diferents fragments de DNA en paral·lel.

Amb el temps, els mètodes NGS han anat millorant el rendiment i baixant els costos, fent-se accessibles per la majoria de laboratoris i essent la metodologia emprada actualment de forma general (Figura 12). Així, el cost per genoma es va situar a uns 1000 dòlars a finals del 2015, i actualment es pot trobar per uns 600 dòlars (segons dades extretes del National Institutes of Health).

A dia d'avui, existeixen múltiples plataformes de seqüenciació i, malgrat que cadascuna té les seves peculiaritats, la majoria engloba els següents passos: preparació de la llibreria, enriquiment (opcional) , seqüenciació i anàlisi bioinformàtica.

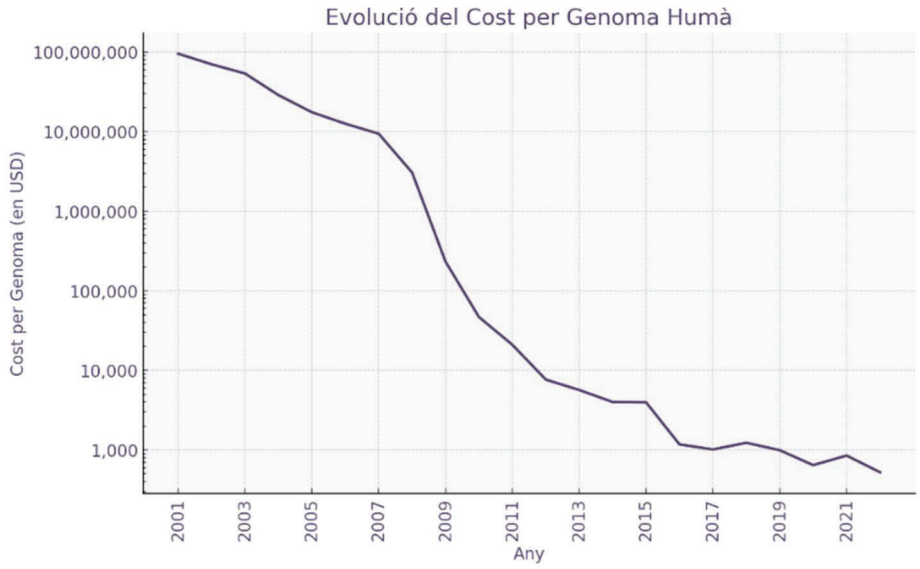


Figura 12: Evolució del cost per genoma humà en dòlars des del 2001 fins al 2022. Dades extretes de National Institutes of Health (NIH) (últim accés agost 2024, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data) .

Preparació de la llibreria

Per la majoria de processos de NGS, un cop aïllat i purificat l'àcid nucleic a seqüenciar, normalment DNA, aquest es fragmenta, ja sigui de forma mecànica o química per obtenir un material de partida adequat al protocol que es vulgui emprar (normalment 100-500 pb). Als fragments obtinguts es lliguen seqüències adaptadores als extrems, que contenen les dianes per uns encebadors així com seqüències índex per identificar la mostra. En funció de l'aplicació de l'NGS i de la quantitat de DNA inicial es pot fer una PCR per augmentar la quantitat de material de partida.

Enriquiment (opcional)

Les tècniques NGS poden analitzar tot el genoma (WGS, de l'anglès *whole genome sequencing*), limitar-se a les regions codificants o exoma (WES, de l'anglès *whole exome sequencing*) o analitzar només un conjunt de gens (panell).

En cas que no es seqüenciï el genoma cal un enriquiment dirigit a les zones d'interès (exoma o panell de gens) per assegurar que estiguin suficientment representades (Singh, 2022). Aquest pas se sol aconseguir per amplificació amb PCR o per captura amb sondes.

Seqüenciació

Els diferents fragments de DNA es seqüencien i formen les lectures. Aquestes poden obtenir-se des d'un sol extrem del fragment de DNA o des dels dos extrems (lectures aparellades). Les lectures aparellades són especialment útils per obtenir més informació dels fragments seqüenciats, ja que faciliten l'alineament al genoma de referència i la detecció de variants estructurals o de regions complexes.

Amb els anys, diferents companyies comercials han desenvolupat plataformes de seqüenciació que es distingeixen principalment pel mètode utilitzat per determinar la seqüència, la precisió d'aquesta i el cost. Avui dia, les tres principals companyies que ofereixen plataformes de seqüenciació de segona generació són Illumina, Thermo Fisher i MGI (subsidiària de BGI) . A la Taula 3 i Figura 13 es descriuen les principals característiques de les seves plataformes.

Taula 3: Principals característiques de les plataformes de seqüenciació de segona generació desenvolupades per Illumina, Thermo Fisher i MGI.

Plataforma (cases comercials)	Amplificació	Seguiment de la reacció	Avantatges	Limitacions
Illumina	Pont de PCR	Monitorització per canvi de fluorescència	Alta precisió, cobertura profunda	Preu
Thermo Fisher	Emulsió de PCR	Detecta l'alliberament d'ions d'hidrogen	Ràpida, cost moderat	Menor precisió en regions complexes
MGI	Cercle rodant (DNA nanoballs)	Síntesi combinatòria d'ancoratge de sondes fluorescents(cPAS) o anticossos marcats amb fluorocroms (coolMPS)	Baix cost, menor biaix d'amplificació	Menor disponibilitat

A les plataformes d'Illumina, la companyia capdevantera en la darrera dècada, s'utilitza una metodologia basada en la seqüenciació per síntesi de lectures curtes. En aquest mètode, els fragments de DNA s'adhereixen a un suport sòlid i s'amplifiquen en clústers per generar múltiples còpies, cadascuna actuant després com una reacció individual de seqüenciació. Durant el procés, s'utilitzen nucleòtids reversiblement bloquejats, que estan marcats amb fluorocroms i contenen un grup químic que impedeix la incorporació de més nucleòtids fins que no es retira. A cada cicle, s'incorpora un únic nucleòtid, el fluorocrom és detectat per un làser, i posteriorment, tant el marcador com el grup bloquejador són eliminats químicament per permetre la següent incorporació. Aquest procés, repetit cicle rere cicle, permet obtenir la seqüència completa de les lectures curtes generades de manera massiva i paral·lela en milions de fragments simultàniament.

Anàlisi bioinformàtica

Engloba el conjunt de passos a nivell bioinformàtic per passar de les dades crues del seqüenciador a la identificació i anotació de variants en la mostra. Aquest apartat serà abordat amb detall en el quart capítol de la tesi.

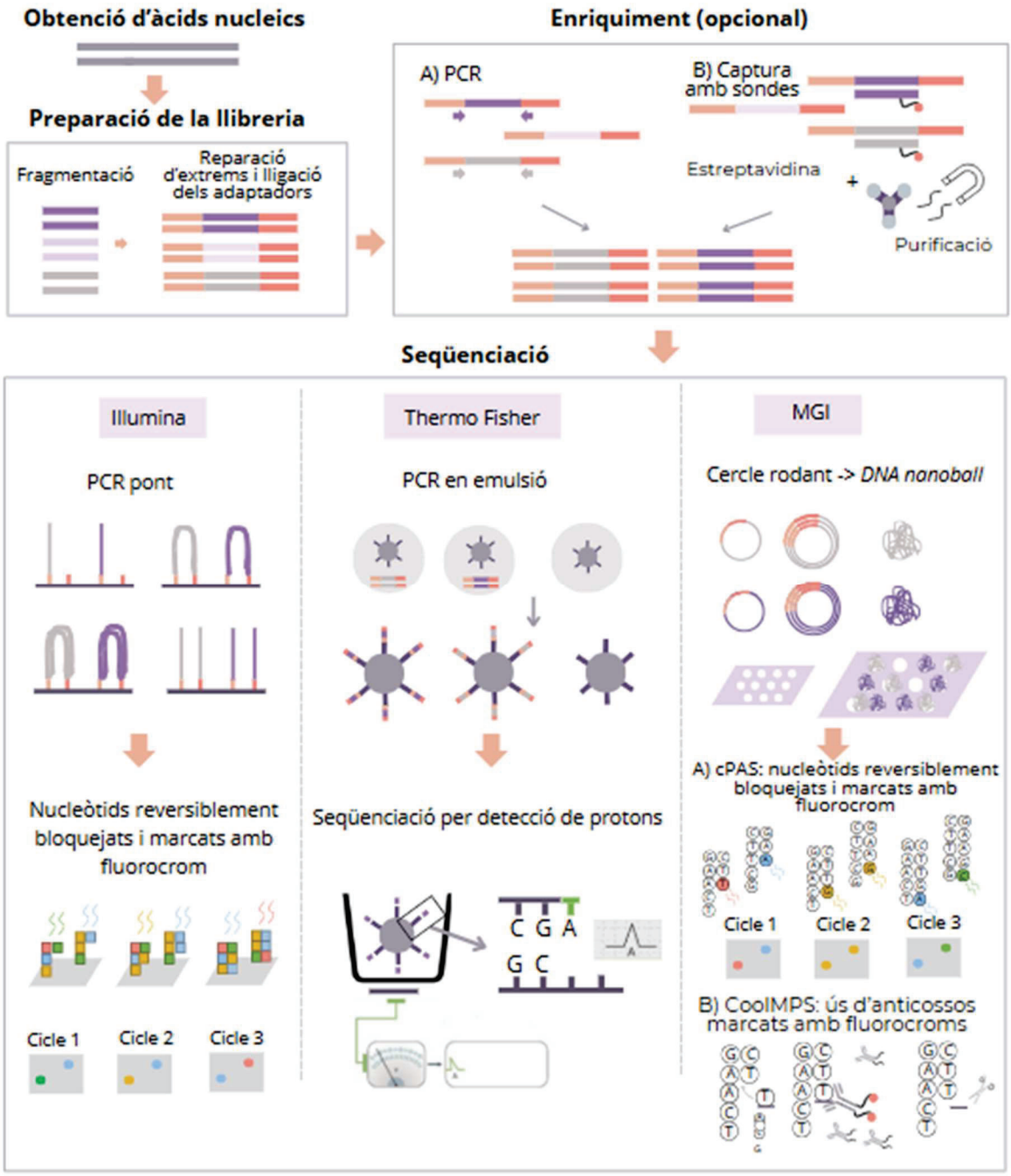


Figura 13: Esquema dels passos principals en el procés de seqüenciació NGS per a les tres principals plataformes tecnològiques: Illumina, Thermo Fisher i MGI.

3.3 Seqüenciació de tercera generació

Durant els darrers anys, s'ha estat treballant en el desenvolupament d'una nova generació de mètodes de seqüenciació, la seqüenciació de tercera generació (TGS). A diferència de les tecnologies de segona generació que produeixen lectures curtes, la TGS busca generar lectures llargues de fins a 30.000 bases (30 kb) a temps real. Les principals companyies que lideren aquest camp són Pacific Biosciences (PacBio) i Oxford Nonopore (ONT).

PacBio utilitza la tecnologia de seqüenciació en temps real de molècula única (SMRT, de l'anglès *Single Molecule Real-Time*), que permet llegir directament el DNA en temps real. Aquesta tecnologia

incorpora un procés anomenat seqüenciació de consens circular (CSS, de l'anglès Circular Consensus Sequencing), en què les molècules de DNA es circularitzen i es seqüencien repetidament per generar lectures d'alta fiabilitat (HiFi, de l'anglès *High-Fidelity*) amb una precisió superior al 99%.Tot i que actualment la seqüenciació HiFi té un cost elevat, està esdevenint cada vegada més competitiva gràcies a la reducció progressiva dels costos.

Per la seva banda, les cel·les de flux utilitzades per la tecnologia d'ONT contenen un conjunt de nanoporus a través dels quals passa el DNA o l'RNA a seqüenciar. Les bases es van identificant en temps real mitjançant la detecció de canvis en la conductància elèctrica generats durant el pas de les molècules. Aquesta plataforma permet seqüenciar DNA i RNA en estat nadiu, sense necessitat de crear llibreries, eliminant així possibles biaixos associats a la PCR. Aquesta tecnologia destaca per ser més econòmica en comparació amb PacBio, però també menys precisa.

3.4 MLPA

La tècnica d'MLPA (de l'anglès *múltiplex ligation probe amplification*) és la metodologia de referència per detectar CNVs amb resolució d'exons en el context del diagnòstic genètic. Aquesta tècnica, ràpida i fiable, permet la quantificació relativa del nombre de còpies de fins a 60 *loci* simultàniament (Figura 14).

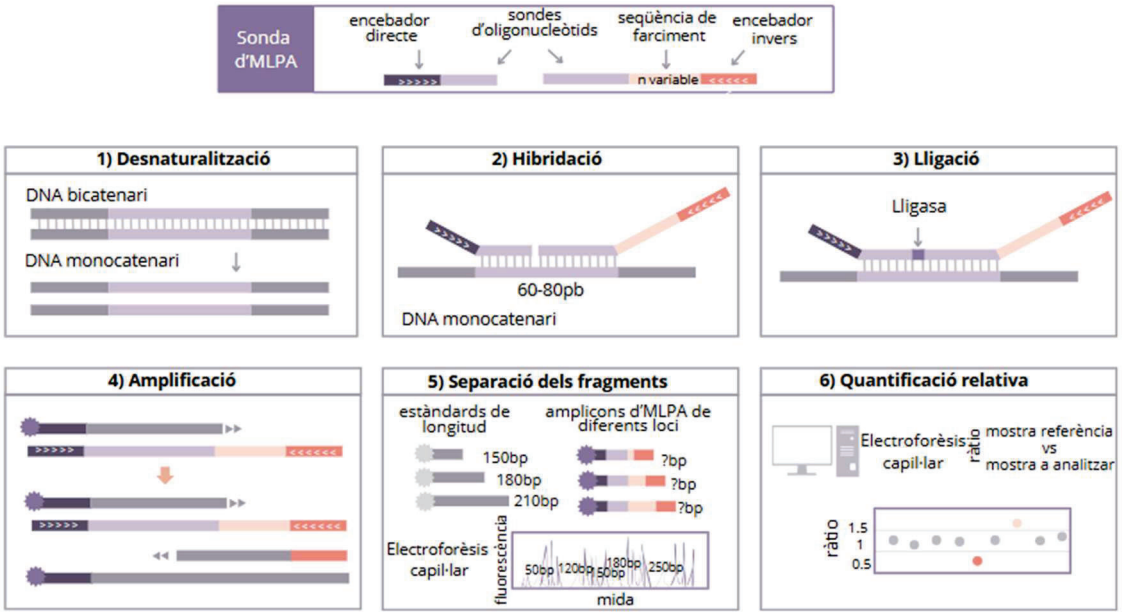


Figura 14: Passos per realitzar la tècnica d'MLPA. Les sondes d'MLPA, formades per dos fragments de DNA de cadena senzilla, hibriden a dianes (d'uns 60-80 nucleòtids) adjacents al DNA que es vol interrogar. El fragment de l'esquerra conté un encebador directe, mentre que el de la dreta inclou una seqüència de farciment i un encebador invers. La seqüència de farciment facilita l'ordenació posterior durant l'electroforesi dels fragments segons la mida. 1) El DNA es desnaturalitza. 2) Les sondes s'incuben tota la nit per hibridar sobre la mostra de DNA. 3) S'afegeix una lligasa altament específica que, en absència de desaparellaments, uneix els dos oligonucleòtids mitjançant un enllaç covalent. D'això en resulta un únic fragment d'una longitud específica. 4) El procés continua amb 35 cicles d'amplificació per PCR, on l'encebador directe està marcat amb fluorescència. 5) Electroforesi capil·lar dels productes amplificats en un seqüenciador, que permetrà la identificació de les sondes que representen cada exó segons la seva mida. 6) Les dades crues són comparades amb les d'altres mostres, idealment sense alteracions, utilitzant un programari bioinformàtic per determinar diferències relatives en el nombre de còpies. Les mostres que es comparen han de ser el més semblants possibles per no afectar el patró d'hibridació i comprometre la fiabilitat de l'anàlisi.

Tot i ser una tècnica robusta i àmpliament utilitzada, l'MLPA no és perfecta. Una de les seves limitacions és la susceptibilitat a generar falsos positius si hi ha un SNV o una altra variant a la regió d'hibridació de la sonda. Aquestes variants poden interferir amb la unió correcta de les sondes i impedir el procés de lligació, simulant una deleció.

3.5. LR-PCR

La tècnica de la PCR s'empra per amplificar el DNA de zones d'interès. Resumidament, es basa en la desnaturalització del DNA a altes temperatures, en la hibridació dels encebadors amb als extrems de la regió d'interès i l'elongació de la cadena per a una DNA polimerasa . Aquest procés es repeteix en diversos cicles gràcies als quals el DNA de la regió s'amplifica exponencialment.

Inicialment, la principal limitació d'aquesta tecnologia era la mida dels fragments que es podien amplificar, que era reduïda. No obstant això, el 1992 en modificar certes condicions de la PCR es va aconseguir amplificar fragments de fins a 5kB. Aquesta es va passar a dir PCR de llarg abast (LR-PCR, de l'anglès *long range polymerase chain reaction*). Des de llavors, s'han perseguit diferents estratègies per intentar amplificar segments més llargs i de manera més eficient (Hogrefe and Borns, 2011; Ignatov *et al.*, 2014; Zhao *et al.*, 2022).

La LR-PCR pot ser molt útil per amplificar específicament regions concretes de gens que tenen alta homologia amb altres gens o pseudogens per després procedir a la seqüenciació. En aquests casos cal buscar encebadors que anellin a regions on les dues seqüències presentin diferències.

4. Anàlisi de dades d'NGS

L'anàlisi bioinformàtica és la via que permet processar i analitzar les dades d'NGS per la seva interpretació biològica i clínica. Es divideix en tres etapes d'anàlisi: primària, secundària i terciària (Moorthie, Hall and Wright, 2013) (Figura 15).

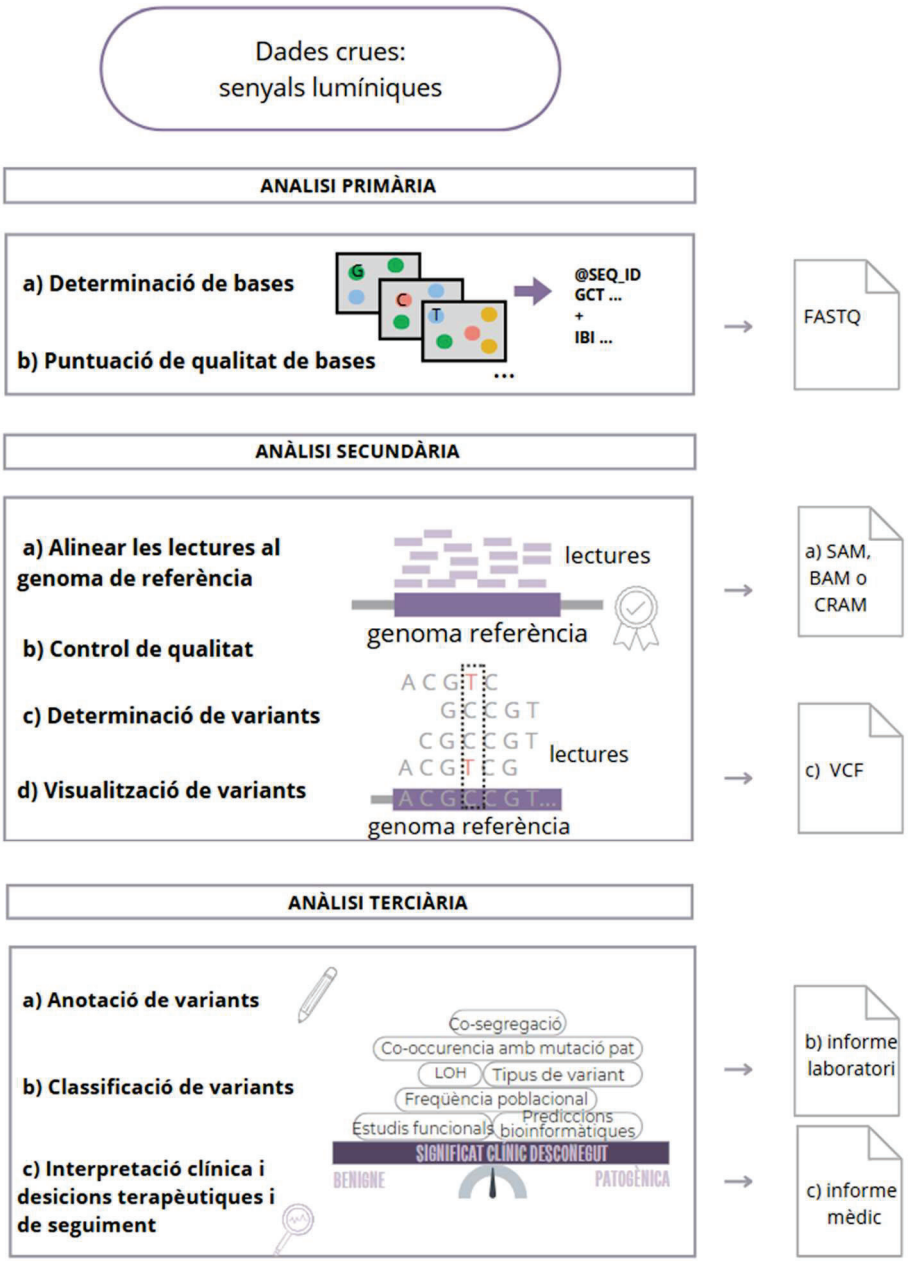


Figura 15: Representació gràfica de les tres etapes en l'anàlisi bioinformàtica amb els principals elements que engloba cadascun. Opcionalment, es poden afegir punts de control de qualitat addicionals a cada nivell, que permeten extreure diferents fitxers i mètriques per a la supervisió de l'anàlisi.

4.1. Anàlisi primària

L'anàlisi primària depèn especialment de la plataforma utilitzada. Aquesta tesi se centra en la tecnologia d'Illumina, que és amb la que s'ha generat les dades analitzades.

Determinació de bases

La tecnologia NGS d'Illumina produeix, com a dades crues, senyals lumíniques. El primer pas de l'anàlisi bioinformàtica consisteix a inferir la seqüència nucleotídica a partir d'aquestes dades. És un pas específic de la plataforma utilitzada, ja que depèn de la química emprada i la probabilitat d'error. Per tant, aquest pas el sol realitzar la mateixa plataforma. Els resultats s'emmagatzemen en un fitxer de format FASTQ, que inclou no només les lectures de la seqüenciació, sinó també la qualitat de cada base inferida. Aquesta qualitat es mesura amb el *Phred Score*, una escala logarítmica que indica la confiança en la lectura de cada base, on valors més alts corresponen a una menor probabilitat d'error. Per reduir la mida del fitxer, es codifica mitjançant caràcters ASCII.

Control de qualitat de bases

Una primera anàlisi estudia la qualitat de les bases identificades i la composició dels nucleòtids per tenir una visió de la qualitat de la carrera de NGS així com de les llibreries seqüenciades (Barbitoff et al., 2024). Existeixen programes específics que faciliten aquestes anàlisis i filtren les lectures que no superen un cert llindar de qualitat, com FASTQC i multiQC.

Addicionalment, i malgrat que s'ha vist que no té un alt impacte en la crida de variants (Barbitoff and Predeus, 2024), si l'anàlisi suggereix que en les lectures hi ha un alt contingut d'adaptadors es poden utilitzar programes que els eliminen. Alguns exemples serien FASTP, cutadapt o Trimmomatic.

4.2. Anàlisi secundària

Alineament

El següent pas consisteix en l'alineament de les lectures contingudes al fitxer FASTQ al genoma de referència, que està en format FASTA. Aquest procés es coneix com a re-seqüenciació.

Existeixen varis programes bioinformàtics per realitzar aquest pas, cadascun d'ells amb els seus avantatges i limitacions. Un dels més emprats per la comunitat bioinformàtica i d'ús gratuït és BWA-MEM (https://bio-bwa.sourceforge.net/).

Els resultats de l'alineament s'emmagatzemen en fitxers SAM, BAM o CRAM (de l'anglès *sequence alignment map, binary alignment map I compressed reference-oriented alignment map*, respectivament).

Tot i que en diagnòstic ja es disposa de la seqüència de referència, i per tant generalment es fa re-seqüenciació, si es volen identificar variants més complexes, es pot optar per la seqüenciació *de novo*. Aquest mètode implica determinar la seqüència completa de DNA sense un genoma de referència previ, sent ideal per capturar totes les variacions genètiques potencials, però sol ser computacionalment més costós.

Control de qualitat de l'alineament

Després de l'alineament, és recomanable realitzar un control de qualitat per garantir que tots els passos previs s'han completat adequadament (Barbitoff et al., 2024). Els paràmetres que s'analitzen són: la cobertura de les regions, la presència de possibles contaminacions externes i l'eliminació de duplicats. Aquest últim pas evita introduir un biaix en la crida de variants (Koboldt, 2020) i es pot fer amb programes com The Genome Analysis ToolKit (GATK) (McKenna *et al.*, 2010), Dopplemark (https://github.com/grailbio/doppelmark) or Sambamba (Tarasov *et al.*, 2015).

Addicionalment, es poden implementar altres controls, com ara recalibrar la puntuació de qualitat de base o realinear localment les zones entorn d'indels per reduir el nombre d'artefactes de seqüenciació i minimitzar així els falsos positius (Van der Auwera *et al.*, 2013). Tanmateix, aquests passos són costosos a nivell computacional i les millores que aporten són modestes (Li and Wren, 2014), motiu pel qual es consideren opcionals.

Addicionalment, existeixen eines que proporcionen mètriques útils per poder identificar problemes de creuament de mostres, com seria el recompte de variants del cromosoma X, la relació general entre variants heterozigotes i homozigotes, i l'estimació de la relació entre mostres. Un exemple n'és Peddy (Pedersen and Quinlan, 2017).

Detecció de variants

Una vegada assegurada la qualitat de l'alineament, es procedeix a identificar les variants de la mostra. Les lectures alineades es comparen amb el genoma de referència per veure en quines posicions difereixen. Les variants trobades es solen registrar en un fitxer en format VCF (de l'anglès *variant call format*), tot i que alguns programes també ofereixen la possibilitat d'emmagatzemar els resultats en un format TSV (de l'anglès *tab-separated values*).

Existeixen diferents estratègies per identificar variants, així doncs, la selecció de l'eina més adequada no és un procés trivial. L'elecció òptima dependrà de múltiples factors, com per exemple la tecnologia usada (WGS, WES o panell), la naturalesa de les mostres (teixits normals o tumorals) i el tipus de variants d'interès (SNV, indels, SVs desequilibrades o equilibrades).

Addicionalment, els programes de detecció de variants normalment permeten acotar les regions d'interès per a l'anàlisi, mitjançant l'ús d'un fitxer BED (de l'anglès *browse extensible data*) on s'especifiquen les coordenades genòmiques que es volen analitzar. Això redueix significativament el nombre de variants obtingudes.

A més, sovint els programes permeten realitzar la crida de variants tant per una única mostra com per a múltiples mostres simultàniament (mode cohort). El segon augmenta el poder estadístic en la detecció de variants, sempre que les mostres tinguin la mateixa naturalesa i hagin estat seqüenciades de manera consistent per evitar biaixos (Barbitoff et al., 2024).

SNVs i indels

Existeixen dues estratègies per cridar aquest tipus de variants: les basades en dades apilades (*pileup*) i les basades en haplotips (Barbitoff et al., 2024).

Les primeres identifiquen les variants a partir d'un fitxer de format *pileup* que conté de manera resumida la informació d'alineament de cada base nitrogenada i la seva cobertura identificant les bases que difereixen de la referència i que superen certa freqüència al·lèlica. Aquest és el mètode que hi ha darrere dels primers programes identificadors de variants que van sorgir com SAMtools (Li *et al.*, 2009) i VarScan (Koboldt *et al.*, 2009). En els últims anys, s'ha observat que aquest enfocament pot ser molt potent si es combina amb algoritmes que es basen en l'aprenentatge profund (*deep-learning* en anglès), com DeepVariant (Poplin *et al.*, 2018).

La segona estratègia consisteix en un assemblatge local de les lectures formant haplotips, amb el posterior alineament i genotipat. Alguns programes que utilitzen aquesta aproximació són: HaplotypeCaller de GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller) , FreeBayes (Garrison and Marth, 2012), Strelka2 (Kim *et al.*, 2018) i Octopus (Cooke, Wedge and Lunter, 2021). Aquí, el paper de la intel·ligència artificial no es dona en la identificació de les variants sinó en la priorització d'aquestes.

SVs

La identificació de SVs mitjançant NGS de lectures curtes és un procés complex. De fet, té una sensibilitat bastant inferior a les tecnologies que utilitzen lectures llargues, tot i que aquestes també tenen els seus inconvenients (Barbitoff et al., 2024).

Existeixen quatre estratègies per detectar variants estructurals mitjançant les lectures curtes: la profunditat de la cobertura (RD, de l'anglès *read depth*), el mapatge d'extrems aparellats (PEM, de l'anglès *paired-end mapping*), les lectures dividides (SR, de l'anglès *split reads*) i l'assemblatge (Escaramís *et al.*, 2015; Mahmoud *et al.*, 2019) (Taula 4).

Aquestes estratègies es diferencien en el tipus d'informació que exploten, però cap d'elles és capaç de detectar totes les SVs per si sola, ja que cadascuna té els seus punts forts i limitacions. Per aquest motiu, cada vegada més eines combinen els diferents mètodes per maximitzar la sensibilitat en la detecció de variants (Wang et al., 2022).

Taula 4: Resum de les diferents estratègies que es poden utilitzar per detectar SVs utilitzant NGS de lectures curtes

Resum	Estratègia			
	RD	**PEM**	**SR**	**As**
Definició	Comparació estadística de la cobertura en la regió d'interès contra l'esperada	Extrems aparellats que alineen de manera incorrecta per orientació, separació o posició	Part de la lectura alinea a una posició i l'altra part, a una altra	Es formen còntigs amb les lectures i després es comparen amb el genoma de referència
Detecten	- CNVs grans (> 1 exó)	- SVs grans (equilibrades i desequilibrades)	- SVs equilibrades i desaquilibrades	Tot tipus de variants estructurals
No detecten	- CNVs petites (< 1 exó) - SVs equilibrades	- SVs petites	- Variants amb punts de tall fora de la regió coberta	/
Punt de tall exacte	No	No, però sí aproximat	Sí	Sí
Comentari	-Cal cobertura alta	- SVs petites no es poden diferenciar de la variabilitat entre lectures -Calen vàries lectures a prop de la regió afectada	- FP d'artefactes de seqüenciació i en regions repetititves	- Mal rendiment en regions complexes del genoma -Computacional-ment costós -Calen vàries lectures a prop de la regió afectada

RD: profunditat de cobertura; PEM: mapatge d'extrems aparellats, SR: lectures dividides; As: assemblatge. FP: fals positiu.

Profunditat de la cobertura (RD)

Aquesta estratègia, que pot ser utilitzada tant per lectures úniques com aparellades, es basa en mètodes estadístics que determinen si en una regió o finestra concreta del genoma la profunditat de cobertura és l'esperada o, per contra, difereix (Figura 16). Aquesta diferència podria indicar la presència d'una SV desequilibrada (Escaramís *et al.*, 2015). Per tant, aquest mètode és útil per detectar CNVs, però no SVs equilibrades, on la profunditat de la cobertura no es veu alterada.

A més, presenta altres limitacions. D'una banda, el nivell de resolució no sobrepassa l'exó, de manera que no és un bon mètode per detectar SVs petites. D'altra banda, no permet determinar el punt de tall exacte, i el rendiment pot veure's compromès per biaixos en la cobertura, provocats per mala qualitat del DNA, efectes de grup de mostres (en anglès *batch effects*), zones repetitives, la composició de bases nitrogenades, l'amplificació de les zones enriquides (si n'és el cas) i per la mida de la finestra escollida (Escaramís *et al.*, 2015).

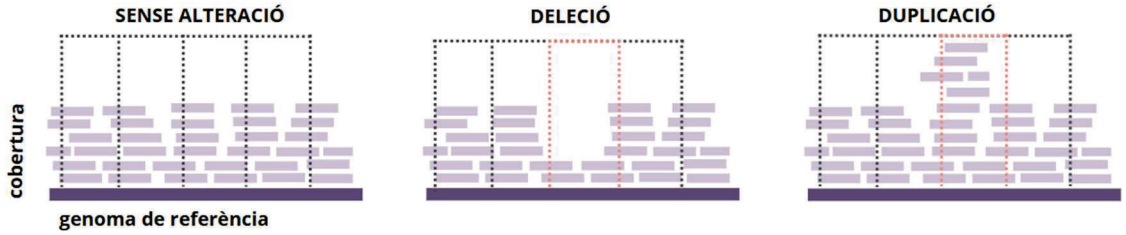


Figura 16: Representació gràfica de l'estratègia de RD. Els rectangles de color lila clar representen les lectures alineades al genoma de referència. Les línies de punts representen les finestres en que es divideix el genoma. Es pot observar com el patró de cobertura cau quan hi ha una deleció i augmenta quan hi ha una duplicació en comparació a les mostres sense alteració.

Mapeig d'extrems aparellats (PEM)

Per l'estratègia PEM es necessiten lectures aparellades. Aquest mètode es basa en la hipòtesi que la distància de dues lectures aparellades alineades al genoma de referència ha de tenir una distribució específica. Per tant, intenta identificar clústers de lectures aparellades que no segueixin la distribució, indicant la presència de SVs. Les discordances poden manifestar-se també en l'orientació entre les lectures, en l'ordre o fins i tot perquè estiguin en cromosomes diferents. En funció del patró d'aquestes discordances, es pot suggerir un tipus concret de SV, tal com s'observa en la Figura 17.

La resolució dels punts de trencament en aquest enfocament dependrà de la mitjana i la desviació estàndard de la mida d'inserció de la llibreria, així com de la cobertura (Escaramís *et al.*, 2015). Tot i que els punts de tall es poden estimar perquè envolten les dues lectures, no es poden determinar amb precisió.

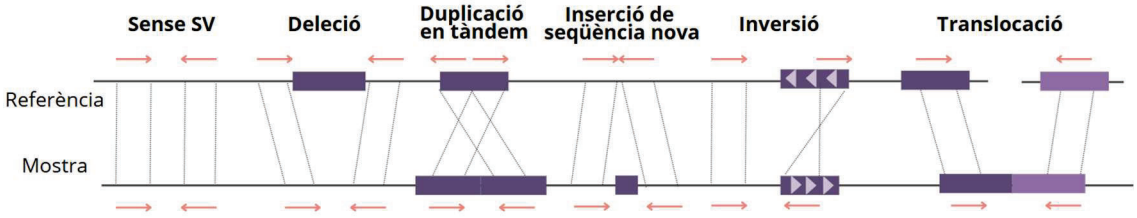


Figura 17: Possibles patrons de les lectures aparellades quan hi ha una SV. Deleció: les lectures aparellades es veuen més separades del que s'espera segons la llibreria. Duplicació en tàndem: les lectures aparellades apareixen en diferent ordre. Inversió: les lectures s'alineen al genoma de referència en el mateix sentit. Insercions de seqüència nova: les lectures aparellades es veuen més juntes del que s'espera. Translocació: una lectura alinea a una regió i l'altra a una de diferent. Adaptat d'Escaramís et al. (2015).

Aquest enfocament presenta alguns inconvenients. Les regions repetitives representen un repte, ja que poden dificultar l'alineació precisa de les lectures aparellades. A més, la presència de SNPs (de l'anglès *single nucleotide polymorphisms*) o altres elements de seqüència propers als punts de trencament pot interferir amb la detecció. Un altre límit important és la identificació d'insercions que superin la mida mitjana d'inserció de la llibreria. Tampoc és efectiva per identificar insercions i delecions petites, perquè les distàncies entre lectures aparellades poden estar dins del rang de variabilitat natural i no mostrar cap patró discordant clar(Rodríguez-Santiago and Armengol, 2012; Escaramís *et al.*, 2015).

Lectures divididas (SR)

Aquesta estratègia es centra en l'anàlisi de la seqüència, on les SVs es detecten quan una part de la lectura s'alinea correctament amb la regió d'interès, mentre que l'altra part queda mal alineada. Aquestes bases mal alineades, conegudes com a *soft-clipped bases*, formen una seqüència que pot alinear a una altra regió del genoma en funció del tipus de SV (Figura 18). Aquest patró permet identificar el punt de trencament de la SV, ja que revela el lloc exacte on la seqüència divergeix del genoma de referència. Així doncs, s'assoleix una resolució a nivell de nucleòtid, la qual no és aconseguida per cap de les estratègies prèviament descrites.

Aquesta estratègia guanya precisió si s'utilitza amb lectures aparellades, perquè la lectura parella pot aportar informació addicional que confirma o complementa la posició del trencament, millorant la fiabilitat en la identificació de la variant.

Tanmateix, la principal limitació de l'estratègia de SR és que les lectures produïdes per NGS són curtes, fet que restringeix l'alineament a una part encara més curta de la lectura. Això provoca una baixa cobertura al voltant del punt de trencament, fent poc fiable la identificació de la variant.

A més, els errors d'alineament poden generar falsos positius i negatius. Per tant, els programes de detecció de variants basats en SR pateixen especialment en la detecció de variants en regions genòmiques ambigües, on els alineadors mostren dificultats, com són les zones repetitives (Wang et al., 2022).

Aquest mètode és especialment útil en WGS, ja que es té accés a tota la seqüència. Mentre que quan s'utilitza per WES o en panells dirigits, el factor limitant és que el punt de trencament ha de trobar-se justament dins de la regió coberta per poder ser detectat.

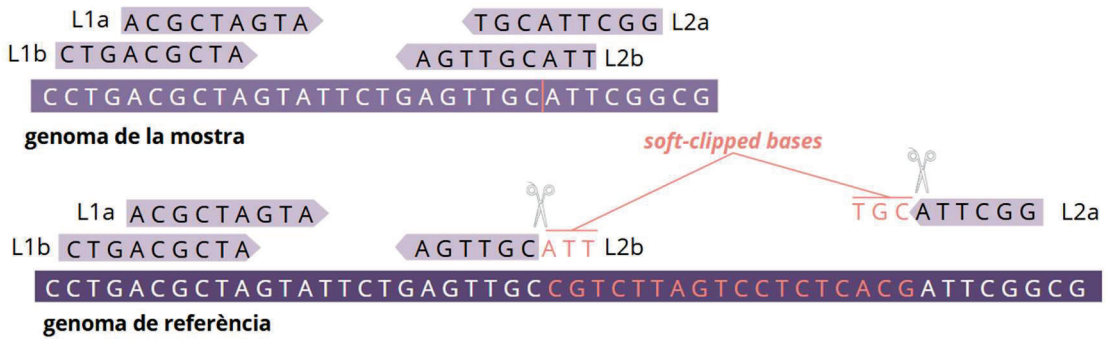


Figura 18: Representació de les soft-clipped bases en un cas de deleció. Quan les lectures del genoma de la mostra s'alineen al genoma de referència, es fragmenten ja que part de la lectura alinea a una regió i part a una altra, patint el fenomen de *soft clipping*. El mètode de SR detecta aquestes bases per identificar els punts de tall exactes, representats a la imatge amb les tisores. L1: lectura 1; L2: lectura 2, ambdues aparellades.

Assemblatge

Per detectar noves variants mitjançant assemblatge de seqüències, s'utilitza l'assemblatge *de novo.* Aquesta estratègia es basa en l'agrupació de les lectures en còntigs sense utilitzar el genoma de referència com a motlle, construint la seqüència peça per peça. De fet, és la manera en que es construeix el propi genoma de referència. Posteriorment, aquests còntigs sí que són comparats amb el genoma de referència o un altre assemblatge per descobrir SVs (Figura 19) (Mahmoud *et al.*, 2019).

Malgrat que en teoria aquest mètode és capaç de detectar tota mena d'SVs, s'ha vist que no té un bon rendiment en regions complexes del genoma, especialment si s'utilitzen lectures curtes. A més, és molt costós a nivell computacional, fet pel qual no sol ser massa utilitzat quan es persegueix una fi diagnòstica. Un altre problema dels mètodes d'assemblatge és que no poden gestionar seqüències haplotípiques, de manera que només es poden detectar variacions estructurals homozigotes (Xi et al., 2012).

D'altra banda, cal esmentar una altra aproximació d'assemblatge, el reassemblatge local, tot i que no està dissenyada específicament per detectar noves variants estructurals, sinó per millorar la precisió en la identificació de variants en àrees complexes del genoma. Aquesta tècnica s'utilitza en regions ja candidates a tenir una SV. N'extreu les lectures que cobreixen la regió d'interès i les realinea de manera local amb el genoma de referència. Això permet corregir possibles errors d'alineament inicials, i en conseqüència, identificar de manera més precisa les variants i determinar-ne el genotip associat.

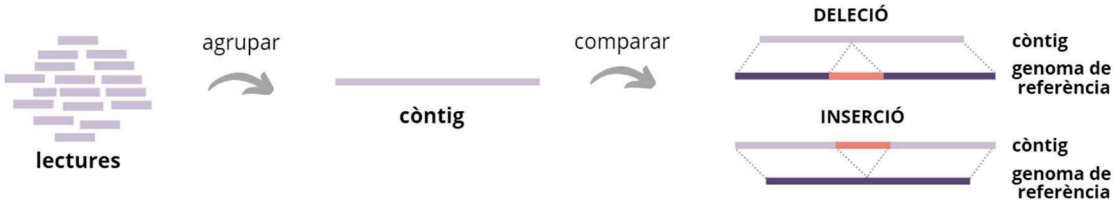


Figura 19: Esquema del funcionament del mètode d'assemblatge. Les lectures s'agrupen en còntigs, que posteriorment es comparen amb el genoma de referència.

A la Taula suplementària 1 (annex) es pot trobar una llista de diferents eines dissenyades per detectar SVs i l'estratègia que segueixen.

4.3. Anàlisi terciària

El primer pas és anotar les variants identificades amb tota la informació rellevant disponible. Es poden fer servir programes com ANNOVAR (Wang, Li and Hakonarson, 2010) o VEP d'Ensembl (McLaren et al., 2016).

Les variants anotades en les zones d'interès han de ser revisades i interpretades per experts, amb l'objectiu d'avaluar la seva rellevància biològica i clínica.

Per poder classificar les variants, cal recollir informació de diferent naturalesa, com ara la freqüència en bases de dades poblacionals, prediccions *in silico* i estudis funcionals, d'entre d'altres. Aquest apartat serà aprofundit en el capítol 7 d'aquesta tesi doctoral.

Finalment, els clínics utilitzen aquesta informació per elaborar un informe amb les recomanacions personalitzades per al o la pacient.

5. Funcionament del Programa de Càncer Hereditari a l'ICO - IDIBELL

El Programa de Càncer Hereditari està integrat per un grup multidisciplinari de professionals de l'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL) i l'Institut Català d'Oncologia (ICO). Es centra en la millora de l'assistència sanitària dels pacients i famílies amb predisposició al càncer, tant en una vessant clínica i diagnòstica com de recerca (Figura 20).

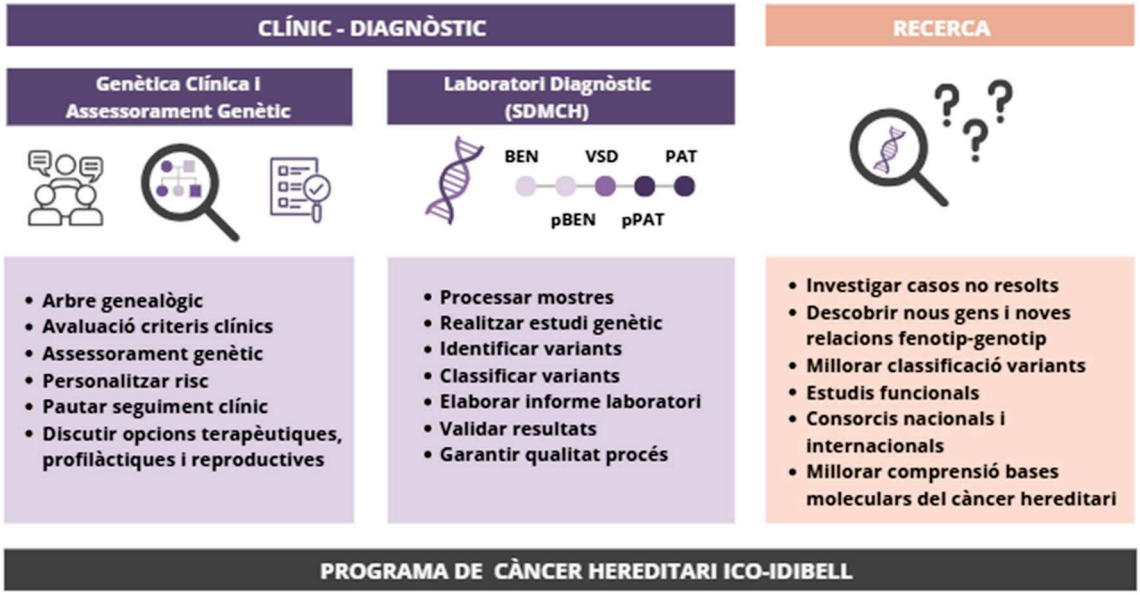


Figura 20: Organització del Programa de càncer hereditari ICO-IDIBELL. Es detallen les principals tasques de cadascuna de les àrees que formen part del programa. SDMCH: Servei de Diagnòstic Molecular de Càncer Hereditari.

A nivell assistencial, el programa busca identificar els pacients que compleixen els criteris de predisposició al càncer hereditari, per tal de fer un càlcul personalitzat del seu risc, implementar mesures de seguiment i de detecció precoç, així com oferir opcions preventives i terapèutiques personalitzades. Les dues activitats principals en aquesta àrea són l'assessorament genètic i el diagnòstic molecular. El programa consta de diverses unitats d'assessorament genètic, que atenen més de la meitat de la població adulta de Catalunya. Així doncs, quan un pacient visitat a la unitat d'assessorament genètic compleix criteris de sospita d'alguna síndrome de càncer hereditari, se li ofereix la possibilitat de fer un estudi genètic. A Catalunya, per tal de garantir l'equitat assistencial, es va establir al 2019 un consens català de criteris clínics per realitzar un estudi genètic així com dels gens a analitzar segons els criteris. Aquest consens s'ha recollit en la Instrucció 03/2021 del CatSalut (https://scientiasalut.gencat.cat/bitstream/handle/11351/8438.3/determinacions_perfil_genetic_sindromes_hereditaries_cancer_adult_pediatria_2023.pdf?sequence=4&isAllowed=y). Aquests criteris i gens s'actualitzen anualment.

L'ICO compta amb el Servei de Diagnòstic Molecular de Càncer Hereditari (SDMCH), situat a l'Hospitalet per centralitzar la realització dels estudis genètics de tots els pacients ICO que compleixin criteris. Aquest servei, amb més de 25 anys d'experiència, realitza estudis genètics segons la seva cartera de serveis per a la identificació i interpretació de variants germinals en gens que predisposen al càncer hereditari.

Des de 2017 els estudis dels casos índex s'estudien per NGS utilitzant un panell de gens dissenyat *adhoc* per l'equip, anomenat I2HCP (panell ICO-IMPPC) i un algoritme d'anàlisi de dades que utilitza VarScan per a la crida de SNVs i petites indels (Castellanos *et al.*, 2017). Amb els anys s'han fet actualitzacions del panell. D'una banda, s'han inclòs nous gens relacionats amb síndromes de càncer hereditari, així com regions no codificants d'interès, com les 5' i 3'UTR o part del promotor (per alguns gens). La versió actual (v3) compta amb 165 gens, 24 dels quals tenen la majoria de les regions UTR cobertes (Taula 5).

Taula 5: Llista del 165 gens inclosos a I2HCP (v3)							
A2ML1	*CDKN1B*	*ERCC4*	*GNAS*	*MLH3*	*POLH*	*RECQL5*	*SPRED1*
AIP	*CDKN1C*	*ERCC5*	*GOT2*	*MN1*	*POT1*	*RET*	*STK11*
AKT1	*CDKN2A*	*ERCC6*	*GPC3*	*GREM1*	*PPM1D*	*RIT1*	*SUFU*
ALK	*CDKN2B*	*ERCC8*	*GRB2*	*MRE11A*	*PPP1CB*	*RNASEL*	*TDP2*
APC	*CDKN2C*	*EXO1*	*HOXB13*	*MSH2*	*PRKAR1A*	*RNF43*	*TGFBR2*
ARAF	*CHEK2*	*EXT1*	*HRAS*	*MSH3*	*PRKN*	*RRAS*	*TMEM127*
ATM	*CTNNA1*	*EXT2*	*KIF1B*	*MSH6*	*PTCH1*	*SBDS*	*TP53*
AXIN2	*CYLD*	*FAN1*	*KIT*	*MUTYH*	*PTEN*	*SDHA*	*TPMT*
BAP1	*DDB1*	*FANCA*	*KLLN*	*NBN*	*PTPN11*	*SDHAF2*	*TSC1*
BARD1	*DDB2*	*FANCB*	*KRAS*	*NF1*	*RAD50*	*SDHB*	*TSC2*
BLM	*DICER1*	*FANCC*	*LZTR1*	*NF2*	*RAD51*	*SDHC*	*TSHR*
BMPR1A	*DNMT3A*	*FANCD2*	*MAD2L2*	*NRAS*	*RAD51B*	*SDHD*	*UBE2T*
BRAF	*EDC4*	*FANCE*	*MAP2K1*	*NTHL1*	*RAD51C*	*SHOC2*	*VHL*
BRCA1	*EGFR*	*FANCF*	*MAP2K2*	*PALB2*	*RAD51D*	*SLX4*	*WRN*
BRCA2	*EGLN1*	*FANCG*	*MAX*	*PDGFB*	*RAF1*	*SMAD4*	*WT1*
BRIP1	*EGLN2*	*FANCI*	*MCPH1*	*PDGFRA*	*RASA1*	*SMARCA4*	*XPA*
BUB1B	*ELAC2*	*FANCL*	*MDH2*	*PHOX2B*	*RASA2*	*SMARCB1*	*XPC*
CBL	*EPCAM*	*FANCM*	*MEN1*	*PIK3CA*	*RB1*	*SMARCE1*	*XRCC2*
CDC73	*ERBB2*	*FH*	*MET*	*PMS2*	*RBBP8*	*SNAI2*	
CDH1	*ERCC2*	*FLCN*	*MITF*	*POLD1*	*RECQL*	*SOS1*	
CDK4	*ERCC3*	*GNAQ*	*MLH1*	*POLE*	*RECQL4*	*SOS2*	

En gris clar s'assenyalen els gens que tenen la regió 5'UTR coberta. Concretament, es cobreix 500 pb abans de l'ATG i 200 pb després del codó terminació, a excepció dels gens *BRCA1, BRCA2* i *MLH1* que es cobreix 1000 pb abans de l'ATG. La resta de gens només es cobreix unes 150 pb de la regió UTR.

L'estudi del panell inclou la utilització del programa DECoN com a eina de cribratge de CNVs, utilitzant paràmetres optimitzats per maximitzar la sensibilitat (Moreno-Cabrera et al., 2020). Només es realitzen estudis de MLPA per confirmar les troballes detectades per DECoN, fet que ha reduït costos i augmentat l'eficiència i rendiment diagnòstic (Moreno-Cabrera *et al.*, 2022).

Malgrat l'alineament i la crida de variants es fa de tots els gens del panell, a nivell assistencial només s'analitza un panell de gens associat al fenotip del pacient/família. Només s'examinen aquelles variants en gens clínicament accionables i relacionats amb el fenotip del pacient i la seva família (Taula 6. En són una excepció els *gens BRCA1/2, MLH1, MSH2 i MSH6* que són analitzats en totes les mostres de manera oportunista a causa de la seva alta rellevància clínica (Feliubadaló *et al.*, 2019).

Quan es detecta una variant patogènica o probablement patogènica en el proband, es poden iniciar estudis en els familiars de primer grau per determinar si són portadors de la mateixa alteració genètica. Aquests estudis de portadors, que es realitzen en cascada, es limiten exclusivament a confirmar la presència o absència de la variant identificada.

A nivell de recerca, el programa té diferents línies que van des de la millora del diagnòstic genètic del càncer hereditari (on s'emmarca aquesta tesi doctoral), a obtenir una millor comprensió de les bases moleculars del càncer hereditari, amb un interès particular en el càncer gastrointestinal, el càncer de mama i ovari, així com altres tumors rars.

Taula 6: Gens de la Instrucció del CatSalut analitzats en funció de fenotip

Gens	Fenotips														
	CMOH	OV	CRC	GAS	PAN	MEL	REN	PR	CMT	PPGL	HP	TS	GIST	AC	TP53
ACD						■									
AIP											○				
APC			●									●		●	
ATM	●			●	●			●							
BAP1						●	●								
BARD1	●														
BMPR1A			●	●											
BRCA1															
BRCA2															
BRIP1	●	●													
CaSR											□				
CDC73											●				
CDH1	○			●											
CDK4						●									
CDKN1B											●				
CDKN2A					●	●									
CHEK2	●							●							
CTNNA1				●											
DICER1												●			
EPCAM			●												
FH							●			●				●	
FLCN							●								
HOXB13								●							
KIF18										●					
KIT													●		
MAX										●					
MEN1										●					
MEN1											●				
MET							●								
MITF						●									
MLH1															
MSH2															
MSH6															
MUTYH			●												
NF1										○					
NTHL1			●												
PALB2	●	●		●	●			●							
PMS2			○												
POLD1			●												
POLE			●												
POT1						●									
PRKAR1A											○	●			
PRSS1					○										
PTEN	●		●	●			●					●			

RAD51C	●	●															
RAD51D	●	●															
RET										●	●						
RNF43			○														
SDHA							●				●				●		
SDHAF2											●						
SDHB							●				●				●		
SDHC							●				●				●		
SDHD							●				●				●		
SMAD4			●	●													
STK11			●	●	●												
TERF2IP						■											
TERT						■											
TMEM127											●						
TP53	○		○	●												●	●
TSC1							○										
TSC2							○										
VHL							●				●						

Els gens oportunistes estan senyalats amb bandes de color lila clar. Les figures pintades de color lila representen els gens que s'han d'analitzar sempre amb aquell fenotip, mentre que les pintades de blanc indiquen els gens que només s'analitzen si el proband compleix unes condicions fenotípiques addicionals. Els cercles representen gens inclosos en el panell I2HCP v3, mentre que els quadrats indiquen gens que encara no estan inclosos en el panell, però que s'inclouran a la propera versió, per cumplir amb la instrucció de CatSalut. Fenotips: CMOH: càncer de mama i ovari hereditari; OV: càncer d'ovari; CRC: càncer colorectal i d'endometri i poliposi; GAS: càncer gàstric; PAN: càncer de pàncrees; MEL: melanoma; REN: càncer renal; PR: càncer de pròstata; CMT: carcinoma medul·lar de tiroide; PPGL: feocromocitoma/paraganglioma; HP: hiperparatiroïdisme primari, o manifestacions de MEN1/MEN4; TS: càncer de tiroide sindròmic; GIST: tumor de l'estroma gastrointestinal; AC: tumor adrenocortical; TP53: síndromes de càncer hereditari relacionats amb TP53. Gens amb anàlisi parcial: CDK4 (exó 2); HOXB13 (variant G84E); MITF (variant E318K); POLD1 (exons 6-13); POLE (exons 7-14), TERT (promotor).

6. Reptes en la identificació de variants utilitzant NGS de lectures curtes

Malgrat els múltiples avantatges que comporta la NGS de lectures curtes, aquesta tècnica també presenta algunes limitacions significatives en la detecció de certs tipus de variants. Principalment, presenta limitacions en la identificació de CNVs petites, reordenaments complexes, indels grans, zones repetitives de baixa complexitat, pseudogens i variants en mosaic (Figura 21) (Lincoln et al., 2021; Barbitoff et al., 2024).

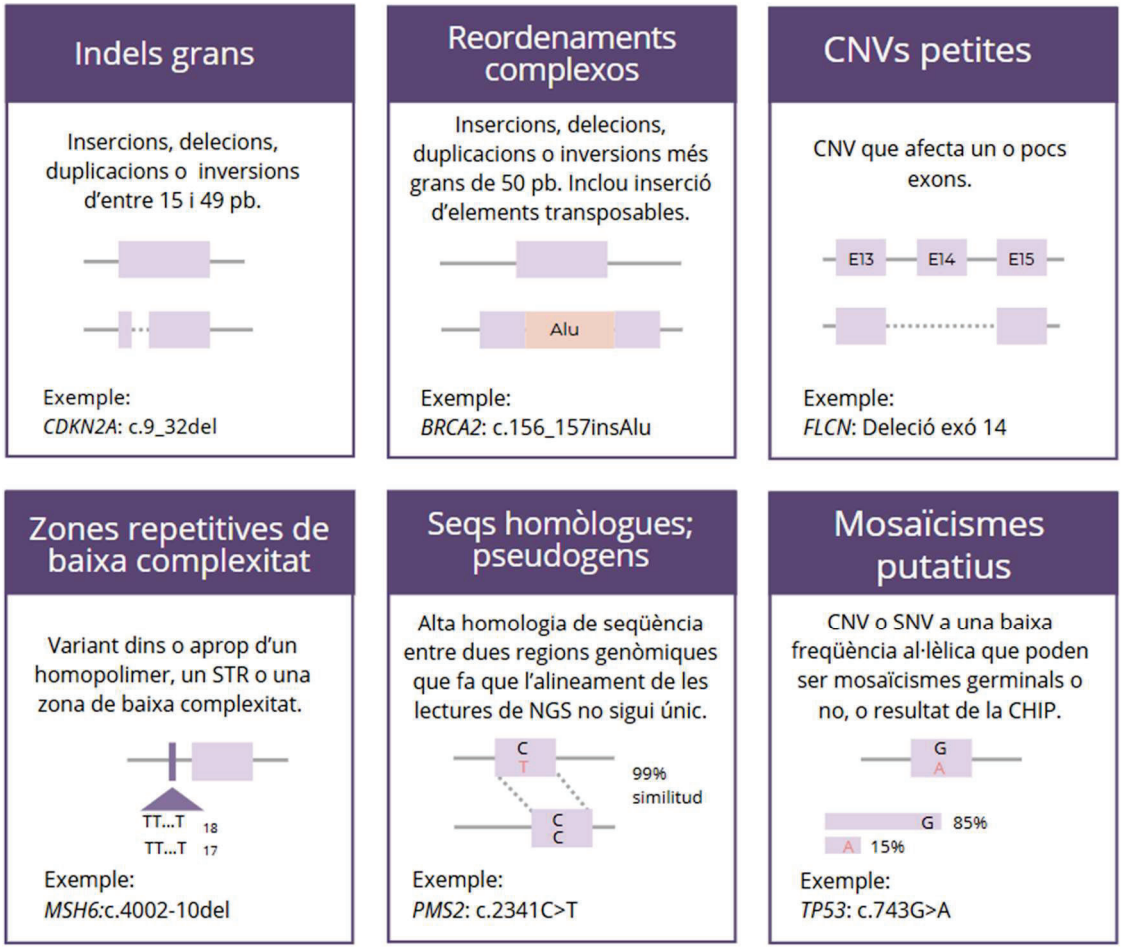


Figura 21: Exemples de variants tècnicament complexes de detectar per NGS. Seqs: seqüències; STR: repeticions curtes en tàndem (de l'anglès short tandem repeats);CHIP: hematopoesis clonal de potencial indeterminat. Adaptat de Lincoln et al. (2021).

De fet, un estudi realitzat en 471.591 pacients que complien criteris clínics per a estudi genètic va determinar que aproximadament 1 de cada 7 variants patogèniques (14%) pertany a aquestes categories difícils de detectar per NGS (Lincoln et al., 2021).

Avui dia, cap algoritme d'identificació de variants pot capturar tota aquesta heterogeneïtat. Per a la detecció d'aquestes variants, cal utilitzar diferents aproximacions bioinformàtiques, cosa que requereix expertesa i temps. Per aquest motiu, la majoria dels laboratoris de diagnòstic molecular no les aborden totes.

En els següents apartats s'aprofundeix en la problemàtica que suposen aquestes variants.

Seqüències homòlogues, pseudogens

Són àrees del genoma amb una alta homologia de seqüència entre elles. Les lectures procedents d'aquestes regions són idèntiques o pràcticament idèntiques. Això fa que sovint no s'alineïn en un únic lloc del genoma i poden arribar a ser descartades per filtres de qualitat. En altres casos, s'alineen en un lloc o l'altre indistintament. Per tant, si un programa detecta una variant en aquestes regions, no es pot confirmar directament la localització d'aquesta. D'altra banda, si no es detecta la variant a nivell bioinformàtic, no es pot assegurar que no hi sigui, ja que podria estar alineada incorrectament en una regió homòloga.

Sovint hi ha bases on les dues regions homòlogues difereixen, anomenades variants paràlogues de seqüència (PSVs, de l'anglès *paralogous sequence variants*). No obstant això, sovint no es poden utilitzar per a diferenciar les regions perquè la recombinació genètica sol ser elevada entre ambdós llocs (Chen *et al.*, 2020).

Un exemple clar d'alta homologia són els gens paràlegs, que es donen per una duplicació de la seqüència. Com que no estan sotmesos a la mateixa pressió selectiva, una de les còpies sovint acumula mutacions i pot acabar adquirint noves funcions o perdre la seva funció original i esdevenir un pseudogèn.

Existeixen diversos gens inclosos en el panell de diagnòstic de l'ICO (I2HCP_v3) que presenten pseudogens (Taula 7).

Taula 7: Llistat de gens relacionats amb HC i inclosos al panell de diagnòstic de l'ICO I2HCP v3 que tenen pseudogens segons GeneReviews

Gen	Pseudogèn	Malaltia causada per variants (probablement) patogèniques al gen
BMPR1A	*BMPR1APS1*	Poliposi juvenil
	BMPR1APS2	
CHEK2	5 pseudogens	Predisposició hereditària a càncer de mama
		Predisposició hereditària a càncer de pròstata
FANCD2	*FANCD2P1*	Anèmia de Fanconi
NF1	11 pseudogens	Neurofibromacotosis tipus 1
PIK3CA	*LOC100422375*	Espectre de sobrecreixement relacionat amb PIK3CA
PMS2	*PMS2CL* i ≥ 13 pseudogens	Síndrome de Lynch
		Deficiència constitucional de la reparació de desajustos (CMMRD)
PRSS1	*PRSS3P1*	Pancreatitis hereditària
	PRSS3P2	
PTEN	*PTENP1*	Síndrome de Cowden
SDHA	4 pseudogens	Paraganglioma-feocromocitoma hereditari
SDHC	5 pseudogens	Paraganglioma-feocromocitoma hereditari
SDHD	7 pseudogens	Paraganglioma-feocromocitoma hereditari
SMAD4	1 pseudogèn	Poliposis juvenil
		Telangièctasia hemorràgica hereditària

Extret: https://www.ncbi.nlm.nih.gov/books/NBK535152/table/resources_Table3.T.genes_with_highly_hom , data d'últim accés agost 2024.

És especialment destacable el cas del gen *PMS2,* que presenta una complexitat considerable en l'anàlisi degut a l'existència de múltiples pseudogens (Nicolaides *et al.*, 1995; Nakagawa *et al.*, 2004). Catorze d'ells comparteixen homologia amb els exons 1-5 i un quinzè, anomenat *PMS2CL*,

comparteix més del 98% d'identitat amb els exons 9 i 11-15. *PMS2CL* és una duplicació parcial invertida localitzada al mateix cromosoma 7, els exons del qual es diferencien únicament per 28 PSVs. A més, tenen una considerable taxa de recombinació genètica que dificulta l'ús d'aquestes posicions per a distingir-los (Hayward *et al.*, 2007; Ganster *et al.*, 2010).

Identificar correctament les variants que corresponen al gen *PMS2* és especialment rellevant, ja que les variants patogèniques monoal·lèliques causen LS, mentre que les bial·lèliques causen deficiència constitucional de la reparació de desajustos (CMMRD, de l'anglès *constitutional missmatch repair deficiency*) (Tabori et al., 2017; Win et al., 2017).

Variants que afecten el nombre de còpies

Com s'ha comentat, existeixen múltiples eines per detectar SVs a partir d'NGS. Moltes d'elles estan específicament dissenyades per detectar CNVs i es basen, principalment, en l'estratègia de profunditat de cobertura. L'elecció de l'eina òptima no és un procés trivial, ja que el seu rendiment depèn, en gran manera, del tipus de tècnica i de les dades utilitzades (Barbitoff et al., 2024).

En aquest context, els articles que avaluen el rendiment de diverses eines són extremadament útils per determinar quina funciona millor i sota quines condicions. Ara bé, sovint moltes d'aquestes avaluacions són realitzades pels mateixos autors de les eines, la qual cosa pot introduir biaixos importants. Per exemple, sovint fan la comparació amb un nombre limitat d'eines i utilitzen només un conjunt de dades.

Fins ara, sota el nostre coneixement, només hi ha tres avaluacions d'eines dedicades a la identificació de CNVs en dades de panell i no realitzades pels mateixos autors. Tanmateix, aquestes també presenten les seves pròpies limitacions. Roca et al., (2019) avaluen les eines utilitzant principalment dades simulades, que poden comportar-se de forma diferent de les dades de pacients reals. Lepkes et al. (2021) només verifiquen experimentalment les CNVs detectades pels programes, fet que no proporciona una visió completa del rendiment de les eines, ja que podrien estar obviant falsos negatius. El tercer estudi, una avaluació realitzada pel nostre grup, aborda aquestes limitacions, perquè avalua cinc eines en quatre conjunts de dades de panell amb resultats de MLPA disponibles (Moreno-Cabrera et al., 2020). La limitació més significativa d'aquest estudi és que únicament cobreix les eines disponibles fins a 2018.

Alteracions de mida mitjana

En aquesta tesi emprarem el terme mida mitjana per referir-nos a les variants que engloben de 20 bp a 1kb. Aquestes variants són complicades d'alinear correctament al genoma de referència, ja que poden o bé comprendre una part important de la lectura o inclús sobrepassar-la (Sedlazeck *et al.*, 2018; Mahmoud *et al.*, 2019). Encara que en alguns casos aquestes variants poden afectar el nombre de còpies, també són difícils de detectar per programes de RD, perquè les finestres d'on s'extreu la cobertura solen ser més grans i no tenen prou poder estadístic (Barbitoff et al., 2024).

Per tant, per detectar aquestes variants, és necessari l'ús d'un programari que combini altres estratègies esmentades en els apartats anteriors, com són el PEM i el SR. Aquesta aproximació no només permet detectar variants de mida mitjana, sinó també SVs equilibrades i altres alteracions complexes com les insercions de transposons. En aquest context, una revisió realitzada per Barbitoff et al. (2024) destaca GRIDSS (Cameron et al., 2017) com a programa prometedor per a aquesta tasca. GRIDSS utilitza la notació *breakend* per presentar les variants, la qual especifica les coordenades exactes dels punts de trencament i com es connecten els diferents fragments de DNA. Tot i que aquesta notació és especialment útil per descriure SVs complexes, presenta l'inconvenient que tots els esdeveniments es reporten com a punts de trencament simples (*BND*). Això dificulta la interpretació del tipus específic de variant i requereix passos addicionals per categoritzar correctament els esdeveniments detectats.

Zones repetitives o de baixa complexitat

Les variants que es troben dins d'un homopolímer, d'un microsatèl·lit o d'una zona de baixa complexitat representen un repte. La repetició i simplicitat de les seqüències en aquestes regions poden induir errors en les lectures, com ara identificar insercions o delecions que no són reals. Aquests errors dificulten l'assignació precisa de variants, ja que les eines poden confondre les variacions reals amb artefactes de seqüenciació.

A més, aquestes zones són propenses a generar variacions amb petites diferències que dificulten l'alineament de les seqüències, fent que sigui complicat definir clarament les variants i comparar-les amb les de les bases de dades poblacionals. Aquesta dificultat impacta en la determinació de la MAF, clau per distingir entre polimorfismes i variants potencialment patogèniques.

A causa de la naturalesa ambigua d'aquestes seqüències, és difícil obtenir dades de referència fiables que permetin, primer, desenvolupar i refinar eines, i després comparar i validar els resultats obtinguts per tal de trobar una eina de referència. Per tant, sovint aquestes regions estan poc representades en els tests de referència, ja que són zones del DNA poc confiables, fent que es sobreestimi la capacitat de les eines de crida de variants (Olson *et al.*, 2023).

Mosaïcisme

Depenent del moment en el desenvolupament on es produeixi una mutació post-zigòtica la línia sanguínia es pot veure o no afectada. Així, moltes variants en mosaic poden presentar baixa o nul·la VAF en sang, per la qual cosa sovint no es detecten per rutina diagnòstica, on se solen descartar les variants per sota del 10-20%, per evitar els falsos positius

Malgrat que avui en dia existeixen programes que ajuden a detectar aquestes variants, com Samovar o LinkedSV, la seva sensibilitat encara no és molt bona. Per tant, sovint la detecció d'aquestes variants s'emmarca més en context de recerca, on s'analitzen manualment en un programa visor del genoma, com pot ser el Integrative Genomics Viewer (IGV), les variants que es troben a una VAF inferior al punt de tall establert per diagnòstic.

D'altra banda, també pot passar a la inversa, que es detecti una mutació que per VAF sembli germinal, però que en realitat sigui deguda al fenomen d'hematopoesi clonal de potencial indeterminat (CHIP, de l'anglès *clonal hematopoiesis of indeterminate potential*). Aquest fenomen es dona quan una cèl·lula mare hematopoètica, que pot convertir-se en altres cèl·lules de la sang, comença a produir cèl·lules amb una determinada variant. Les variants associades a CHIP no es transmeten a la descendència ja que estan restringides a la línia sanguínia. Aquest fenomen s'ha descrit més freqüent en persones d'edat avançada, tot i que també es poden associar a tractaments com la radioteràpia i la quimioteràpia, així com al consum de tabac (Genovese *et al.*, 2014). Atès que la majoria dels tests genètics es fan a partir d'una mostra de sang, aquestes variants es poden identificar i donar lloc a interpretacions errònies. En aquests casos, que se solen sospitar davant VAFs baixes en variants de gens associats al fenomen, com *TP53* i *CHEK2*, la solució no pot ser bioinformàtica, sinó que caldria considerar la realització d'un estudi genètic a partir de fibroblasts, o altres teixits amb poca contaminació sanguínia.

7. Classificació de variants

Inicialment, la classificació de variants es basava en les poques dades que es disposava i en mètodes qualitatius i intuïtius. Això afavoria la subjectivitat i causava diferències notables entre laboratoris. Tan aviat com les anàlisis genètiques van començar a estendre's en l'àmbit clínic, es va reconèixer la necessitat de disposar de guies de classificació consensuades, tant per homogeneïtzar la interpretació de les variants com per guiar la presa de decisions clíniques.

7.1. Sistemes de classificació de variants

Previs a les guies ACMG

Plon et al. (2008) van començar a abordar aquesta necessitat amb el desenvolupament d'un sistema de classificació de variants, basat en la seva probabilitat de ser patogèniques, que les dividia en cinc classes (Taula 8). A més, van ser pioners en associar recomanacions clíniques específiques a cada nivell de classificació, tant per al seguiment del pacient estudiat com per iniciar estudis directes en cascada a la resta de familiars, o estudis en un context d'investigació per ajudar a classificar la variant.

En els anys següents, van començar a crear-se les primeres guies específiques en el camp del càncer hereditari, dedicades als principals gens de les síndromes més comunes de predisposició al càncer: *BRCA1* i *BRCA2* per al càncer de mama i ovari hereditaris (Spurdle *et al.*, 2012) i els gens MMR (*mismatch repair*) per la LS (Thompson *et al.*, 2014).

Taula 8: Sistema de classificació proposat per Plon et al., 2008 i recomanacions associades a cada classificació.

Sistema de classificació			Recomanacions clíniques		
Classe	**Descripció**	**Probabilitat de patogenicitat**	**Test clínic a familiars**	**Mesures de seguiment**	**Test genètic en context d'investigació**
1	No patogènic / sense significat clínic	<0,001	No	No	No
2	Probablement no patogènic / poc significat clínic	0,001-0,049			Pot ser útil per acabar de classificar la variant
3	Incertesa /desconegut	0,05-0.949	Sí	Basades en la història familiar	
4	Probablement patogènic	0,95-0,99		Alt Risc	
5	Patogènic	>0,99			No

Adpatada de Plon et al., 2008.

Guies ACMG

Per tal d'estandarditzar la interpretació de les variants genètiques, és a dir, establir un sistema per arribar a una classificació final coherent entre laboratoris, *l'American College of Medical Genetics*

and Genomics (ACMG) i l'Association of Molecular Pathology (AMP) van elaborar el 2015 unes guies de classificació generals (Richards et al., 2015).

Aquestes guies també adopten un sistema de cinc categories per classificar les variants: patogènica, probablement patogènica, de significat clínic desconegut, probablement benigna i benigna, i proposen probabilitats associades a la patogenicitat. No obstant això, la diferència principal respecte al sistema de Plon et al. (2008) radica en que les guies ACMG indiquen com realitzar la classificació, organitzant les evidències disponibles en criteris definits segons la seva procedència i naturalesa. A més, a cada criteri li associen un pes a favor de patogenicitat o benignitat. Per identificar els diferents criteris fan servir codis. Aquests comencen per una P o una B per indicar si és un criteri cap a patogenicitat o benignitat, seguit d'una lletra que n'indica el pes: A: autònom (de l'anglès *standalone*), VS: molt fort (de l'anglès *very strong*), S: fort (de l'anglès *strong*), M: moderat, P: de suport i finalment acaben per un número per distingir-los (Taules 9 i 10).

Tipus d'evidències

En funció de la seva naturalesa, les evidències es classifiquen en els següents 8 grups:

- **Dades de poblacions**: aquestes evidències s'extreuen de bases de dades poblacionals i estudis de cas-control, i permeten comparar les freqüències de les variants amb població control, partint de la suposició que les variants patogèniques estan sotmeses a pressió selectiva i/o que es trobaran més freqüentment en individus afectes. Aquí s'inclouen els criteris PM2, BA1, BS1 i BS2.
- **Computacionals o predictives**: Fan prediccions sobre la conseqüència de la variant en el gen, utilitzant algoritmes i models bioinformàtics o comparant-ho amb la classificació d'altres variants en la mateixa posició. Inclouen els criteris PVS1, PS1, PM4, PM5 i PP3, BP4 i BP7.
- **Dades d'estudis funcionals**: Evidències provinents d'assajos experimentals que avaluen l'impacte de la variant en la funció de la proteïna i del domini en què es localitza. Engloben els criteris PS3, PM1, PP2, BS3 i BP3.
- **Dades de segregació**: Informació sobre com la variant es transmet dins de les famílies afectades. Inclouen els criteris PP1 i BS4.
- **Dades al·lèliques:** Informació sobre la situació al·lèlica (en *cis* o *trans*) de la nostra respecte d'altres variants en pacients o sans per la malaltia. Engloben PM3 i BP2.
- **Variants *de novo*:** variants que apareixen en un individu afectat sense que estiguin presents en els seus progenitors sans, indicant una possible causa de malaltia. Inclouen els criteris PS2 i PM6.
- **Altres bases de dades**: Informació de la classificació de patogenicitat de la variant en bases de dades reputades. Engloben els criteris PP5 i BP6.
- **Altres criteris**: valoren si la variant es troba en pacients que tenen el fenotip específic associat al gen o en individus que ja tenen altres variants que explicarien aquest fenotip. Engloben PP4 i BP5.

Combinació de criteris

Un cop atorgats els criteris, s'estableixen regles per combinar-los i així assolir la classificació final de la variant. En la Figura 22 es mostra els diferents escenaris que contemplen les guies ACMG.

.

Taula 9: Criteris a favor de patogenicitat contemplats en les guies ACMG amb la seva descripció associada		
Pes	**Codi**	**Descripció**
Molt Fort	PVS1	Variant nul·la (sense sentit, *frameshift*, canònica ±1 o 2 del lloc d'empalmament, codó d'inici, deleció d'un o múltiples exons) en un gen on la pèrdua de funció és un mecanisme conegut de malaltia.
Fort	PS1	Canvi d'aminoàcid igual a una variant patogènica establerta prèviament, independentment del canvi de nucleòtid. Per variants que no afecten l'empalmament.
	PS2	Variant *de novo* (maternitat i paternitat confirmades) en un pacient amb la malaltia i sense història familiar.
	PS3	Estudis funcionals *in vitro* o *in vivo* ben establerts donen suport a un efecte perjudicial en el gen o producte genètic.
	PS4	La prevalença de la variant en individus afectats és significativament augmentada en comparació amb la prevalença en controls.
Moderat	PM1	Localitzada en un punt calent mutacional i/o domini funcional crític i ben establert sense variació benigna.
	PM2	Absent en controls (o a una freqüència extremadament baixa si és recessiu) en Exome Sequencing Project, 1000 Genomes Project, o Exome Aggregation Consortium.
	PM3	Detectada en trans amb una variant patogènica (únicament en trastorns recessius).
	PM4	Canvis en la longitud de la proteïna com a resultat de delecions/insercions en pauta en una regió no repetitiva o variants de pèrdua de stop.
	PM5	Nou canvi *missense* en un residu d'aminoàcid on un canvi missense diferent determinat com a patogènic s'ha vist previament. Per variants que no afecten l'empalmament.
	PM6	Assumit *de novo*, però sense confirmació de paternitat i maternitat.
Suport	PP1	Co-segregació amb la malaltia en múltiples membres afectats de la família en un gen que clarament causa la malaltia.
	PP2	Variant *missense* en un gen que té poques *missense* benignes i en el qual les variants *missense* són un mecanisme comú de malaltia.
	PP3	Múltiples línies d'evidència computacional donen suport a un efecte perjudicial en el gen o producte genètic (conservació, evolució, impacte en l'empalmament).
	PP4	El fenotip del pacient o la història familiar és altament específica per a una malaltia amb una única etiologia genètica.
	PP5	Una font reputada considera la variant patogènica, però l'evidència no està disponible per al laboratori per fer una avaluació independent.

Adaptat i resumit de Richards et al. (2015). S'ha mantingut l'essència de cada criteri però s'han obviat les excepcions.

Taula 10: Criteris a favor de benignitat contemplats en les guies ACMG amb la seva descripció associada		
Pes	**Codi**	**Descripció**
Autònom	BA1	La freqüència al·lèlica és >5% en Exome Sequencing Project, 1000 Genomes Project o Exome Aggregation Consortium.
Fort	BS1	La freqüència al·lèlica és superior a l'esperada per al trastorn.
	BS2	Observada en un adult sa respecte un trastorn recessiu (homozigot), dominant (heterozigot) o lligat a l'X (hemizigot), amb penetrància completa esperada a una edat primerenca.
	BS3	Estudis funcionals ben establerts *in vitro* o *in vivo* no mostren cap efecte perjudicial en la funció de la proteïna o en l'empalmament.
	BS4	Manca de segregació en membres afectats d'una família.
Suport	BP1	Variant *missense* en un gen on es coneixen principalment variants truncants com a causa de la malaltia.
	BP2	Observada en *trans* amb una variant patogènica per a un gen/trastorn dominant completament penetrant o observada en *cis* amb una variant patogènica en qualsevol patró d'herència.
	BP3	Delecions/insercions en pauta en una regió repetitiva sense una funció coneguda.
	BP4	Múltiples línies d'evidència computacional suggereixen cap impacte en el gen o producte genètic (conservació, evolució, impacte en l'empalmament, etc.).
	BP5	Variant trobada en un cas amb una base molecular alternativa per a la malaltia.
	BP6	Una font reputada considera la variant benigna, però l'evidència no està disponible per al laboratori per realitzar una avaluació independent.
	BP7	Variant sinònima per a la qual els algorismes de predicció d'empalmament no prediuen cap impacte en la seqüència consens d'empalmament ni la creació d'un nou lloc, i el nucleòtid no és altament conservat.

Adaptat i resumit de Richards et al. (2015). S'ha mantingut l'essència de cada criteri però s'han obviat les excepcions.

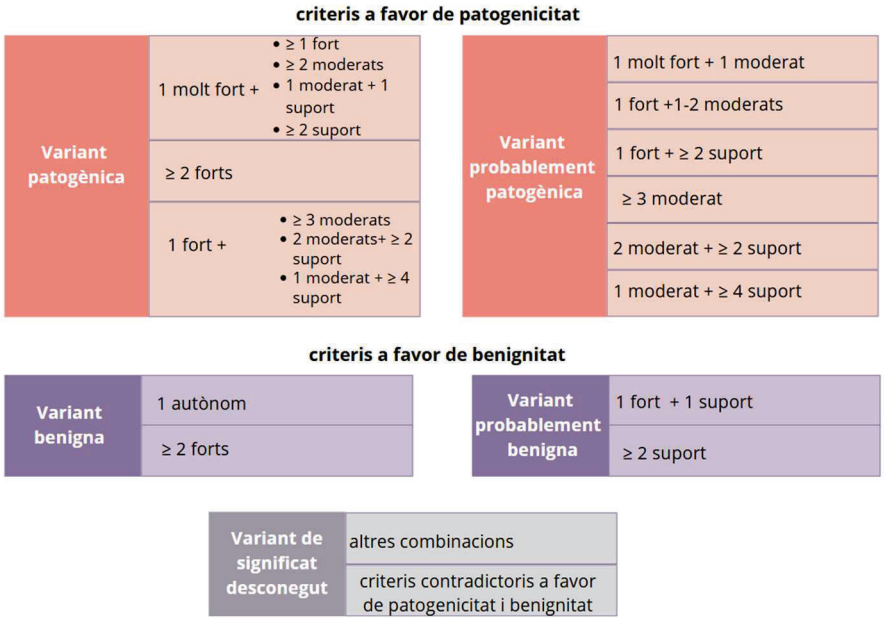


Figura 22: Regles per combinar els criteris i assolir la classificació final de la variant segons les guies ACMG/AMP. Adaptat de Richards et al. (2015).

Adaptacions de les guies ACMG/AMP

No obstant, la combinació/integració dels criteris proposada per les guies ACMG presenta certes limitacions. En primer lloc, no contempla algunes combinacions possibles. En segon lloc, mostra incongruències en la probabilitat de patogenicitat d'algunes combinacions. En tercer lloc, qualsevol criteri contradictori ja és motiu per classificar la variant com a variant de significat desconegut (VSD), no admetent cap tipus de discrepància. A més, el sistema és poc intuïtiu, per tant, cal estar constantment revisant les combinatòries.

Per tal d'abordar aquestes limitacions, Tavtigian et al. (2018) van demostrar primer que les guies ACMG/AMP són compatibles amb una formulació bayesiana de probabilitats, assumint quatre nivells de força (suport, moderada, forta i molt forta) i una probabilitat de patogenicitat escalada exponencialment (Taula 10). Posteriorment, van convertir les categories de força en un sistema de puntuació amb nombres naturals que facilita el càlcul de la classificació final (Tavtigian et al., 2020). Així doncs, cada criteri complert a favor de patogenicitat amb una força de suport, moderada, forta o molt forta, suma 1,2,4 i 8 respectivament a la puntuació final. D'altra banda, cada criteri a favor de benignitat resta aquests valors. Aquesta suma d'exponents determina la classificació final de la variant (Taules 11 i 12).

Taula 11: Càlculs de probabilitat per obtenir la força del criteri

Força del criteri	Odds combinades de patogenicitat	Suma d'exponents
Suport	2,08 ($2,08^1$)	1
Moderada	4,33 ($2,08^2$)	2
Forta	18,7 ($2,08^4$)	4
Molt forta	350 ($2,08^8$)	8

Extret de Garrett et al. (2021).

Taula 12: Suma dels exponents i probabilitat posterior de patogenicitat associada a cada (sub)classe de la classificació.

	Tavtigian et al. (2020)		Garrett et al. (2021) CanVIG-UK		
Classe	Suma dels exponents	Límits basats en probabilitats posteriors (PP) de patogenicitat*	Sub classe	Suma dels exponents	Límits basats en probabilitats posteriors (PP) de patogenicitat*
Patogènica	≥10	PP >0,99	-	≥10	PP > 0,99
Probablement patogènica	6-9	0,99≥ PP > 0.90	-	6-9	0,99 ≥ PP >0,90
Significat clínic desconegut	0-5	0,10 ≤ PP ≤0,90	VSD calenta	5	0,90 ≥ PP >0,812
			VSD càlida	4	0,812 ≥ PP >0,675
			VSD tèbia	3	0,675 ≥ PP >0,5
			VSD freda	2	0,5≥ PP >0,325
			VSD gelada	1	0,325≥ PP >0,188
				0	0,188≥ PP >0,1
Probablement benigna	(-1)-(-6)	0,001 ≤ PP < 0,10	-	(-1)-(-5)	0.1≤ PP <0,001
Benigna	≤-7	PP <0,001	-	≤-6	<0,001

Adaptat de Tavtigian et al. (2020); Garrett et al., (2021). VSD: variant de significat clínic desconegut; *assumint una probabilitat a priori de patogenicitat de 0.1.

Gràcies a aquests càlculs, Tavtigian et al. (2018) van evidenciar que una de les combinacions considerada probablement patogènica segons les guies ACMG/AMP computaria en realitat com a patogènica (combinació d'una evidència molt forta i 1 moderada) i a la inversa, una considerada patogènica computaria com a probablement patogènica (≥2 evidències fortes). A més, van permetre que combinacions que inclouen criteris a favor i en contra de la patogenicitat puguin donar classificacions de (probablement) benigna o patogènica, mentre que en la proposta original d'ACMG/AMP es quedaven com a VSDs.

Aquesta aproximació va ser adoptada, amb algunes petites modificacions, pel grup d'interpretació de variants del Regne Unit (CanVIG-UK, de l'anglès *Cancer Variant Interpretation Group UK*) (Garrett et al., 2021). Addicionalment, CanVIG-UK va introduir l'estratificació de les VSDs (que van de 0 a 5 punts). Això possibilita prioritzar les VSDs amb puntuacions més altes, considerades calentes, on només un criteri addicional amb força de suport seria suficient per classificar-les com probablement patogèniques (Garrett et al., 2021) (Taula 12).

CanVIG-UK també va suggerir que alguns criteris eren incompatibles en funció del tipus de variant i que altres podien ser repetitius, ja que no eren independents, i per tant no haurien de combinar-se. Per tant, van definir quines combinacions estaven permeses (Figura 23) (Garrett et al., 2021).

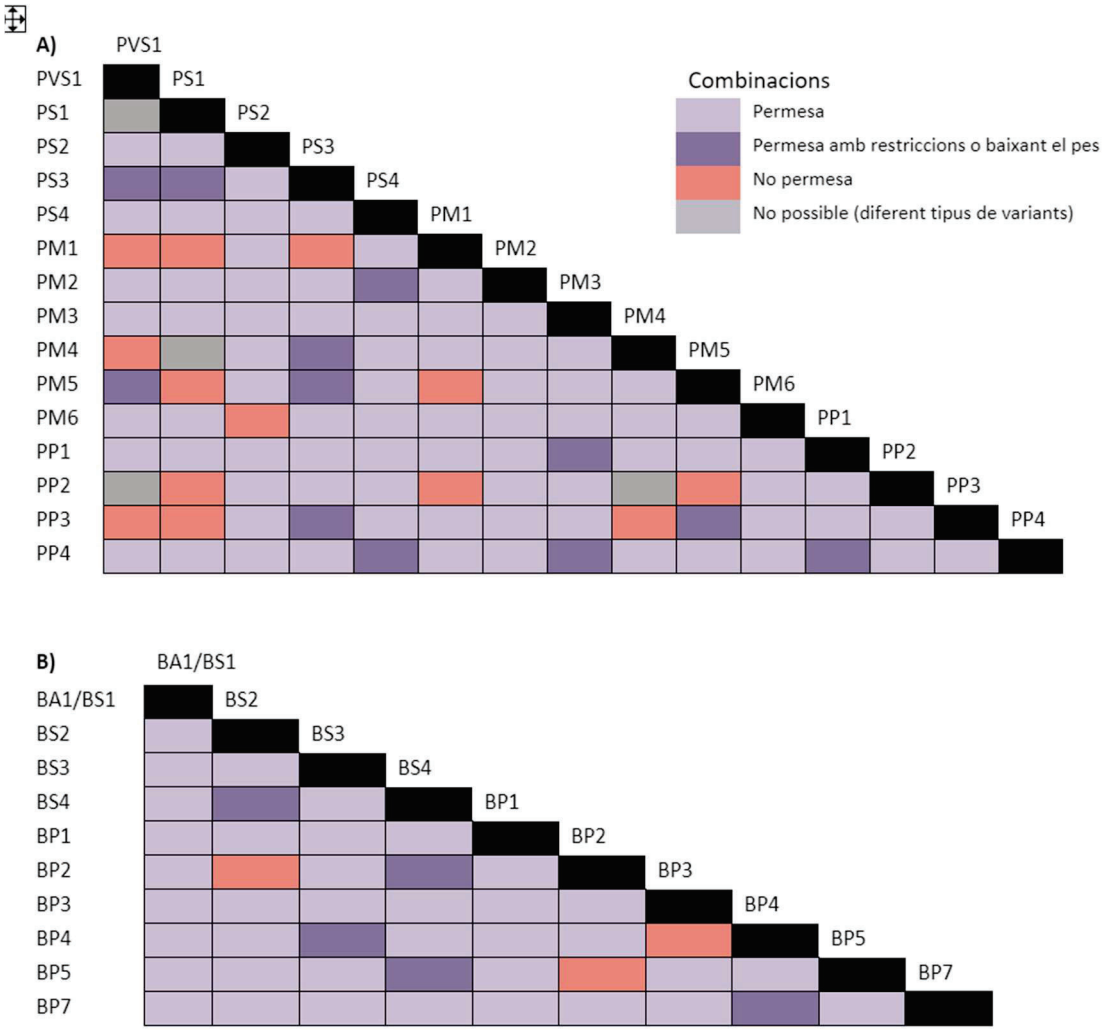


Figura 23: Combinacions permeses a favor de A) patogenicitat i B) benignitat. Adaptat de Garrett et al, no s'ha considerat el criteri PP5 ni BP6, tal i com proposen (Biesecker et al., 2018). En lila clar es veuen les combinacions permeses, en lila fosc les que estan permeses però hi assignen restriccions, amb rosa les que no s'haurien de combinar degut a repeticions i en gris les que no són possibles degut a incompatibilitat amb tipus de variant.

Guies ClinGen

Malgrat els esforços per estandarditzar els criteris entre laboratoris, moltes de les recomanacions de les guies ACMG són generalistes, qualitatives i no prou precises, el que comporta diferències en la interpretació clínica de les variants entre laboratoris.

En aquest context va emergir *The Clinical Genome Resource* (ClinGen: https://clinicalgenome.org/), una iniciativa que canvià el focus generalista per un de més concret, centrat en cada gen, permetent així avançar de forma més específica. Per fer-ho possible, van constituir panells d'experts per gens o conjunts de gens relacionats, encarregats de definir la rellevància clínica i associar-los a un fenotip o malaltia. Posteriorment, s'han anat creant panells d'experts per curar variants, que estableixen guies específiques de gen, les quals són sotmeses a una avaluació contínua.

Els diversos panells d'experts cobreixen una àmplia varietat de camps, com per exemple, malalties cardiovasculars, errors innats del metabolisme, càncer hereditari, malalties immunològiques, trastorns neurològics i del desenvolupament, d'entre d'altres.

En data de novembre de 2024, en el camp del càncer hereditari hi ha les següents guies creades o en preparació: *APC, ATM, BRCA1, BRCA2, CDH1, DDX41*, *DICER1, MLH1, MSH2 MSH6, MUTYH, PALB2, PMS2, PTEN, RAD51C, RUNX1, TP53 i VHL* (Taula 13).

Taula 13: Llistat de gens amb guies ClinGen específiques.

Gens	Data primera versió	Versió actual i data
APC	10/1/2023	v.2.1.0 (24/11/2023)
ATM	19/1/2022	v.1.3.0 (27/3/2024)
BRCA1	9/8/2023	v.1.1.0 (21/12/2023)
BRCA2	9/8/2023	v.1.1.0 (21/12/2023)
CDH1	19/9/2018	v.3.1.0 (7/12/2022)
DDX41	En preparació	-
DICER1	5/5/2022	v.1.3.0 (30/1/2024)
MLH1	v.1.0.0 (8/9/2024)	v.1.0.0 (8/9/2024)
MSH2	v.1.0.0 (8/9/2024)	v.1.0.0 (8/9/2024)
MSH6	v.1.0.0 (8/9/2024)	v.1.0.0 (8/9/2024)
MUTYH	En preparació	-
PALB2	17/3/2023	v.1.1.0 (28/11/2023)
PMS2	v.1.0.0 (8/9/2024)	v.1.0.0 (8/9/2024)
PTEN	17/8/2018	v.3.1.0 (14/3/2024)
RAD51C	En preparació	-
RUNX1	10/7/2019	v.2.0.0 (15/9/2021)
TP53	6/8/2019	v.2.2.0 (30/09/2024)
VHL	29/2/2024	v.1.0.0 (29/2/2024)

Es mostra la data de la primera versió i la de la versió actual (revisat a novembre del 2024).

A part, ClinGen ha estat proactiu en l'actualització i refinament d'alguns criteris. D'una banda, ha suggerit eliminar els criteris PP5 i BP6, ja que considerava que les fonts amb reputació que no estiguin directament vinculades a l'evidència de suport no haurien d'utilitzar-se com a criteri (Biesecker *et*

al., 2018). D'altra banda, ha proposat considerar els diferents tipus de pèrdua de funció, elaborant un arbre de decisió que considerava el tipus de variant, la seva localització i altres evidències, assignant diferent forces possibles al criteri PVS1 (Abou Tayoun et al., 2018). També ha publicat recomanacions específiques per criteris com el PM3, per ajustar la força basada en observacions en *trans*, el PS3 i BS3, per avaluar assaigs funcionals segons la seva robustesa (Brnich *et al.*, 2019) i, PP3 i BP4, per calibrar les eines computacionals emprades en la classificació de variants *missense* per així maximitzar-ne la fiabilitat (Pejaver *et al.*, 2022) . A més, entre altres, ha desenvolupat recomanacions per interpretar CNVs (Riggs *et al.*, 2020) , variants que afecten l'empalmament, incloent l'ús de models de predicció validats (Walker *et al.*, 2023) i per la curació i classificació d'al·lels de risc i de baixa penetrància (Schmidt *et al.*, 2024).

7.2. Implementació de la classificació de variants en laboratoris clínics

La classificació de les variants suposa un gran repte pel diagnòstic genètic, donat que només una correcta classificació permet un assessorament genètic i personalització del risc encertats.

A banda de ser un procés manual complex i laboriós que, com s'ha comentat, es basa en l'acumulació, interpretació i combinació de les dades seguint guies específiques de gen, és també un procés iteratiu. Així doncs, la classificació de les variants s'ha d'anar revisant periòdicament perquè contínuament es publiquen estudis que aporten noves evidències per a la classificació.

Si aquest procés, ja llarg i complicat per una sola variant, l'extrapolem a les moltes variants noves que es detecten en un sol pacient en emprar tècniques NGS, es fa evident que constitueix el principal coll d'ampolla de les unitats de diagnòstic genètic. Aquest increment de feina és únicament assumible o bé incrementant el personal o automatitzant part del procés, per tal d'agilitzar-lo. Per aquest motiu, molts laboratoris de diagnòstic genètic han optat per adquirir solucions comercials que facilitin aquest procediment. Malgrat que aquests programes permeten automatitzar gran part de l'anàlisi, no tots els criteris poden ser extrets automàticament.

Alguns d'aquestes solucions sovint formen part de l'anàlisi terciària oferta per empreses privades com Emedgene d'Illumina (https://www.illumina.com/products/by-type/informatics-products/emedgene.html), Moon d'Invitae (https://diploid.atlassian.net/wiki/spaces/MOON/pages/360571/Release+notes), eVaI d'enGenome (https://evai.engenome.com/) o programes interns de Fabric Genomics (https://fabricgenomics.com/fabric-gem/) i BluePrint Genetics (https://blueprintgenetics.com/variant-classification/).

Existeixen també d'altres programes que són oberts o amb llicències gratuïtes amb restriccions A la Taula 14 es pot veure una comparativa detallada dels diferents programes amb opcions gratuïtes. La majoria d'aquestes eines utilitzen les guies genèriques de l'ACMG/AMP, com per exemple, Intervar (Li and Wang, 2017), PathoMAN (Joseph, Ravichandran and Offit, 2017), ClinGen Pathogenicity Calculator (Patel *et al.*, 2017), Varsome (Kopanos *et al.*, 2019), CharGer (Scott *et al.*, 2019), TAPES (Xavier, Scott and Talseth-Palmer, 2019), Franklin (https://franklin.genoox.com/clinical-db/home) o la recent eina GeneBe (Stawiński and Płoski, 2024).

També hi ha programes que es centren en un conjunt de gens o fenotips específics. Per exemple, CardioClassifier (Whiffin *et al.*, 2018) i CardioVai (Nicora *et al.*, 2018) per malalties cardíaques hereditàries, Variant Interpretation Platform for genetic Hearing Loss (VIP-HL) (Peng *et al.*, 2021) i GenOtoScope (Melidis *et al.*, 2022) per malalties de pèrdua d'audició, mentre que Cancer Predisposition Sequencing Reporter (CPSR) (Nakken *et al.*, 2021) i Cancer SIGVAR (Li et al., 2021) per gens de predisposició al càncer. CPSR segueix l'algoritme SherLoc, que proposa millores a les guies

originals de l'ACMG (Nykamp *et al.*, 2017). Innovadorament, Cancer-SIGVAR inclou les actualitzacions de Tayoun et al. (2018) i semi-automatitza les guies gen-específiques de ClinGen per *PTEN* (Mester *et al.*, 2018), *CDH1* (Lee *et al.*, 2018), *RASopaties* (Gelb *et al.*, 2018), *RUNX1* (Luo *et al.*, 2019) i pèrdua d'oïda (Oza *et al.*, 2018). No obstant, aquestes guies s'han anat actualitzant i el programa no ha estat modificat per considerar les noves versions.

Al nostre coneixement, no existeix cap eina que semi automatitzi les guies específiques d'alguns dels gens més rellevants en el camp de l'HC, com són *ATM*, *BRCA1*, *BRCA2*, *CHEK2*, els gens MMRs, *PALB2* i *TP53*.

Taula 14: Taula comparativa de les eines de classificació de variants disponibles gratuïtament				
	Eines			
	Varsome	**ClinGen Pathogenicity Calculator**	**PathoMAN**	**InterVar**
Llenguatge de programació	no detallat	R i JavaScript + ExtJS	no detallat	python
Llicència gratuïta amb restriccions?	amb restriccions	sí	sí	sí
Tipus de llicència	programari comercial	GNU Affero General Public License 3.0	ús personal o de recerca, no per a fins comercials	GNU General Public License (GPL) versió 3
Entrada	variants úniques amb nomenclatura de DNA codificant o genòmica	variants úniques en nomenclatura HGVS	variants úniques o càrrega en lot amb un arxiu csv (només en genòmic)	arxiu preanotat o arxiu vcf
Assemblatge del genoma	GRCh37 i GRCh38	no trobat	GRCh37	GRCh37 i GRCh38
Sortida	recurs web interactiu	Informe HTML interactiu	variants soles: (web), en lot, al correu	InterVar (línia de comandament): tabular i wInterVar: web
Format de fitxer de descàrrega	txt (tabulat)	xls	pdf	csv
Regles de classificació - assignació de criteris	Directrius ACMG/AMP amb algunes consideracions	Directrius originals ACMG/AMP (Richards et al., 2015)	Directrius originals ACMG/AMP (Richards et al., 2015)	Directrius originals ACMG/AMP (Richards et al., 2015)
Regles de classificació - combinació de criteris	Sistema de punts naturalment escalat basat en una formulació bayesiana: - Tavtigian et al., 2018 i 2020	Directrius originals ACMG/AMP (Richards et al., 2015)	Directrius originals ACMG/AMP (Richards et al., 2015)	Directrius originals ACMG/AMP (Richards et al., 2015)
Segueix guies específiques per gen?	no	no	no	no
Accés	https://varsome.com/	https://calculator.clinicalgenome.org/site/cg-calculator	https://pathoman.mskcc.org/	https://github.com/WGLab/InterVar https://wintervar.wglab.org
Data de llançament	2016	2016	2016	**2016**
Última actualització: versió i data	12.1.0 (juliol 2024)	3.0.01 (març 2018)	no detallat	v2.2.1 (agost 2021)

	Eines			
	GeneBe	**Franklin**	**TAPES**	**CharGer**
Llenguatge de programació	python	no detallat	python3	python 2.7
Llicència gratuïta amb restriccions?	sí	Sí, també pagament	sí	sí
Tipus de llicència	drets d'autor, marques registrades i altres lleis tant de Polònia com de països estrangers	no trobat	MIThg	General Public License (LGPL) v3.0
Entrada	variants úniques o en lot en format VCF, SPDI, HGVS	arxiu vcf, microarray o variants úniques amb nomenclatura de DNA codificant o genòmica	arxiu vcf, gzipped vcf o vcf binari (BCF)	arxiu vcf
Assemblatge del genoma	GRCh37, GRCh38 i T2T	GRCh37 i GRCh38	GRCh37 i GRCh38	GRCh37
Sortida	recurs web interactiu o via API	recurs web interactiu	Línia de comandament	tsv o html
Format de fitxer de descàrrega	via API vcf anotat	no possible	vcf anotat	no possible
Regles de classificació - assignació de criteris	Directrius originals ACMG/AMP i recomanacions adicionals de SVI de Clingen	Directrius originals ACMG/AMP amb algunes consideracions	Directrius originals ACMG/AMP	Directrius originals ACMG/AMP (Richards et al., 2015)
Regles de classificació - combinació de criteris	Sistema de punts naturalment escalat basat en una formulació bayesiana: Tavtigian et al., 2018 i 2020	no detallat	Dona una probabilitat (de 0 a 1) de ser patogènica bastat en una formula bayesiana: Tavtigian et al., 2018	Puntuació pròpia
Segueix guies específiques per gen?	no	no	no	no
Accés	https://genebe.net/https://genebe.net/https://pypi.org/project/genebe/	https://franklin.genoox.com/clinical-db/home	https://github.com/a-xavier/tapes	https://github.com/ding-lab/CharGer
Data de llançament	**2024**	no trobat	2019	2017
Última actualització: versió i data	v.0.1.10 (juliol 2024)	v.75.3 (juliol 2024)	V-0-1-2 (agost 2023)	v0.5.4 (octubre 2019)

	Eines				
	GenOtoScope	**VIP-HL**	**Cardio-Classifier**	**CardioVai**	**Cancer-SIGVAR**
Llenguatge de programació	pyhton, HTML5 i CSS	Java i Python, la web en HTML 5 i JavaScript	Perl i PHP	no detallat	no detallat
Llicència gratuïta amb restriccions?	sí	sí, per recerca	sí	no detallat	sí, només per a ús de recerca
Tipus de llicència	GNU Affero General Public License.	no detallat	https://www.cardioclassifier.org/FinalLicenceAgreement.pdf	no detallat	no trobat
Entrada	vcf	coordenada genòmica, nomenclatura HGVS, gen, dbSNP i proteïna	fitxer vcf o variant simple	coordenada genòmica o nomenclatura HGVS	fitxer vcf o variants nomenclatura genòmica o de DNA codificant
Assemblatge del genoma	GRCh37	GRCh37	no detallat	no detallat	GRCh37
Sortida	recurs web i línia de comandament	recurs web interactiu	recurs web interactiu	recurs web interactiu	recurs web
Format de fitxer de descàrrega	tsv i html	no possible	no possible	txt	no possible
Regles de classificació - assignació de criteris	Guies ACMG/AMP actualitzades: Richards et al., 2015 i Tayoun et al., 2018	Basat en 13 regles Clingen de pèrdua d'oïda (Oza et al., 2018)	Guies ACMG/AMP actualitzades: Biesecker et al., 2018,llindars propis per a cada criteri gen-específics.	Guies originals ACMG/AMP amb certes modificacions (ex: introduir criteri BP8)	Guies ACMG/AMP actualitzades: Richards et al., 2015 i Tayoun et al., 2018
Regles de classificació - combinació de criteris	Guies originals ACMG/AMP (Richards et al., 2015)	Guies originals ACMG/AMP (Richards et al., 2015)	Guies originals ACMG/AMP (Richards et al., 2015)	Guies originals ACMG/AMP (Richards et al., 2015)	Adaptació de les regles ACMG/AMP per optimitzar les regles de gestió de conflictes. - Richards et al., 2015 - Ellard et al., 2019
Segueix guies específiques per gen?	Sí, guies ClinGen de pèrdua d'audició (Oza et al., 2018)	Sí, guies ClinGen de pèrdua d'audició (Oza et al., 2018)	sí, condicions cardíaques hereditàries	sí, guies *MYH7* v1	sí: Guies específiques de ClinGen: *CDH1* v1, *PTEN* v1, *RUNX1* v1, Pèrdua d'audició i RASopaties v1
Accés	https://genotoscope.mh-hannover.de/	http://hearing.genetics.bgi.com/	https://www.cardioclassifier.org	http://cardiovai.engenome.com/	http://cancersigvar.bgi.com/
Data de llançament	2022	2020	2021	2018	no detallat
Última actualització: versió i data	no detallat	no detallat	v0.2.0 (2017)	no detallat	2021

HIPÒTESIS

El principal supòsit d'aquesta tesi és que una aproximació bioinformàtica acurada pot millorar significativament la interpretació dels resultats obtinguts amb les dades d'NGS de lectures curtes en el context del diagnòstic genètic del càncer hereditari. A partir d'aquest supòsit general, es plantegen les següents hipòtesis específiques:

- Les lectures dels gens que tenen pseudogens d'alta homologia poden alinear-se incorrectament, fet que pot provocar la no identificació de variants rellevants. Una solució bioinformàtica podria augmentar la sensibilitat en la detecció de variants patogèniques, permetre utilitzar l'NGS com a tècnica primària de cribratge i limitar l'ús de les LR-PCRs únicament per confirmar els casos positius. Aquest enfocament seria especialment útil per a casos com el gen *PMS2*, on l'alta homologia amb el pseudogèn *PMS2CL* representa un repte diagnòstic .

- L'aparició de noves eines de detecció de CNVs, juntament amb versions més recents de les ja existents, pot suposar una millora en la sensibilitat, l'especificitat i el rendiment diagnòstic global. Una comparativa metòdica amb conjunts de dades adequats pot ajudar a avaluar les diferents aproximacions, identificar les eines amb millors prestacions i determinar els paràmetres clau per optimitzar-ne l'ús.

- Els *pipelines* actuals utilitzats en el nostre laboratori de diagnòstic no estan dissenyats per a la detecció de SVs de mida mitjana equilibrades ni insercions de transposons, ja que es basen exclusivament en l'estratègia de profunditat de cobertura. L'ús d'eines que integrin diverses estratègies de detecció permetria identificar aquestes variants i així millorar el rendiment diagnòstic.

- Algunes famílies amb sospita de càncer hereditari sense causa genètica identificada podrien explicar-se per variants en les regions 5'UTR, les quals, tot i estar parcialment cobertes en el panell I2HCP, no s'analitzen per rutina diagnòstica.

- L'automatització del procés de classificació de variants, amb un programa que implementi les guies de l'ACMG i les recomanacions específiques per gen de ClinGen, podria agilitzar el procés i reduir el risc d'errors manuals, augmentant la consistència en la classificació de variants.

OBJECTIUS

L'objectiu general d'aquesta tesi **és millorar el diagnòstic genètic del càncer hereditari mitjançant una contribució bioinformàtica.** El present treball es focalitza en la implementació de noves eines bioinformàtiques així com la implementació de noves metodologies per la interpretació dels resultats de seqüenciació massiva. L'objectiu final és que aquestes eines puguin ser integrades a la pràctica diagnòstica, permetent una major precisió en la interpretació biològica de les dades genètiques i per tant una millora en la personalització del risc en els pacients estudiats que podran beneficiar-se d'un maneig clínic més acurat.

Objectius específics

1. Desenvolupar solucions bioinformàtiques per millorar l'anàlisi de les dades de seqüenciació massiva i facilitar la identificació de variants potencialment no detectades amb les anàlisis prèvies, amb la finalitat d'aplicar aquestes estratègies tant de manera retrospectiva com prospectiva. En concret s'ha treballat en els següents aspectes:

    1.1. Implementar un codi per millorar la precisió en la detecció de variants de *PMS2* utilitzant dades de panells NGS de rutina i reduir el nombre de mostres que s'han d'estudiar mitjançant PCR llarga (LR-PCR).
    1.2. Avaluar una bateria d'eines dissenyades per a la detecció de CNVs en dades de panell.
    1.3. Identificar tipus de variants que actualment es poden estar perdent amb els *pipelines* utilitzats.
    1.4. Explorar les regions 5'UTR i prioritzar a nivell bioinformàtic possibles variants reguladores de l'expressió gènica.

2. Optimitzar el procés de classificació de variants mitjançant automatització d'alguns passos:
    2.1. Semi-automatitzar i ponderar l'assignació del màxim nombre d'evidències ACMG, amb criteris adaptats a cada gen, per agilitzar la classificació i reduir el possible error manual.
    2.2. Implementar l'opció d'introduir manualment i enregistrar les evidències que no es poden automatitzar.
    2.3. Generar una classificació de la variant basada en el conjunt d'evidències, que segueixi els algoritmes definits per les guies internacionals.
    2.4. Produir una eina bioinformàtica en format d'accés obert per la comunitat.

INFORME DE LES DIRECTORES

A continuació es descriu el paper de la doctoranda en cadascun dels treballs presentats, juntament amb els factors d'impacte (IF) de les revistes on s'han publicat els resultats. Cap dels articles publicats ha estat presentat com a part d'altres tesis doctorals.

Article publicat 1: ***Open-Source Bioinformatic Pipeline to Improve PMS2 Genetic Testing Using Short-Read NGS Data***

Elisabet Munté, Lídia Feliubadaló, Jesús Del Valle, Sara González, Mireia Ramos-Muntada, Judith Balmaña, Teresa Ramon y Cajal, Noemí Tuset, Gemma Llort, Juan Cadiñanos, Joan Brunet, Gabriel Capellá, Conxi Lázaro* i Marta Pineda*.

Revista: *The Journal of Molecular Diagnostics*; volum 26; número 8.

Data: Agost 2024; doi: 10.1016/j.jmoldx.2024.05.005

IF 2023 = 3,4; Categoria: *Pathology*; Q1/D3; Rang: 22/88

Contribució de la doctoranda: revisió de bibliografia sobre articles de *PMS2*, disseny, elaboració i optimització del codi, validació del codi, implementació del codi a mostres d'una cohort de HC i anàlisi dels resultats, redacció del manuscrit.

Article publicat 2: ***Detection of germline CNVs from gene panel data: benchmarking the state of the art***

Elisabet Munté[¥], Carla Roca[¥], Jesús del Valle, Lídia Feliubadaló, Marta Pineda, Bernat Gel, Elisabeth Castellanos, Bárbara Rivera, David Cordero, Víctor Moreno, Conxi Lázaro* i José Marcos Moreno-Cabrera*.

[¥]Primera autoria compartida.

Revista: *Briefings in Bioinformatics*

Data: Desembre 2024; doi: 10.1093/bib/bbae645

IF 2023 = 6,8; Categoria: *Biochemical research methods*; Q1/D1; Rang= 4/85

Contribució de la doctoranda: revisió sistemàtica dels articles publicats (conjuntament amb Carla Roca), selecció de les eines per identificar CNVs que compleixen els requisits (conjuntament amb Carla Roca), creació del codi CNVbenchmarkeR2 per 5 eines clearCNV, Atlas-CNV, Cobalt, CNVkit, VisCap, creació de tot el codi d'avaluació dels paràmetres, creació del codi de comparació de les eines per parelles, anàlisis estadístiques, anàlisi i presentació dels resultats, redacció de la metodologia del manuscrit (conjuntament amb Carla Roca).

Article publicat 3: ***vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines***

Elisabet Munté, Lidia Feliubadaló, Marta Pineda, Eva Tornero, Maribel Gonzalez, José Marcos Moreno-Cabrera, Carla Roca, Joan Bales Rubio, Laura Arnaldo, Gabriel Capellá, Jose Luis Mosquera* i Conxi Lázaro*.

Revista: *Bioinformatics*; volum 39; número 3.

Data: Març 2023; doi: 10.1093/bioinformatics/btad128

IF 2022 =5,8 ; Categoria: *Biochemical research methods* -SCIE; Q1/D2; Rang= 8/77

Contribució de la doctoranda: revisió de les diferents bases de dades per classificar variants, extracció de resultats de les bases de dades, disseny i preparació del paquet, validació dels resultats, manteniment del paquet, comparació amb Cancer Sigvar, escriptura del manuscrit.

Resultat no publicat 1: ***Optimizing GRIDSS for clinical use: a targeted NGS filtering strategy for germline structural variant detection***

Elisabet Munté[¥], Paula Rofes[¥]**,** Miriam Millán-Castillo, Ares Solanes, Xavier Muñoz, Olga Campos, Mònica Salinas, Raquel Cuesta, Lídia Feliubadaló, Jesús del Valle* i Conxi Lázaro*.

[¥]Primera autoria compartida.

Revista: No publicat

Contribució de la doctoranda: Cerca bibliogràfica de programes identificadors de SVs. Selecció de GRIDSS. Analitzar les mostres per GRIDSS i RepeatMasker, disseny dels filtres per prioritzar les variants d'interès, aplicació dels filtres al conjunt de dades i revisió visual de totes les variants seleccionades. Participació en la redacció del manuscrit (conjuntament amb Paula Rofes).

Resultat no publicat 2: ***Identifying potential pathogenic variants in 5'UTR regions within a hereditary cancer cohort***

Elisabet Munté, Lidia Feliubadaló, Alexandra Martin-Geary, Conxi Lázaro, Nicola Whiffin

Revista: No publicat

Contribució de la doctoranda: estada formativa al grup Computational Rare Disease Genomics dirigit per la professora Whiffin (Oxford), revisió bibliogràfica de literatura, re-anàlisi de les mostres per incloure les regions 5'UTR en la crida de les variants. Aplicació dels filtres descrits en la literatura per prioritzar les variants, anàlisi dels resultats, redacció del manuscrit.

Conxi Lázaro, PhD

Lídia Feliubadaló, PhD

ARTICLES

Els resultats d'aquesta tesi doctoral es presenten en cinc articles (tres publicats i dos no publicats) en els quals la doctoranda és primera autora. Addicionalment, també s'ha col·laborat en l'anàlisi bioinformàtica d'altres estudis relacionats amb el diagnòstic del càncer hereditari, els articles resultants dels quals s'han adjuntat com a annexos d'aquesta tesi doctoral.

Resultats considerant els objectius de la tesi:

Per tal de garantir la claredat, la secció de resultats s'estructura d'acord amb els objectius específics.

1. Desenvolupar solucions bioinformàtiques per millorar l'anàlisi de les dades de seqüenciació massiva i facilitar la identificació de variants potencialment no detectades amb les anàlisis prèvies, amb la finalitat d'aplicar aquestes estratègies tant de manera retrospectiva com prospectiva. En concret s'ha treballat en els següents aspectes:
    1.1. Implementar un codi per millorar la precisió en la detecció de variants de *PMS2* utilitzant dades de panells NGS de rutina i reduir el nombre de mostres que s'han d'estudiar mitjançant PCR llarga (LR-PCR).

Article publicat 1: ***Open-Source Bioinformatic Pipeline to Improve PMS2 Genetic Testing Using Short-Read NGS Data***

Elisabet Munté, Lídia Feliubadaló, Jesús DelValle, Sara González, Mireia Ramos-Muntada, Judith Balmaña, Teresa Ramon y Cajal, Noemí Tuset, Gemma Llort, Juan Cadiñanos, Joan Brunet, Gabriel Capellá, Conxi Lázaro* i Marta Pineda*.

Revista: The Journal of Molecular Diagnostics; volum 26; número 8.

Data: Agost 2024; doi: 10.1016/j.jmoldx.2024.05.005

1.2. Avaluar una bateria d'eines dissenyades per a la detecció de CNVs en dades de panell

Article publicat 2: ***Detection of germline CNVs from gene panel data: benchmarking the state of the art***

Elisabet Munté[¥], Carla Roca[¥], Jesús del Valle, Lídia Feliubadaló, Marta Pineda, Bernat Gel, Elisabeth Castellanos, Bárbara Rivera, David Cordero, Víctor Moreno, Conxi Lázaro* i José Marcos Moreno-Cabrera*.

[¥]Primera autoria compartida.

Revista: Briefings in Bioinformatics; volum 26; número 1

Data: Gener 2025 ; doi: 10.1093/bib/bbae645

1.1. Identificar tipus de variants que actualment es poden estar perdent amb els *pipelines* utilitzats.

Resultat no publicat 1: ***Optimizing GRIDSS for clinical use: a targeted NGS filtering strategy for germline structural variant detection***

Elisabet Munté[¥], Paula Rofes[¥], Míriam Millán-Castillo, Ares Solanes, Xavier Muñoz, Olga Campos, Mònica Salinas, Raquel Cuesta, Lídia Feliubadaló, Jesús del Valle* i Conxi Lázaro*.

[¥]Primera autoria compartida.

1.3. Explorar les regions 5'UTR i prioritzar a nivell bioinformàtic possibles variants reguladores de l'expressió gènica

Resultat no publicat 2: ***Identifying potential pathogenic variants in 5'UTR regions within a hereditary cancer cohort***

Elisabet Munté, Lidia Feliubadaló, Alexandra Martin-Geary, Conxi Lázaro i Nicola Whiffin

2. Optimitzar el procés de classificació de variants mitjançant automatització d'alguns passos:
   2.1. Semi-automatitzar i ponderar l'assignació del màxim nombre d'evidències ACMG, amb criteris adaptats a cada gen, per agilitzar la classificació i reduir el possible error manual.
   2.2. Implementar l'opció d'introduir manualment i enregistrar les evidències que no es poden automatitzar.
   2.3. Generar una classificació de la variant basada en el conjunt d'evidències, que segueixi els algoritmes definits per les guies internacionals.
   2.4. Produir una eina bioinformàtica en format d'accés obert per la comunitat científica.

Article publicat 3: ***vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines***

Elisabet Munté, Lidia Feliubadaló, Marta Pineda, Eva Tornero, Maribel Gonzalez, José Marcos Moreno-Cabrera, Carla Roca, Joan Bales Rubio, Laura Arnaldo, Gabriel Capellá, Jose Luis Mosquera* i Conxi Lázaro*.

Revista: Bioinformatics; volum 39; número 3.

Data: Març 2023; doi: 10.1093/bioinformatics/btad128

Article publicat 1

Open-Source Bioinformatic Pipeline to Improve *PMS2* Genetic Testing Using Short-Read NGS Data

Elisabet Munté, Lídia Feliubadaló, Jesús Del Valle, Sara González, Mireia Ramos-Muntada, Judith Balmaña, Teresa Ramon y Cajal, Noemí Tuset, Gemma Llort, Juan Cadiñanos, Joan Brunet, Gabriel Capellá, Conxi Lázaro* i Marta Pineda*.

J Mol Diagn. 2024 Aug;26(8):727-738

DOI: 10.1016/j.jmoldx.2024.05.005

The Journal of Molecular Diagnostics, Vol. 26, No. 8, August 2024

ELSEVIER

the Journal of Molecular Diagnostics

jmdjournal.org

Open-Source Bioinformatic Pipeline to Improve *PMS2* Genetic Testing Using Short-Read NGS Data

Check for updates

Elisabet Munté,[*†] Lídia Feliubadaló,[*†‡] Jesús Del Valle,[*†‡] Sara González,[*†] Mireia Ramos-Muntada,[†‡] Judith Balmaña,[§¶] Teresa Ramon y Cajal,[‖] Noemí Tuset,[**] Gemma Llort,[††] Juan Cadiñanos,[‡‡] Joan Brunet,[*‡§§] Gabriel Capellá,[*†‡] Conxi Lázaro,[*†‡] and Marta Pineda[*†‡]

From the Hereditary Cancer Program, Catalan Institute of Oncology, L'Hospitalet de Llobregat; Hereditary Cancer Group,[†] Molecular Mechanisms and Experimental Therapy in Oncology Program, Institut d'Investigació Biomèdica de Bellvitge, L'Hospitalet de Llobregat; Ciber Oncología,[‡] Instituto Salud Carlos III, Madrid; the High Risk and Cancer Prevention Group,[§] Vall d'Hebron Institute of Oncology, Barcelona; the Medical Oncology Department,[¶] University Hospital of Vall d'Hebron, Barcelona; the Familial Cancer Clinic,[‖] Medical Oncology Service, Hospital Sant Pau, Barcelona; the Medical Oncology Department,** Hospital Universitari Arnau de Vilanova, Lleida; the Department of Medical Oncology Parc Taulí,[††] Hospital Universitari Parc Taulí Sabadell, Barcelona; the R&D Laboratory,[‡‡] Fundación Centro Médico de Asturias—IMOMA, Oviedo; and the Precision Oncology Group (OncoGir_Pro),[§§] Institut d'Investigació Biomèdica de Girona, Girona, Spain*

Accepted for publication May 13, 2024.

Address correspondence to Marta Pineda, Ph.D., or Conxi Lázaro, Ph.D., Catalan Institute of Oncology, L'Hospitalet de Llobregat, Hereditary Cancer Program, Avinguda de la Granvia de l'Hospitalet, 199, L'Hospitalet de Llobregat, 08908 Barcelona, Spain. E-mail: mpineda@iconcologia.net or clazaro@iconcologia.net.

The molecular diagnosis of mismatch repair—deficient cancer syndromes is hampered by difficulties in sequencing the *PMS2* gene, mainly owing to the PMS2CL pseudogene. Next-generation sequencing short reads cannot be mapped unambiguously by standard pipelines, compromising variant calling accuracy. This study aimed to provide a refined bioinformatic pipeline for *PMS2* mutational analysis and explore *PMS2* germline pathogenic variant prevalence in an unselected hereditary cancer (HC) cohort. *PMS2* mutational analysis was optimized using two cohorts: 192 unselected HC patients for assessing the allelic ratio of paralogous sequence variants, and 13 samples enriched with *PMS2* (likely) pathogenic variants screened previously by long-range genomic DNA PCR amplification. Reads were forced to align with the *PMS2* reference sequence, except those corresponding to exon 11, where only those intersecting gene-specific invariant positions were considered. Afterward, the refined pipeline's accuracy was validated in a cohort of 40 patients and used to screen 5619 HC patients. Compared with our routine diagnostic pipeline, the PMS2_vaR pipeline showed increased technical sensitivity (0.853 to 0.956, respectively) in the validation cohort, identifying all previously *PMS2* pathogenic variants found by long-range genomic DNA PCR amplification. Fifteen HC cohort samples carried a pathogenic *PMS2* variant (15 of 5619; 0.285%), doubling the estimated prevalence in the general population. The refined open-source approach improved *PMS2* mutational analysis accuracy, allowing its inclusion in the routine next-generation sequencing pipeline streamlining *PMS2* screening. *(J Mol Diagn 2024, 26: 727—738; https://doi.org/10.1016/j.jmoldx.2024.05.005)*

Supported by Fundació La Marató de Televisió de Catalunya (TV3) grant 202028-30; Fundación Mutua Madrileña grant AP183612023; Ministerio de Ciencia e Innovación (PID2019-111254RB-100); Carlos III National Health Institute, funded by Fondo Europeo de Desarrollo Regional (FEDER) "A way to build Europe" (PI23/00017 and PI19/00553); Carlos III National Health Institute and the European Union (EU) Next Generation EU/Mecanismo para la Recuperación y la Resiliencia (PMP22/00064); Centro de Investigación Biomédica en Red Cáncer (CIBERONC) grant CB16/12/00234; and Department of Research and Universities of the Generalitat de Catalunya and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (2023SGR01112). Also supported by the Acció Instrumental de Formació de Científics i Tecnòlegs grant SLT017/20/000129 of the Departament de Salut de la Generalitat de Catalunya. The CERCA program/Generalitat de Catalunya and Fundación María Cristina Masaveu Peterson provided institutional support.

C.L. and M.P. contributed equally to this work.

Copyright © 2024 Association for Molecular Pathology and American Society for Investigative Pathology. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (*http://creativecommons.org/licenses/by-nc-nd/4.0*).
https://doi.org/10.1016/j.jmoldx.2024.05.005

Lynch syndrome (LS) is a common, dominantly inherited, cancer-predisposing condition caused by germline pathogenic variants affecting the function of mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*).[1] Despite its incomplete penetrance, individuals harboring an MMR pathogenic variant have increased chances of developing colorectal and endometrial cancers, among others.[2] Biallelic pathogenic alterations damaging the same MMR genes cause constitutional mismatch repair deficiency (CMMRD), a very rare (1 in 1,000,000) and severe condition that predisposes to multiorgan cancers—mainly brain, hematologic, and colorectal—usually with childhood onset.[3–5] Its penetrance is more than 90% at the age of 20.

The estimated population frequency of pathogenic *PMS2* variant carriers is the highest among the four MMR genes (1 in 714; 0.140%).[1] Accordingly, *PMS2* is the most frequently mutated gene in CMMRD syndrome, accounting for nearly 60% of cases.[5] In contrast, *PMS2* is the least frequently mutated gene in the LS series, probably owing to its lowest penetrance in heterozygous carriers and the former use of clinical criteria for LS tumor screening. Nevertheless, the cancer risk varies widely even among heterozygous carriers from the same family.[2,6,7]

Gene panels using targeted next-generation sequencing (NGS) of short reads are the tests most used in the field of hereditary cancer (HC) because of their optimal balance between cost and benefit. However, short-read–based NGS has significant limitations in the identification of variants in complex regions.[8] Indeed, *PMS2* gene analysis presents a major challenge mainly because of the existence of multiple pseudogenes.[9,10] Specifically, there are 14 pseudogenes located at the 5′ end, spanning exons 1 through 5, and an additional 15th pseudogene located at the 3′ end, known as PMS2CL. Remarkably, the PMS2CL pseudogene is an inverted partial duplication located on the same chromosome 7 that exhibits notable sequence homology (>98% identity) with exons 9 and 11 to 15 of the *PMS2* gene. Some bases, called paralogous sequence variants (PSVs), differ in *PMS2*-PMS2CL reference sequence.[11,12] It has been proven that sequence exchange (recombination and gene conversion) is a frequent event observed between these two loci.[13,14] This makes it difficult to discriminate reliably whether an identified variant is located in the gene or the pseudogene.[13,14]

Genomic DNA long-range PCR (LR-PCR) amplification and gene-specific cDNA amplification using primers located in less-homologous regions, and DNA/cDNA long-read sequencing, can analyze *PMS2* specifically.[7,15–20] Nevertheless, these techniques are labor-intensive, complicate routine diagnostic workflows, and present many technical challenges, which question the feasibility of implementing them in large cohorts. Bioinformatic approaches partially can palliate these difficulties. In this sense, Gould et al[11] proposed a workflow in which gene and pseudogene variants were forced to align with the *PMS2* gene reference sequence. By this means, they identified seven PSV positions in *PMS2* exon 11, where none of the 707 ethnically diverse patients from their cohort differed from the gene and pseudogene reference sequences. These positions, hereinafter referred to as invariant PSVs, were demonstrated to be useful in identifying the origin of variants identified in NGS reads overlapping them.[11] Despite these advances, to our knowledge, there is no free open-source pipeline available to analyze *PMS2* accurately. Thus, the inclusion of the *PMS2* gene in routine NGS diagnostic pipelines remains a challenge for most genetic testing laboratories.

To address this need, PMS2_vaR is presented, the first free open-source pipeline written in R, which integrates and upgrades the previously reported strategy. The aim of this study is to increase the accuracy of *PMS2* variant detection using routine NGS panel data, in addition to reducing the number of samples that need to undergo LR-PCR. This study also aimed to assess the prevalence of *PMS2* pathogenic variants in an HC cohort upon implementation.

Materials and Methods

Study Cohorts

Table 1 and Supplemental Figure S1 provide an overview of the cohorts used.

The PMS2_vaR pipeline was optimized using samples from two cohorts: optimization cohort A, comprising 192 HC patients used to assess the allelic ratio of PSVs in unselected samples; and optimization cohort B, enriched in samples harboring PMS2 (likely) pathogenic variants, composed of 13 cancer patients in whom blood DNA was analyzed previously by *PMS2* LR-PCR, enabling the identification of *PMS2* variants. For validation purposes, a LS suspicion cohort of 40 patients analyzed previously by PMS2 LR-PCR was used to determine the pipeline's accuracy.

Finally, a large HC cohort of 5619 patients was studied. According to clinical phenotypes, the cohort comprised 13 LS-suspected patients showing exclusive PMS2 loss of expression in tumors (blood samples were not analyzed previously by LR-PCR), 36 patients diagnosed with early onset colorectal cancer (age, <50 years at diagnosis) showing MMR-conserved protein expression or with no available immunohistochemistry information, 798 patients fulfilling other LS suspicion criteria (Amsterdam criteria or MMR expression loss but not exclusively of PMS2), and 4772 patients tested for suspicion of other HC syndromes.

Sample Collection and Preprocessing

DNA samples were obtained from peripheral blood leukocytes of individuals with HC suspicion referred to the Molecular Diagnostics Service at the Institut Català d'Oncologia from its network of genetic counseling units. Informed written consent for both diagnostic and research purposes was obtained from this cohort of patients. The study protocol was approved by the Ethics Committee of the Catalan Institute of Oncology—Bellvitge University Hospital (PR278/19).

Table 1 Summary of Cohorts Analyzed

Group	Subgroup	Clinical and molecular criteria	*n*	Recommendation for *PMS2* LR-PCR analysis	Purpose
Optimization cohort	Cohort A	HC suspicion, unselected	192	No	Determine allele ratio
	Cohort B	LS suspicion; IHC: PMS2^{-}; *PMS2* LR-PCR analysis previously performed Enriched in PMS2 (L)PAT variants	13	Previously performed	Determine pipeline accuracy
Validation cohort		LS suspicion; *PMS2* LR-PCR analysis previously performed	40	Previously analyzed	Determine pipeline accuracy
Hereditary cancer cohort		LS suspicion; IHC: PMS2^{-}; *PMS2* LR-PCR analysis not performed	13	Yes, in samples with an identified *PMS2* (likely) pathogenic variant*	Determine prevalence in this subgroup
		Early onset CRC suspicion; IHC: conserved expression	36	Yes, in samples with an identified *PMS2* (likely) pathogenic variant*	Determine prevalence in this subgroup
		LS suspicion (Amsterdam criteria or IHC MMR expression loss excluding *PMS2*)	798	Yes, in samples with an identified *PMS2* (likely) pathogenic variant*	Determine prevalence in this subgroup
		Other HC suspicions	4772	Yes, in samples with an identified *PMS2* (likely) pathogenic variant*,†,‡	Determine prevalence in this subgroup

*Pathogenic paralogous sequence variants (c.1864_1865del and c.1730dup) will be considered for LR-PCR testing only if the variant allele frequency is >60%.

†Pseudogenic exon 13 c.2186_2187del and c.2243_2246del *PMS2* variants will be considered for LR-PCR testing only if the tumor molecular characteristics are indicative of a *PMS2* alteration (microsatellite instability or exclusive IHC PMS2 loss) or when IHC analysis is not possible.

‡(Likely) pathogenic variants called by the general approach but filtered out after the refined E11 approach will not be tested unless the tumor molecular characteristics are indicative of a *PMS2* alteration (microsatellite instability or exclusive IHC PMS2 loss) or when IHC analysis is not possible.

CRC, colorectal cancer; HC, hereditary cancer; IHC, immunohistochemistry; (L)PAT, (likely) pathogenic; LR-PCR, long-range PCR; LS, Lynch syndrome; MMR, mismatch repair.

Routine Diagnostics Pipeline

Genetic testing was conducted on peripheral blood DNA using NGS custom panel ICO-IMPPC Hereditary Cancer Panel (I2HCP).[21] This panel encompasses a comprehensive selection of 122 to 168 genes (depending on the version used) associated with HC susceptibility. The bioinformatics approach used for the routine diagnostics pipeline was described previously.[21,22] The selection of genes for analysis was based on the phenotype of each patient[23] and their family, following the Catalan Health Service guidelines.

PSV Determination

A list of 31 exonic base differences was obtained by comparing the *PMS2* (NM_000535.7; *https://www.ncbi.nlm.nih.gov/nuccore/1519311653*, last accessed May 15, 2024) and PMS2CL (NR_002217.1; *https://www.ncbi.nlm.nih.gov/nuccore/NR_002217.1*, last accessed May 15, 2024) sequences using the BlastN tool from the National Center for Biotechnology Information. Because some variants were consecutive, they were considered compound variants, thus the final list was 28 PSVs, 23 of which were located within exon 11. Supplemental Table S1 contains the complete list of all 28 PSVs.[12]

Bioinformatic Pipeline Development

The PMS2_vaR pipeline was conceived for the R statistical computing environment (v4.2.1; *https://github.com/emunte/PMS2_vaR*, last accessed March 22, 2024). It requires the installation of the following software: SAMtools (v1.10; *https://www.htslib.org*), Picard (v.2.26.4.jar; *https://github.com/broadinstitute/picard*), BWA (0.7.17; *https://github.com/lh3/bwa*), and VarDictJava (v1.8.3; *https://github.com/AstraZeneca-NGS/VarDictJava*). It also uses functions from both R/Bioconductor and CRAN packages (*https://cran.r-project.org*; see the required libraries in the GitHub space).

The pipeline consists of two scripts: modify_reference and run_PMS2_vaR (Figure 1).

modify_reference Script

Given a human reference genome sequence FASTA file and its PMS2CL FASTA sequence, the workflow generates a PMS2CL-masked reference genome in which the PMS2CL genomic sequence is replaced by Ns (any base). This file is needed as an input file for the Run_PMS2_vaR algorithm. This step only needs to be executed once (per human reference genome version).

run_PMS2_vaR Script

To feed the pipeline, the user is required to provide several input files, including a text file containing paths to the BAM

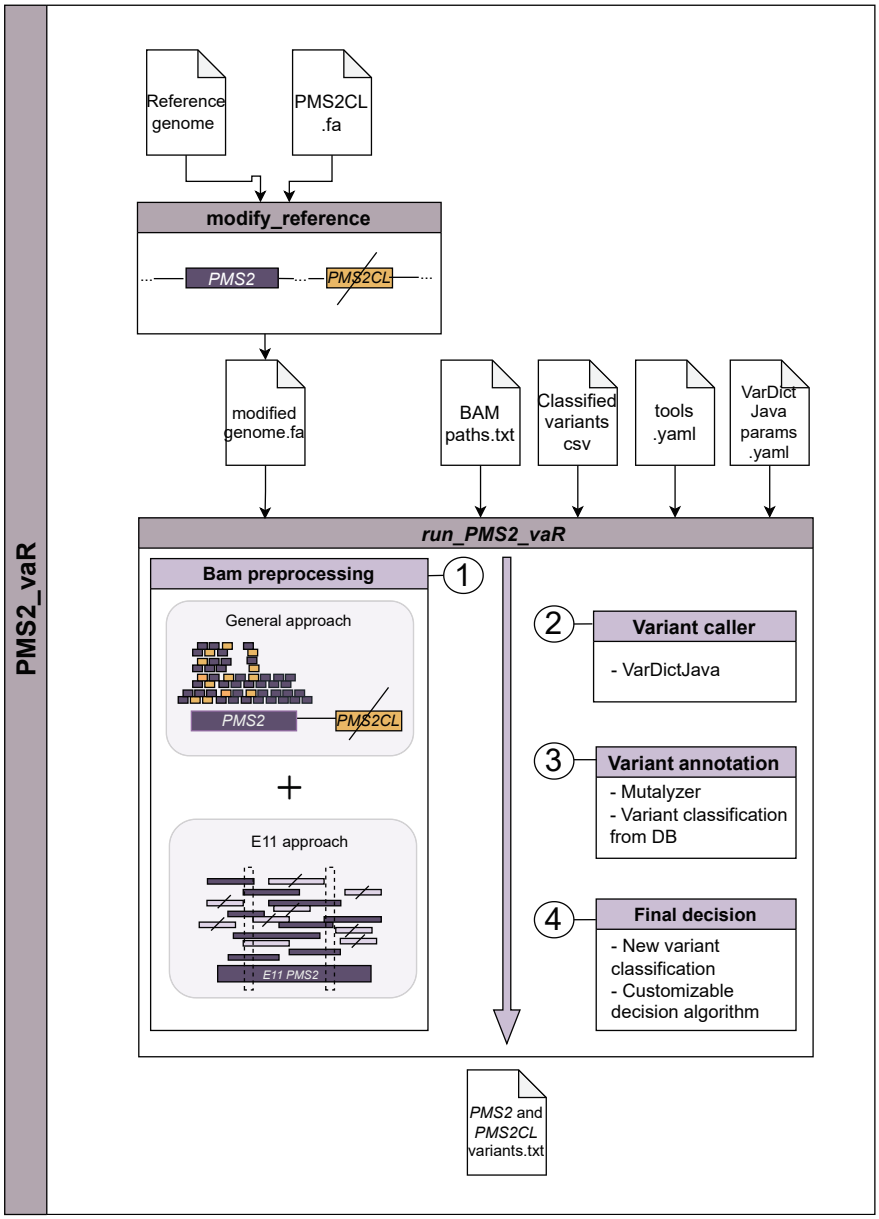


Figure 1 Schematic representation of the PMS2_vaR pipeline. The modified reference genome sequence file is obtained using the "modify_reference" script. Given a BAM file, a classified variants csv file and the yaml files specifying the tool's paths and parameters for VardictJava, the run_PMS2_vaR script produces a text file containing all *PMS2* and PMS2CL variants. The steps to construct the final *PMS2* candidate variants list are detailed in *run_PMS2_vaR Script*. .fa, FASTA file.

files, a yaml file detailing the paths to the necessary tools (SAMtools, Picard, BWA, and VarDictJava), another yaml file specifying the parameters to be used with VarDictJava, and a comma delimited file listing classified *PMS2* variants. The template for these files is available at *https://github.com/emunte/PMS2_vaR*.

In the general approach, to obtain the list of candidate variants that need to be validated further by LR-PCR, gene and pseudogene reads in the highly homologous regions were aligned with the PMS2CL-masked human genome reference sequence. To this end, first, reads aligning with *PMS2* or PMS2CL in the standard pipeline BAM were selected using SAMtools. The resulting BAM file was transformed into paired-end FASTQ files using Picard software. Afterward, the FASTQ files were realigned with the modified reference genome using BWA-MEM. The SAM file was converted to a BAM file, sorted, and indexed using SAMtools.

Subsequently, exon (E)11 was analyzed based on the approach of Gould et al[11] (hereinafter called the E11 approach). Following their recommendations, only reads that intersected with any of the seven invariant PSVs were included (Supplemental Table S1). Read names overlapping the corresponding positions were obtained using SAMtools and the reads then were filtered by name using Piccard. These were aligned to the standard (nonmasked) reference genome. The resulting E11 BAM was merged with the BAM obtained for the other exons with the general approach.

Variant calling was performed for both approaches using VarDictJava. The following parameters were modified: i)

Table 2 Primer Sequences for *PMS2* Amplification

Target	Template	Forward primer	Reverse primer
LR1 (exons 1—5)	gDNA	5′-ACGTCGAAAGCAGCCAATGGGAGTT-3′*,†	5′-CTTCCACCTGTGCATACCACAGGCT-3′*,†
LR2 (exons 7—9)	gDNA	5′-GGTCCAGGTCTTACATGCATACTGT-3′*,†	5′-CTGACTGACATTTAGCTTGTTGACA-3′*,†
LR3 (exons 11—15)	gDNA	5′-GCGTTGATATCAATGTTACTCCAGA-3′*,†	5′-CCTTCCATCTCCAAAACCAGCAAGA-3′*,†
Exon 1	LR1	5′-M13F-ACGTCGAAAGCAGCCAATGGGAGTT-3′*,†	5′-M13R-CAGGTAGAAAGGAAATGCATTCAGT-3′*,†
Exon 2	LR1	5′-ACAGTGTTGAGTCATTTCCCACAGT-3′*,†	5′-TTCTTAGCATAACACCTGCCTGGCA-3′*,†
Exon 3—4	LR1	5′-M13F-CTGGGCTAGTAAATAGCCAGAAAG-3′†	5′-M13R-TATGACTTAGATTGGCAGCGAGACA-3′*,†
Exon 5	LR1	5′-M13F-CTTGATTATCTCAGAGGGATCGTCA-3′*,†	5′-M13R-TCTCACTGTGTTGCCCAGTCCTAAT-3′*,†
Exon 6	gDNA	5′-M13F-TGCTTCCCTTGATTTGTGCGATGAT-3′*,†	5′-M13R-CATTCTACTGGAAGGGACAATGGA-3′
Exon 7	gDNA	5′-M13F-ACCCACGAGTTTGACATTGCAGTGA-3′*	5′-M13R-AAAAGACACGAAACTATTAGCCTTAGA-3′
Exon 8	gDNA	5′-M13F-AGATTTGGAGCACAGATACCCGTGA-3′*,†	5′-M13R-TGCGGTAGACTTCTGTAAATGCACA-3′*,†
Exon 9	LR2	5′-M13F-CCTTCTAAGAACATGCTGGTTGGTT-3′*,†	5′-M13R-ATCTCATTCCAGTCATAGCAGAGCT-3′*,†
Exon 10	gDNA	5′-M13F-AATTAGCCAGTGTGGTGGCACTTG-3′†	5′-M13R-AGCTTTAGAAGCTGTTTGTACAC-3′†
Exon 11a	LR3	5′-M13F-TCACATAAGCACGTCCTCTCACCAT-3′*,†	5′-M13R-GAATGGCAGTCCACATCTGAAAAAG-3′
Exon 11b	LR3	5′-M13F-CAGAGCGGAGGTGGAGAAGGAC-3′	5′-M13R-GTGAAACCCTGTTTCCACCAAAAAT-3′
Exon 12	LR3	5′-M13F-GCCAAGATTGTGCCATTGCACTGTA-3′*	5′-M13R-AGTAGATACAAGGTCTTGCTGTGTT-3′*,†
Exon 13	LR3	5′-M13F-TTGTTTTCATTTCATTTCTGCTG-3′	5′-M13R-ATGTTAGCCAGGCTGGTCTCAAACT-3′*,†
Exon 14	LR3	5′-M13F-GCTTTCAAGTGAAACGTGTTTGTCA-3′	5′-M13R-GCACGTAGCTCTCTGTGTAAAATGA-3′*,†
Exon 15	LR3	5′-M13F-GCTGAGATCTAGAACCTAGGCTTCT-3′*,†	5′-M13R-ACACACGAGCGCATGCAAACATAGA-3′*,†

*Primers from Clendenning et al.[15]
†Primers from Vaughn et al.[16]
LR, long range; gDNA, genomic DNA; M13F sequence, 5′-TGTAAAACGACGGCCAGT-3′; M13R sequence, 5′-CAGGAAACAGCTATGACC-3′.

the minimum allele frequency was set to 0.1 to accommodate the factual tetra-allelic situation because the new alignments (without the PMS2CL sequence in the reference) combine four alleles in *PMS2* E9, 11 to 15, and nearby positions (Supplemental Figure S2); ii) the region of interest (-R) was set to chr7:6012350-6049257 for hg19 and chr7:5972719-6009626 for hg38; iii) the minimum phred score (-q) was set to 15; and iv) the number of mismatches allowed in a read (-m) was set to 10 for greater permissiveness. In addition to variants with all filters passed, variants tagged for mean mismatches in reads $\geq$5.25 (NM5.25) or for being adjacent to an insertion variant (InIns) were kept. In a diploid variant calling situation, these filters would point to likely false-positive variants, however, this study tried to be conservative.

The two variant calling files were converted into txt files using the vcfR R package and were merged into the same document. This allows the user to verify whether the variant was found by one or both approaches. The pipeline followed the decision algorithm described in *Results* to suggest whether LR-PCR should be performed or not for each variant.

PMS2 Variant Validation by LR-PCR

Candidate variants in *PMS2* (NM_000535.7, NG_008466.1, *https://www.ncbi.nlm.nih.gov/nuccore/NG_008466.1*, last accessed May 15, 2024) identified by vaR_PMS2 were analyzed using previously described LR-PCR procedures.[15,16] A schematic representation detailing the annealing positions of all the primers used can be found in Supplemental Figure S3. In brief, amplicons encompassing entire exons 1 to 5 (long-range amplicon LR1), 9 (LR2), and 11 to 15 (LR3) were generated using LaTaq polymerase (TaKaRa Bio, Inc, Otsu, Shiga, Japan) and the corresponding primers are listed in Table 2. Amplification of LR-PCR products was confirmed by agarose gel electrophoresis. The LR3 product was purified by gel extraction to avoid pseudogene amplification from genomic DNA instead of from the LR-PCR product in the following exon-specific PCR. LR-PCR products (or purified products) were diluted 1:10 and 1 μL of this dilution was used as the template for nested exon-specific amplifications. Exon-specific PCRs were performed using DreamTaq DNA polymerase (Thermo Fisher Scientific, Waltham, MA) and the corresponding primers (Table 2). For exons 6, 7, 8, and 10, genomic DNA was used as the PCR template. Amplification was confirmed by agarose gel electrophoresis and PCR products were sequenced using the Big Dye Terminator v.3. 1 Cycle Sequencing kit (Applied Biosystems, Waltham, MA) and an Applied Biosystems 3130XL Genetic Analyzer.[15,16]

Variant Classification

A list of 129 *PMS2* classified variants was provided to feed the pipeline (Supplemental Table S2). Variants initially were classified using the vaRHC R package[24] and subsequently curated by the Catalan Institute of Oncology

Hereditary Cancer Molecular Diagnostics Service. The draft version of the InSiGHT-ClinGen–specific MMR variant classification guidelines were followed (*https://www.insight-group.org/content/uploads/2021/11/DRAFT_Nov_2021_TEMPLATE_SVI.ACMG_Specifications_InSiGHT_MMR_V1.pdf*, last accessed March 22, 2024). Users have the flexibility to incorporate additional classified variants or modify the classification of existing ones, tailoring the system to their specific requirements (see GitHub for further details).

Assessment of Routine and PMS2_vaR Pipelines Performance

The performance of the routine and the PMS2_vaR pipelines was analyzed against the results obtained from LR-PCR in both the optimization and validation cohorts. A comprehensive set of performance metrics was computed, including accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The McNemar test was used to determine significance, with a predefined P value of 0.05. Statistical analyses were conducted using R v.4.2.2, leveraging the CRAN package DTComPair v.1.2.2 (*https://CRAN.R-project.org/package=DTComPair*). To assess the reduction in long and short PCRs [(LR)-PCR] workload achieved by using PMS2_vaR, the location of each candidate variant identified in the HC *PMS2* cohort was examined to determine the precise PCR reaction required for validation. This result was compared with the total number of PCR reactions needed in the *PMS2* analysis by (LR)-PCR.

Results

Bioinformatic Pipeline Development: Data Processing and Variant Calling

A bioinformatic pipeline was developed to identify *PMS2* variants from multigene panel NGS data with high accuracy. First, reads aligning to *PMS2* or PMS2CL were converted into FASTQ files and realigned with a human reference genome with the PMS2CL sequence masked. This forced both *PMS2* and PMS2CL reads to map to the *PMS2* reference. Subsequently, the PSV variant allele frequency (VAF) was assessed from a cohort of unselected samples (optimization cohort A). Given that PSVs are positions where gene and pseudogene reference sequences diverge, it would be expected that when one of these variants is called, the reads that support it come from the pseudogene. Consequently, PSVs would have an expected VAF of approximately 50% (present in two of four alleles). However, in the analysis of samples of the optimization cohort A, the observed VAF ranged from 35% to 45% for most PSVs (Supplemental Figure S4). This reduction suggested that the probes have a slightly weaker affinity for PMS2CL regions harboring PSVs according to the reference genome. Moreover, some PSVs deviated strongly from the expected proportions: three exhibited a VAF below 25% across multiple samples, indicating that these variants likely are pseudogene polymorphisms, and another three displayed VAFs exceeding 60%, suggesting that they likely are gene polymorphisms (Supplemental Figure S4).

In addition to this general approach, a refined method for analyzing exon 11 subsequently was introduced, including only reads overlapping invariant PSVs for this exon. Results from the two approaches were integrated into the same data frame.

Decision-Making Algorithm

An algorithm was designed to assess the presence of variants in MMR genes and to recommend if a called *PMS2* variant would need confirmation by (LR-)PCR analysis (Figure 2). For each *PMS2* candidate variant, the algorithm first assessed if the variant passed the quality filters and if it was in a region of interest (this study's setting was a coding region $\pm$ 20 bp). Next, it checked if it was a PSV. Among the 28 exonic PSVs listed (Supplemental Table S1), 2 were classified as (likely) pathogenic variants if located in the gene (c.1730dup and c.1864_1865del). These two variants were called in all samples from the optimization cohort A and, in most cases, they corresponded to the pseudogene reference sequence rather than being a gene variant.[11] To avoid the need for LR-PCR analysis of each sample, paralogous variants were regarded exclusively as potential gene candidates if their VAF exceeded 60% (Table 1 and Figure 2). The selected threshold presumes that candidate variants at PSV positions should be present in at least three of the four alleles of *PMS2* and PMS2CL (Supplemental Figure S2).

For non-PSV variants or those PSVs with VAFs $\geq$60%, the algorithm examined whether the variant was classified or not. If the variant was considered pathogenic or likely pathogenic and also was called with the E11 approach, LR-PCR was recommended (Figure 2). However, in samples with any called *PMS2* pathogenic variant discarded by the E11 approach, LR-PCR analysis would be recommended only when the family phenotype strongly indicated Lynch or CMMRD syndromes (ie, loss of PMS2 protein expression or microsatellite instability). Variants of unknown significance did not undergo LR-PCR because they are not currently clinically actionable. They were reported only if the variant was detected with both approaches, with a disclaimer clarifying that they were not validated by LR-PCR. Finally, benign or likely benign variants were not reported.

Apart from PSVs, some common pseudogenic variants also can be found in the gene, albeit at very low population frequencies. Specifically, exon 13 recurrent variants c.2186_2187del (p.Leu729Glnfs*6) and c.2243_2246del (p.Lys748Metfs*19) are of particular interest because they

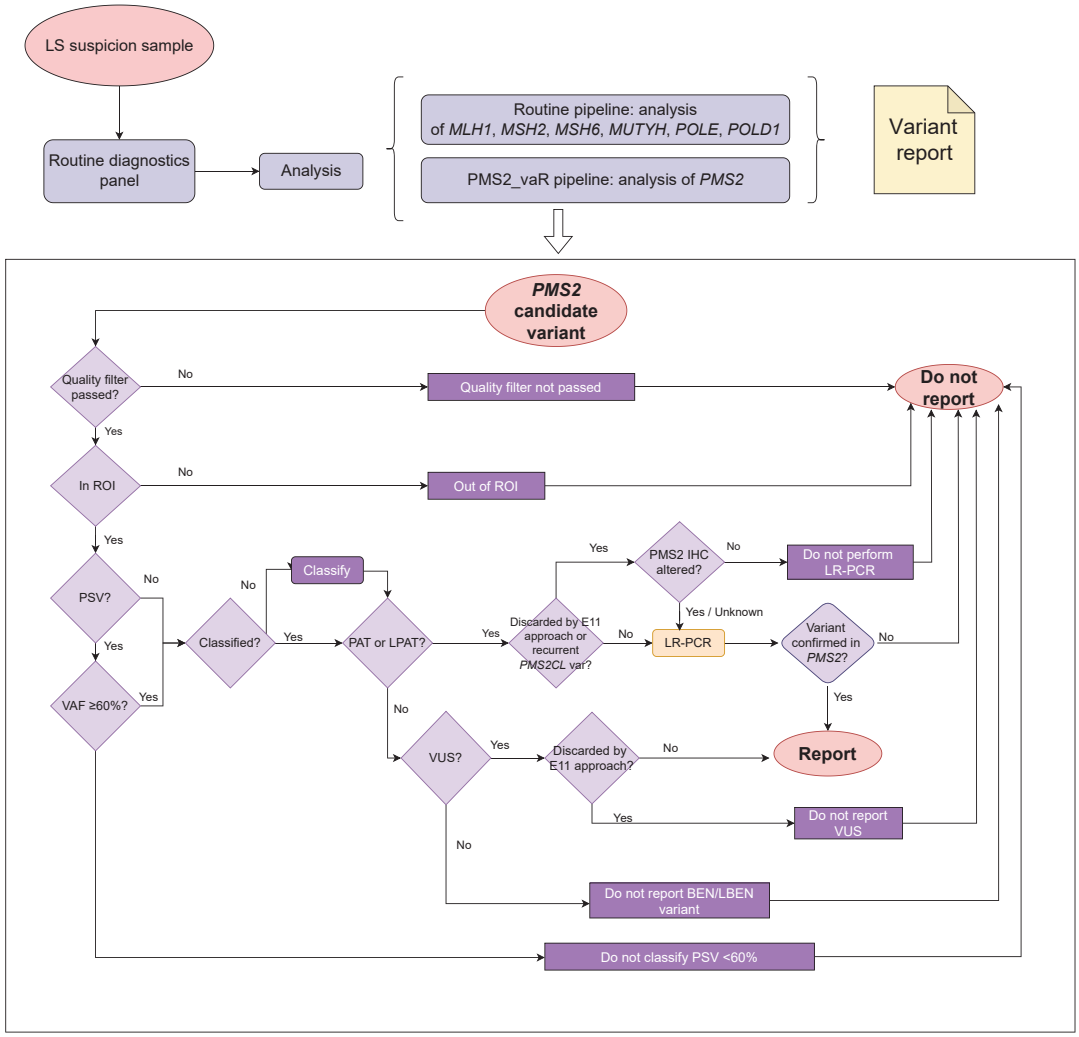


Figure 2 Algorithm used to analyze samples with suspected Lynch syndrome (LS). It assesses the presence of variants in MMR genes using a panel approach. For *PMS2* gene analysis, the PMS2_vaR pipeline indicates when long-range (LR)-PCR analysis should be recommended to confirm a *PMS2* called variant. Recurrent PMS2CL pathogenic variants c.2186_2187del and c.2243_2246del will be examined only if PMS2 expression is lost in tumors. BEN, benign; IHC, immunohistochemistry; LBEN, likely benign; LPAT, likely pathogenic; PAT, pathogenic; PSV, paralogous sequence variant; ROI, region of interest; VAF, variant allele frequency; VUS, variant of unknown significance.

attain a (likely) pathogenic classification within the gene context. The presence of these two variants within the *PMS2* gene was identified in 0.01% (1 of 7593) and 0.07% (2 of 2739) of HC-suspected patients in whom the variants had been called by NGS, respectively.[25] Taking this into account, LR-PCR was performed only on samples that harbored these two recurrent pseudogenic variants when the clinical criteria and tumor molecular characteristics of the carriers indicated a potential *PMS2* alteration or when the VAF was ≥60%.

Assessment of PMS2_vaR Pipeline Performance in the Optimization and Validation Cohorts

To assess the accuracy of the newly developed PMS2_vaR pipeline, the optimization cohort B, including

Table 3 Accuracy, Sensitivity, Specificity, Positive Predicted Value, and Negative Predicted Value Obtained by the Previous Routine Diagnostic Pipeline and the Refined Pipeline (General + E11 Approach) in ROIs (±20 bp)

	Accuracy		Sensitivity		Specificity		PPV		NPV	
Pipeline	Op B	Val	Op B	Val	Op B	Val	Op B	Val	Op B	Val
Diagnostics pipeline	0.9993	0.9993	0.9067	0.8528	0.9994	0.9996	0.7391	0.7818	0.9998	0.9997
PMS2_vaR pipeline	0.9994	0.9996	0.9733	0.9561	0.9994	0.9997	0.7449	0.8435	0.9999	0.9999

NPV, negative predictive value; Op B, optimization cohort B; PPV, positive predictive value; ROI, region of interest; Val, validation cohort.

Table 4 Detailed Information of (Likely)-Pathogenic *PMS2* Variants Identified by PMS2_vaR in HC Clinical Phenotype Cohorts

Group	ID	Variant	Protein	Location	Personal phenotype
LS suspicion with exclusive PMS2 IHC^{-}	1	c.312del	p.(Phe104Leufs*8)	E04	ENDO (54 y), CRC (67 y)
	2	c.584C>A	p.(Ser195*)	E06	Sebaceoma
	3	c.706-1G>T	p.?	I06	CRC (44 y)
	4	c.1144+2T>G	p.?	I10	CRC (73 y)
	5	c.1687C>T	p.(Arg563*)	E11	ENDO (48 y)
Early onset CRC with IHC conserved					
Other LS suspicion criteria (Amsterdam criteria or MMR expression loss but not exclusive of PMS2)	6	c.717_723dup	p.(Phe242Hisfs*9)	E07	CRC (49 y), BBC (51 y)
	7	c.904-1G>A	p.?	I08	CRC (43 y)
	8	c.988+1G>A	p.?	I09	ENDO (62 y)
	9	c.1145-1_1145del	p.?	I10—E11	CRC (68 y)
	10	c.1882C>T	p.(Arg628*)	E11	CRC(44 y)
	11	c.1239dup	p.(Asp414Argfs*44)	E11	CRC (37 y)
Other HC syndromes	12	c.137G>T	p.Ser46Ile	E02	OV (66 y)
	13	c.137G>T	p.Ser46Ile	E02	BR (33 y)
	14[†]	c.989-2A>G	p.?	I09*	PAN (68 y)
	15[†]	c.989-2A>G	p.?	I09	LG (18 y) BR (43 y)
	16	c.2341C>T	p.(Gln781*)	E14	BR (49 y), PAN (49 y)

(*table continues*)

Tumors that were not confirmed by medical reports have the suffix _nc (not confirmed). Cancer family history is broken down by first-degree relatives and second- and third-degree relatives. Each bullet point refers to an individual.

[†]These two individuals are family members.

BBC, basocellular carcinoma; BileDuc, bile duct cancer; BL, bladder cancer; BR, breast cancer; CNS, central nervous system; CRC, colorectal cancer; E, exon; ENDO, endometrial cancer; FDR, first-degree relative; HC, hereditary cancer; I, intron; ID, identification; IHC, immunohistochemistry; KID, kidney cancer; LG, lung cancer; LK, leukemia; LS, Lynch syndrome; MMR, mismatch repair; OV, ovarian cancer; PAN, pancreatic cancer; PV, pathogenic variant; SDR, second-degree relative; STO, stomach cancer; TDR, third-degree relative; U, unknown.

samples previously analyzed by LR-PCR, was analyzed. An increased sensitivity in variant identification compared with the routine diagnostic pipeline was found, increasing it from 0.907 (95% CI, 0.841 to 0.972) to 0.973 (95% CI, 0.937 to 1), while maintaining specificity (0.999) (Table 3). This improvement allowed us to identify all pathogenic variants (Supplemental Table S3), but not two benign polymorphic PSVs with a VAF below 60% (Supplemental Figure S5). These variants were ignored intentionally according to the decision-making algorithm, and in agreement with their benign classification, to reduce the number of LR-PCR confirmations needed.

In the analysis of 40 samples from the validation cohort, the PMS2_vaR pipeline improved sensitivity significantly from 0.853 (95% CI, 0.807 to 0.899) to 0.956 (95% CI, 0.930 to 0.983), in comparison with the routine diagnostic pipeline (McNemar test; score = 24; $P = 9.634 \times 10^{-7}$). Again, all pathogenic variants were identified (Table 3 and Supplemental Table S3). As in the optimization cohort B, there were variants (10 in this case) that were not called by the PMS2_vaR pipeline (Supplemental Figure S5), and all

Table 4 (*continued*)

Family history of FDR	Family history of SDR or TDR	MMR expression in proband's tumors	Comments	True variant?	Prevalence
• BR (50 y), ENDO (55 y) • OV (58 y), ENDO (58 y) • PAN (56 y)	No	PMS2^{-}	*BRCA2* PV carrier IHC conserved of the FDR OV cancer	Yes	38.462% (5/13)
No	No	PMS2^{-}		Yes	
• ENDO (46 y) • PR (79 y)	• STO_nc (55 y)	PMS2^{-}		Yes	
No	• LK_nc (3 y)	PMS2^{-}		Yes	0% (0/16)
• ENDO (55 y)	• CRC (U y) • ENDO (75 y)	MSH6 and PMS2^{-}		Yes	0.627% (5/798)
• BR (61 y) • BL (77 y)	• STO (54 y) • CRC (58 y) • CRC (83 y) • CRC_nc (55 y) • STO_nc (37 y)	U	MMR conserved expression in the SDR STO and CRC (58 y) cancers	Pseudogenic	
• BileDuc_nc (58 y)	• ENDO_nc (40 y) • CNS_nc (30 y)	MLH1 and PMS2^{-}		Yes	
• ENDO (58 y) • CRC (61 y)	No	MSH6 and PMS2^{-}		Yes	
No	No	MSH6 and PMS2^{-}		Yes	
No	• BR_nc (55 y) • PAN_nc (55 y)	MSH6 heterogenous expression and PMS2^{-}		Yes	
No	• KID (44 y) • BR (70 y) • BR_nc (55 y) • PAN_nc (63 y) • CRC_nc (80 y)	PMS2 heterogenous expression, MLH1 conserved		Yes	0.105% (5/4772)
No	• CRC_nc (44 y)	Conserved		Yes	
• BR (43 y)	• CRC_nc (60 y)	Conserved		Yes	
• BR (61 y), CRC (64 y) • PAN (68 y)	• CRC _nc (66 y)	Conserved (LG and BR)	MMR conserved expression in FDR CRC and PAN tumors	Yes	
• PR (77 y)	• ENDO_nc (74 y)	Conserved (PAN)	*BRCA2* germline PV carrier (proband)	Yes	

of them corresponded to polymorphic PSVs with a VAF below 60%, classified as benign following Insight-ClinGen MMR-specific guidelines.

Prevalence of *PMS2* Pathogenic Variants in a Hereditary Cancer Cohort

The implementation of the refined PMS2_vaR pipeline in samples from a HC cohort of 5619 patients identified 16 samples harboring a putative (likely)-pathogenic *PMS2* variant (0.285%) (Table 4). Subsequent (LR-)PCR analysis confirmed a *bona fide PMS2* variant in 15 of these 16 cases: five patients harbored tumors showing exclusive PMS2 loss with immunohistochemistry, five patients met other LS suspicion criteria [four displayed tumor DNA mismatch repair protein Msh6 (MSH6)/mismatch repair endonuclease PMS2 (PMS2) loss and one exhibited DNA mismatch repair protein Mlh1 (MLH1)/PMS2 loss], and five individuals were tested for other HC suspicions (PMS2 expression was later reported as heterogeneous in one ovarian tumor and MMR expression was conserved in the remaining four tumors) (Table 4).

Only variant c.904-1G>A in intron 8, found in a patient with an unavailable tumor, was found to be pseudogenic (case

7) (Table 4). Colorectal and stomach cancers of their relatives showed preserved MMR protein expression. The alignment of the region, assessed with the Integrative Genomics Viewer, showed that the variant was in *cis* with PSVs, supporting its pseudogenic origin (Supplemental Figure S6).

Recurrent pseudogenic exon 13 variants, c.2186_2187del and c.2243_2246del, were detected in 39 and 30 samples of the HC cohort, respectively, at a VAF ranging from 12.28% to 34.67% (Supplemental Table S4). None of them had clinical criteria or tumor molecular characteristics suggesting a *PMS2* alteration, thus LR-PCR was not performed according to the proposed algorithm.

Before implementing PMS2_vaR, 3 LR-PCRs and 15 short PCRs of *PMS2* (Supplemental Figure S3) were conducted on each sample exhibiting exclusive loss of PMS2 in immunohistochemistry. Therefore, the analysis of the 13 PMS2-suspected samples of the HC cohort (Table 1) resulted in 39 LR-PCR and 195 short PCRs. With the implementation of PMS2_vaR, only five samples were recommended for PCR analysis (two LR-PCRs and five short PCRs), resulting in a reduction of 95% of LR-PCRs and 99% of short PCRs.

Discussion

Gene panels are used widely for genetic testing purposes in HC. However, they face challenges when detecting variants in genes that share high homology with pseudogenes.[8] LR-PCR using primers outside the highly homologous regions is the gold standard for discriminating these cases.[7,15,16] Nonetheless, because of its complexity and high costs, it becomes unfeasible as a screening tool in most clinical contexts. In this work, PMS2_vaR was developed, a pipeline designed to address this clinical need in the mutational analysis of the *PMS2* gene. This open-source code uses data already available as the output of a standard panel testing analysis and requires the installation of a few commonly used bioinformatic tools, making it easy to implement in diagnostic pipelines. The pipeline identifies candidate *PMS2* variants and classifies them according to the variant classification list provided. Only samples carrying putative (likely) pathogenic *PMS2* variants are recommended for subsequent LR-PCR analysis.

Our results demonstrated substantial clinical improvements, significantly increasing sensitivity for variant identification from 0.853 to 0.956 in the validation cohort while preserving specificity. Notably, all pathogenic variants were identified. PSVs were regarded as potential gene variants only if their VAF was over 60%, reducing the number of samples requiring confirmation by LR-PCR or cDNA analysis. As an illustration, in the HC cohort, consisting of 5619 samples, the pipeline only recommended (LR-)PCR analysis in 16 cases (0.28%), a number that can be handled in a routine clinical setting. The implementation of PMS2_vaR significantly reduced the number of required PCR analyses, highlighting its efficiency within the diagnostic workflow. By selectively targeting candidate variants identified by PMS2_vaR, this study was able to streamline the analysis process, minimizing unnecessary PCR reactions and conserving valuable resources. Moreover, this also accelerates the diagnostic process, ultimately reducing the turnaround times of the reports.

The selection of invariant positions to filter candidate gene variants in exon 11 was based on the analysis of 707 patient samples.[11] In rare cases, this method may lead to erroneous variant assignments owing to gene conversion–related sequence exchange. Therefore, *PMS2* gene variants potentially might be lost when following the PMS2_vaR algorithm. To reduce this possibility, *PMS2* (likely) pathogenic variants filtered out after the E11 approach should be confirmed by LR-PCR if tumor molecular characteristics are indicative of a *PMS2* alteration. A similar strategy is recommended for the recurrent pseudogenic c.2186_2187del and c.2243_2246del variants in exon 13.

The integration of the PMS2_vaR pipeline into the daily diagnostics routine may produce a notable clinical impact by improving the identification of CMMRD and LS. Because of the complexity of clinically diagnosing CMMRD,[4,5] an accurate and prompt diagnosis is essential for genetic counseling, surveillance recommendations, as well as therapeutic decisions.[3] In contrast, the identification of germline *PMS2* monoallelic variant carriers is more controversial because of its relatively lower penetrance compared with the other MMR genes,[2,26,27] although significant phenotypic variability has been observed among monoallelic carriers, even between individuals from the same family.[7,26,27] The use of effective screening tools for accurate *PMS2* variant detection will help in refining the LS phenotype associated with germline alterations in this gene.

MMR genes, including *PMS2,* are considered clinically actionable because pathogenic variant identification has high benefits for the patient and family in clinical practice.[28] The American College of Medical Genetics proposed reporting secondary incidental findings in these genes in clinical exome and genome sequencing analyses.[29] For panel testing under HC suspicion, analysis of the *MLH1*, *MSH2*, *MSH6*, *BRCA1*, and *BRCA2* genes has been recommended as opportunistic testing in adult cancer patients, regardless of the main clinical phenotype.[23] Expanding this framework to encompass *PMS2* requires the availability of optimized pipelines such as PMS2_vaR.

Nine of the 10 *PMS2* pathogenic variant carriers identified in the LS suspicion cohorts harbored tumors displaying isolated PMS2 or MSH6/PMS2 loss patterns, highly indicative of PMS2 deficiency, as the main driver of carcinogenesis. In contrast, in the HC cohort, four of the five *PMS2* carriers identified presented tumors showing conserved MMR expression. Although an immunohistochemistry test can yield false-negative results, especially

for missense MMR variants,[30] this also could agree with recent findings describing that some individuals carrying *PMS2* pathogenic variants may develop MMR-proficient tumors.[31] Nevertheless, the prevalence of *PMS2* pathogenic variants was enriched in the HC cohort (0.285%) compared with the estimated prevalence in the general population (0.140%).[1]

As a limitation, the PMS2_vaR pipeline has not been optimized to detect copy number variants in the *PMS2* gene. However, the assessment of copy number variant detection tools tailored for panel data using the modified BAM files obtained by PMS2_vaR represents a promising strategy for the future. Of note, one of the major strengths of the PMS2_vaR tool is that it can be adapted for *PMS2* variant calling in the analysis of NGS panels, exomes, and genomes. Moreover, there is potential for extension to other genes in the same situation through necessary code adjustments (eg, the *PRSS1* gene in the context of HC gene panels).

Conclusions

We have developed a pipeline to improve the accuracy of *PMS2* genetic testing by using standard NGS diagnostic workflows. The results show that its use reduces the number of samples that need to undergo LR-PCR and clearly improves the identification of *PMS2* variant carriers.

Acknowledgments

We thank the patients who participated in this study and the members of the Genetic Counseling Units and the Genetic Diagnostics Laboratory of the Hereditary Cancer Program of the Catalan Institute of Oncology.

Disclosure Statement

None declared.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2024.05.005*.

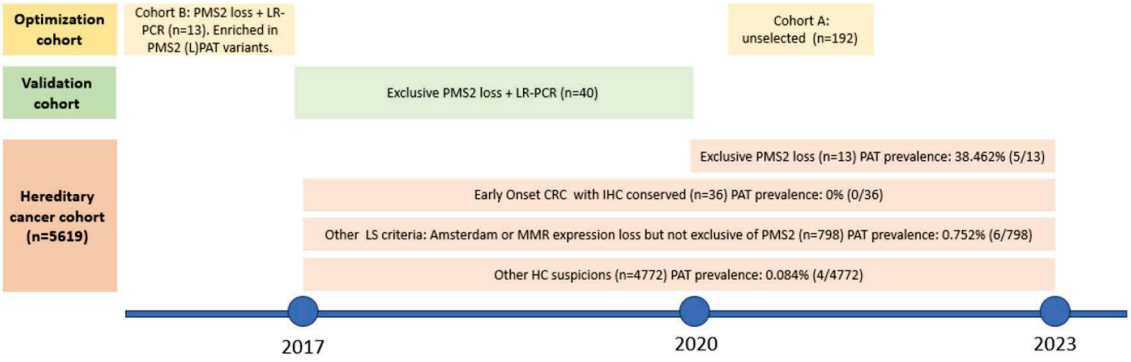
References

1. Win AK, Jenkins MA, Dowty JG, Antoniou AC, Lee A, Giles GG, Buchanan DD, Clendenning M, Rosty C, Ahnen DJ, Thibodeau SN, Casey G, Gallinger S, Le Marchand L, Haile RW, Potter JD, Zheng Y, Lindor NM, Newcomb PA, Hopper JL, MacInnis RJ: Prevalence and penetrance of major genes and polygenes for colorectal cancer. Cancer Epidemiol Biomarkers Prev 2017, 26:404
2. Dominguez-Valentin M, Sampson JR, Seppälä TT, ten Broeke SW, Plazzer JP, Nakken S, et al: Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. Genet Med 2020, 22:15–25
3. Tabori U, Hansford JR, Achatz MI, Kratz CP, Plon SE, Frebourg T, Brugieres L: Clinical management and tumor surveillance recommendations of inherited mismatch repair deficiency in childhood. Clin Cancer Res 2017, 23:e32–e37
4. Bakry D, Aronson M, Durno C, Rimawi H, Farah R, Alharbi QK, Alharbi M, Shamvil A, Ben-Shachar S, Mistry M, Constantini S, Dvir R, Qaddoumi I, Gallinger S, Lerner-Ellis J, Pollett A, Stephens D, Kelies S, Chao E, Malkin D, Bouffet E, Hawkins C, Tabori U: Genetic and clinical determinants of constitutional mismatch repair deficiency syndrome: report from the constitutional mismatch repair deficiency consortium. Eur J Cancer 2014, 50:987–996
5. Wimmer K, Kratz CP, Vasen HFA, Caron O, Ruiz-Ponte C, Slavc I, Burkhardt B, Brugieres L: Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium "Care for CMMRD" (C4CMMRD) on behalf of the EU-Consortium Care for CMMRD (C4CMMRD). J Med Genet 2014, 51:283–293
6. Andini KD, Nielsen M, Suerink M, Helderman NC, Koornstra JJ, Ahadova A, Kloor M, Mourits MJE, Kok K, Sijmons RH, Bajwa–ten Broeke SW: PMS2-associated Lynch syndrome: past, present and future. Front Oncol 2023, 13:547
7. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, Dowty JG, White DM, Hopper JL, Baglietto L, Jenkins MA, de la Chapelle A: The clinical phenotype of Lynch syndrome due to germline PMS2 mutations. Gastroenterology 2008, 135:419
8. Lincoln SE, Hambuch T, Zook JM, Bristow SL, Hatchell K, Truty R, Kennemer M, Shirts BH, Fellowes A, Chowdhury S, Klee EW, Mahamdallie S, Cleveland MH, Vallone PM, Ding Y, Seal S, DeSilva W, Tomson FL, Huang C, Garlick RK, Rahman N, Salit M, Kingsmore SF, Ferber MJ, Aradhya S, Nussbaum RL: One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genet Med 2021, 239:1673–1680
9. Nicolaides NC, Carter KC, Shell BK, Papadopoulos N, Vogelstein B, Kinzler KW: Genomic organization of the human PMS2 gene family. Genomics 1995, 30:195–206
10. Nakagawa H, Lockman JC, Frankel WL, Hampel H, Steenblock K, Burgart LJ, Thibodeau SN, De La Chapelle A: Mismatch repair gene PMS2: disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation. Cancer Res 2004, 64: 4721–4727
11. Gould GM, Grauman PV, Theilmann MR, Spurka L, Wang IE, Melroy LM, Chin RG, Hite DH, Chu CS, Maguire JR, Hogan GJ, Muzzey D: Detecting clinically actionable variants in the 3′ exons of PMS2 via a reflex workflow based on equivalent hybrid capture of the gene and its pseudogene. BMC Med Genet 2018, 19:176
12. Jansen AML, Tops CMJ, Ruano D, van Eijk R, Wijnen JT, ten Broeke S, Nielsen M, Hes FJ, van Wezel T, Morreau H: The complexity of screening PMS2 in DNA isolated from formalin-fixed paraffin-embedded material. Eur J Hum Genet 2020, 28:333
13. Hayward BE, De Vos M, Valleley EMA, Charlton RS, Taylor GR, Sheridan E, Bonthron DT: Extensive gene conversion at the PMS2 DNA mismatch repair locus. Hum Mutat 2007, 28:424–430
14. Ganster C, Wernstedt A, Kehrer-Sawatzki H, Messiaen L, Schmidt K, Rahner N, Heinimann K, Fonatsch C, Zschocke J, Wimmer K: Functional PMS2 hybrid alleles containing a pseudogene-specific missense variant trace back to a single ancient intrachromosomal recombination event. Hum Mutat 2010, 31:552–560
15. Clendenning M, Hampel H, LaJeunesse J, Lindblom A, Lockman J, Nilbert M, Senter L, Sotamaa K, De La Chapelle A: Long-range PCR facilitates the identification of PMS2-specific mutations. Hum Mutat 2006, 27:490–495
16. Vaughn CP, Robles J, Swensen JJ, Miller CE, Lyon E, Mao R, Bayrak-Toydemir P, Samowitz WS: Clinical analysis of PMS2:

mutation detection and avoidance of pseudogenes. Hum Mutat 2010, 31:588−593
17. van der Klift HM, Mensenkamp AR, Drost M, Bik EC, Vos YJ, Gille HJJP, Redeker BEJW, Tiersma Y, Zonneveld JBM, García EG, Letteboer TGW, Olderode-Berends MJW, van Hest LP, van Os TA, Verhoef S, Wagner A, van Asperen CJ, ten Broeke SW, Hes FJ, de Wind N, Nielsen M, Devilee P, Ligtenberg MJL, Wijnen JT, Tops CMJ: Comprehensive mutation analysis of PMS2 in a large cohort of probands suspected of Lynch syndrome or constitutional mismatch repair deficiency syndrome. Hum Mutat 2016, 37:1162−1179
18. Etzler J, Peyrl A, Zatkova A, Schildhaus HU, Ficek A, Merkelbach-Bruse S, Kratz CP, Attarbaschi A, Hainfellner JA, Yao S, Messiaen L, Slave I, Wimmer K: RNA-based mutation analysis identifies an unusual MSH6 splicing defect and circumvents PMS2 pseudogene interference. Hum Mutat 2008, 29:299−305
19. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, et al: Benchmarking challenging small variants with linked and long reads. Cell Genom 2022, 2:100128
20. Schwenk V, Leal Silva RM, Scharf F, Knaust K, Wendlandt M, Häusser T, Pickl JMA, Steinke-Lange V, Laner A, Morak M, Holinski-Feder E, Wolf DA: Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. J Med Genet 2023, 60:747−759
21. Castellanos E, Gel B, Rosas I, Tornero E, Santín S, Pluvinet R, Velasco J, Sumoy L, Del Valle J, Perucho M, Blanco I, Navarro M, Brunet J, Pineda M, Feliubadaló L, Capellá G, Lázaro C, Serra E: A comprehensive custom panel design for routine hereditary cancer testing: preserving control, improving diagnostics and revealing a complex variation landscape. Sci Rep 2017, 71:1−12
22. Moreno-Cabrera JM, del Valle J, Feliubadaló L, Pineda M, González S, Campos O, Cuesta R, Brunet J, Serra E, Capellà G, Gel B, Lázaro C: Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer. J Med Genet 2020, 59:75−78
23. Feliubadaló L, López-Fernández A, Pineda M, Díez O, del Valle J, Gutiérrez-Enríquez S, Teulé A, González S, Stjepanovic N, Salinas M, Capellá G, Brunet J, Lázaro C, Balmaña J; Catalan Hereditary Cancer Group: Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. Int J Cancer 2019, 145:2682−2691
24. Munté E, Feliubadaló L, Pineda M, Tornero E, Gonzalez M, Moreno-Cabrera JM, Roca C, Bales Rubio J, Arnaldo L, Capellá G, Mosquera JL, Lázaro C: vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines. Bioinformatics 2023, 39:btad128
25. Pan S, Brown A, Leclair B, Elias M, Chen D, Kidd J, Bowles K, Coffee B, Roa B, Mancini-Dinardo D: Common variants in PMS2CL that can present in PMS2 as pathogenic variants with extremely low frequencies. Houston TX, American Socierty of Human Genetics, 2019, [abstract 2424], October 15-19, 2019
26. Ten Broeke SW, Klift HMV, Tops CMJ, Aretz S, Bernstein I, Buchanan DD, et al: Cancer risks for PMS2-associated Lynch syndrome. J Clin Oncol 2018, 36:2961
27. Ten Broeke SW, Brohet RM, Tops CM, Van Der Klift HM, Velthuizen ME, Bernstein I, Munar GC, Garcia EG, Hoogerbrugge N, Letteboer TGW, Menko FH, Lindblom A, Mensenkamp AR, Moller P, Van Os TA, Rahner N, Redeker BJW, Sijmons RH, Spruijt L, Suerink M, Vos YJ, Wagner A, Hes FJ, Vasen HF, Nielsen M, Wijnen JT: Lynch syndrome caused by germline PMS2 mutations: delineating the cancer risk. J Clin Oncol 2015, 33:319−325
28. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Sue Richards C, Vlangos CN, Watson M, Martin CL, Miller DT; on behalf of the ACMG Secondary Findings Maintenance Working Group: Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 2017, 19:249−255
29. Miller DT, Lee K, Abul-Husn NS, Amendola LM, Brothers K, Chung WK, Gollob MH, Gordon AS, Harrison SM, Hershberger RE, Klein TE, Richards CS, Stewart DR, Martin CL: ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med 2023, 25:100866
30. Hechtman JF, Rana S, Middha S, Stadler ZK, Latham A, Benayed R, Soslow R, Ladanyi M, Yaeger R, Zehir A, Shia J: Retained mismatch repair protein expression occurs in approximately 6% of microsatellite instability-high cancers and is associated with missense mutations in mismatch repair genes. Mod Pathol 2020, 33:871−879
31. Ranganathan M, Sacca RE, Trottier M, Maio A, Kemel Y, Salo-Mullen E, Catchings A, Kane S, Wang C, Ravichandran V, Ptashkin R, Mehta N, Garcia-Aguilar J, Weiser MR, Donoghue MTA, Berger MF, Mandelker D, Walsh MF, Carlo M, Liu YL, Cercek A, Yaeger R, Saltz L, Segal NH, Mendelsohn RB, Markowitz AJ, Offit K, Shia J, Stadler ZK, Latham A: Prevalence and clinical implications of mismatch repair-proficient colorectal cancer in patients with Lynch syndrome. JCO Precis Oncol 2023, 7:e2200675

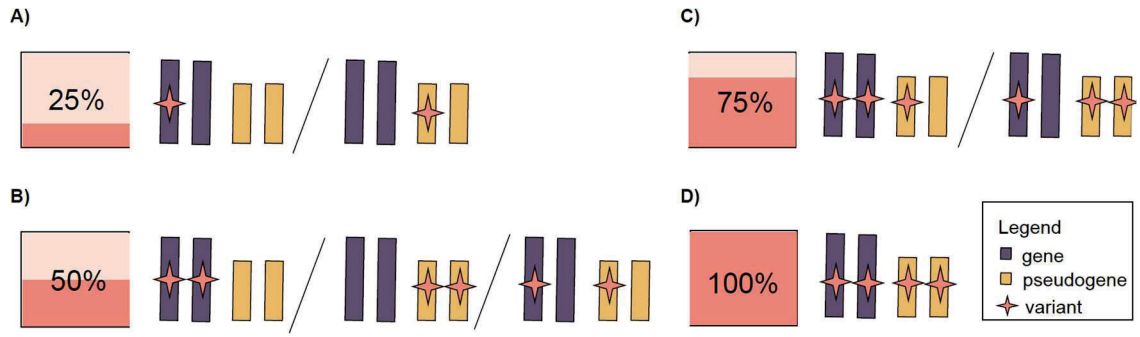
Supplemental Figure S1

Patient cohorts used for the analysis, stratified by sample acquisition year and subgroups. The prevalence of identified *PMS2* variants is indicated for each subgroup (number of carriers/number of samples analyzed). CRC, colorectal cancer; HC, hereditary cancer; IHC, immunohistochemistry; (L)PAT, (likely) pathogenic; LR-PCR, long-range PCR; LS, Lynch syndrome; MMR, mismatch repair.



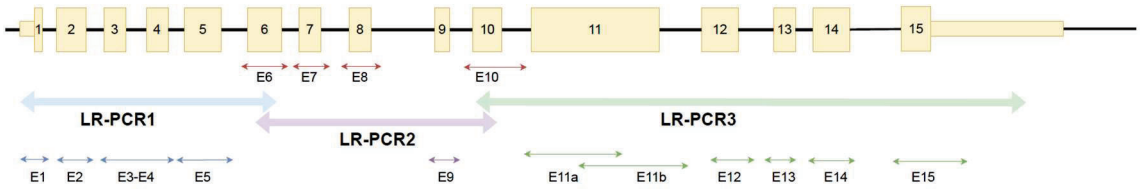
Supplemental Figure S2

Schematic representation of theoretical variant allele frequencies (VAFs) in a tetra-allelic situation, caused by forcing PMS2CL reads to map to *PMS2* reference (exon 9 and exons 11 to 15). A: When the VAF is approximately 25%, the variant can be in either one allele of the gene or the pseudogene. **B:** When the VAF is approximately 50%, the variant can be present in both gene alleles, in both pseudogene alleles, or in one of each. **C:** When the VAF is approximately 75%, three alleles are involved, two in the gene and one in the pseudogene or *vice versa*. **D:** When VAF is approximately 100%, the variant is present in both alleles of both the gene and the pseudogene.



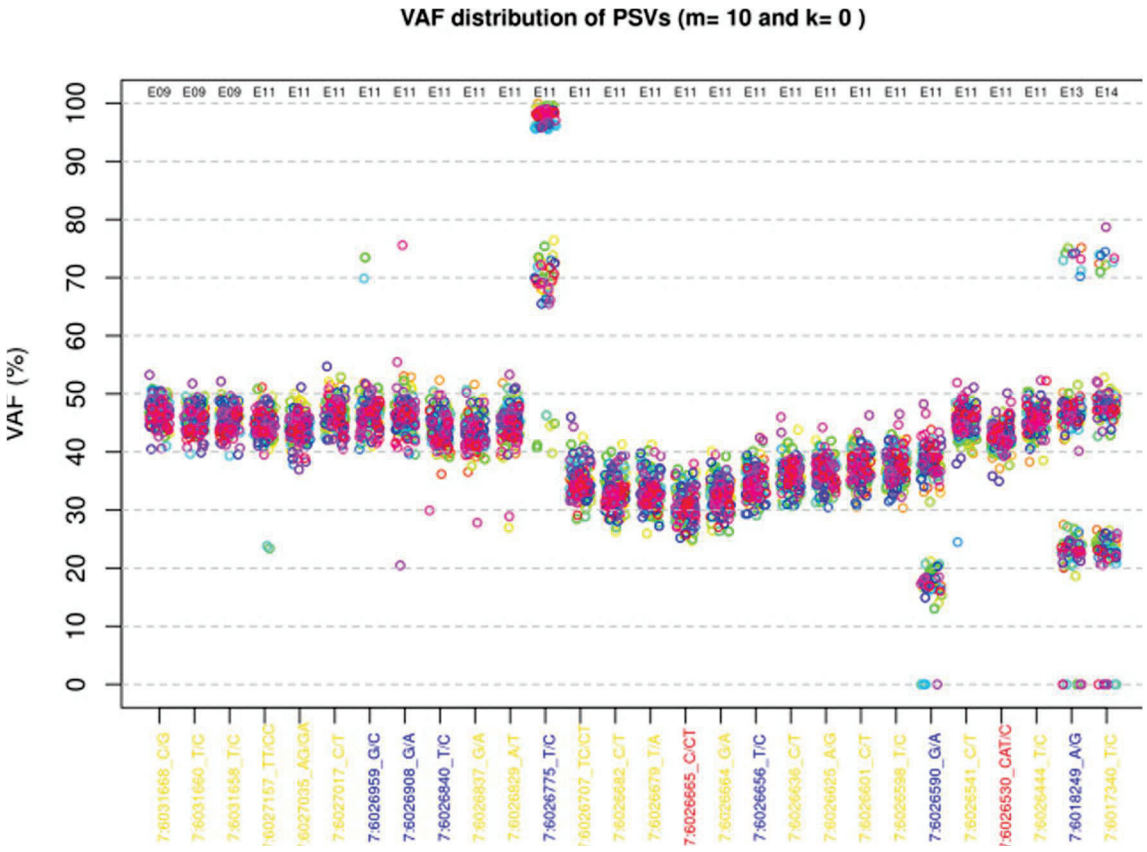
Supplemental Figure S3

Schematic representation detailing the annealing positions of the primer's sequences for *PMS2* amplification. E, exon; LR-PCR, long-range PCR.



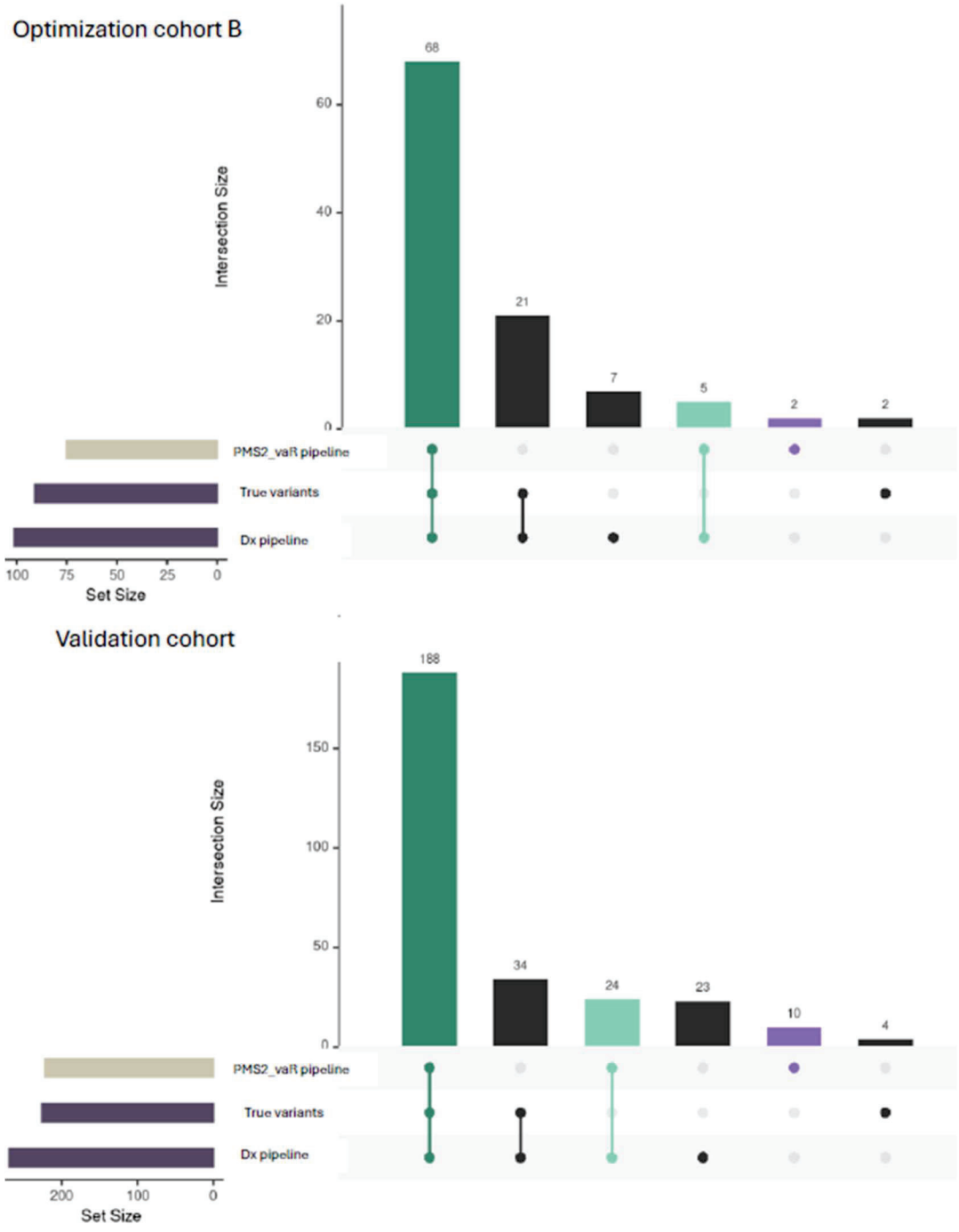
Supplemental Figure S4

Variant allele frequency (VAF) distribution of paralogous sequence variants (PSVs) in optimization cohort A (*n* = 192). Pathogenic or likely pathogenic variants are shown in red, variants of unknown significance are shown in yellow, and likely benign and benign variants are shown in blue. E, exon.



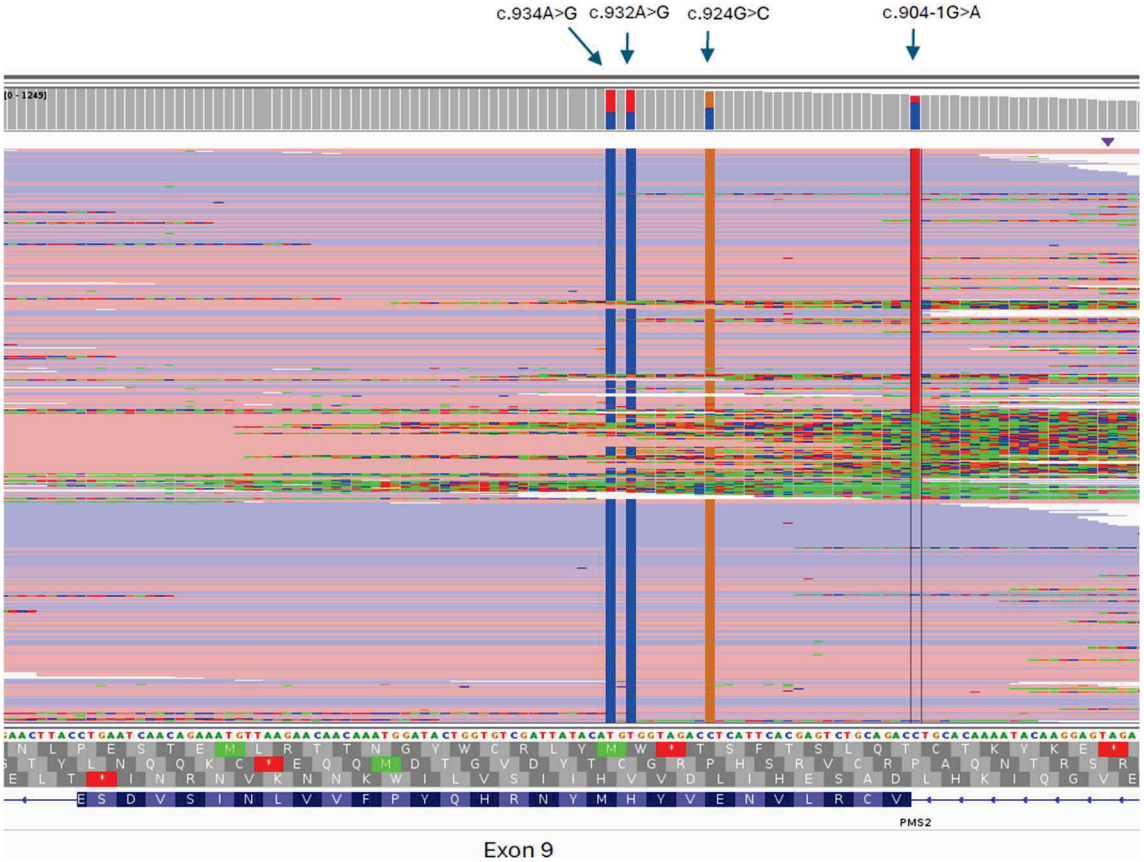
Supplemental Figure S5

Upset plot comparing the performance of the diagnostic and PMS2_vaR pipelines against the true variants detected by long-range PCR. True variants detected by both pipelines are shown in dark green. True variants detected only by the PMS2_vaR pipeline are shown in light green. True variants not detected by either pipeline are shown in purple. False-positive variants detected by either of the two pipelines are shown in black. There were no true variants detected only by the diagnostic pipeline. Dx, diagnostic.



Supplemental Figure S6

Visualization of reads mapped to exon 9 and part of intron 8 in patient 7 using the Integrative Genomics Viewer. The variant c.904-1G>A is observed in *cis* with paralogous sequence variants, specifically c.924G>C, c.932A>G, and c.934A>G, indicating a pseudogene origin.



Article publicat 2

Detection of germline CNVs from gene panel data: benchmarking the state of the art

Elisabet Munté¥, Carla Roca¥, Jesús del Valle, Lídia Feliubadaló, Marta Pineda, Bernat Gel, Elisabeth Castellanos, Bárbara Rivera, David Cordero, Víctor Moreno, Conxi Lázaro* i José Marcos Moreno-Cabrera*

BIB. 2025 Jan;26(1)

DOI: 10.1093/bib/bbae645

OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae645

https://doi.org/10.1093/bib/bbae645

Problem Solving Protocol

Detection of germline CNVs from gene panel data: benchmarking the state of the art

Elisabet Munté[1,2,‡], Carla Roca[1,2,‡], Jesús Del Valle[1,3], Lidia Feliubadaló[1,3], Marta Pineda[1,3], Bernat Gel[4], Elisabeth Castellanos[5,6], Barbara Rivera[1,7,8], David Cordero[9,10,11], Víctor Moreno[11,12,13,14], Conxi Lázaro[1,3,*], José Marcos Moreno-Cabrera[1,9,12,*]

[1]Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL-ONCOBELL, Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Spain
[2]Doctoral Programme in Biomedicine, University of Barcelona (UB), Casanova 143, 08036 Barcelona, Spain
[3]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Instituto de Salud Carlos III, Monforte de Lemos 5, 28029 Madrid, Spain
[4]Hereditary Cancer Group, Germans Trias i Pujol Research Institute (IGTP), Can Ruti Campus, Camí de les Escoles s/n, 08916 Badalona, Barcelona, Spain
[5]Clinical Genomics Research Group, Germans Trias i Pujol Research Institute (IGTP), Can Ruti Campus, Camí de les Escoles s/n, Badalona, Barcelona, Spain
[6]Genetics Department, Germans Trias i Pujol University Hospital (HUGTiP), Can Ruti Campus, Carretera de Canyet s/n, 08916 Badalona, Barcelona, Spain
[7]Lady Davis Institute and Segal Cancer Centre, Jewish General Hospital, 3755 Chemin de la Côte-Sainte-Catherine, Montreal, QC, Canada
[8]Gerald Bronfman Department of Oncology, McGill University, 5100 de Maisonneuve Blvd. West, Suite 720 Montreal, QC, Canada
[9]Unit of Bioinformatics for Precision Oncology (UBOP), Catalan Institute of Oncology, Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[10]Preclinical and Experimental Research in Thoracic Tumors (PReTT), ONCOBELL Program, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[11]Consorcio de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Monforte de Lemos 5, 28029 Madrid, Spain
[12]Oncology Data Analytics Program (ODAP), Catalan Institute of Oncology, Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[13]Colorectal Cancer Group, ONCOBELL Program, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[14]Department of Clinical Sciences, Faculty of Medicine and Health Sciences, Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona (UB), Freixa Llarga s/n, 08907 L'Hospitalet de Llobregat, Barcelona, Spain

*Corresponding authors. José Marcos Moreno-Cabrera, Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL-ONCOBELL, 08908 L'Hospitalet de Llobregat, Spain. E-mail: jmmoreno@iconcologia.net; Conxi Lázaro, Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL-ONCOBELL, 08908 L'Hospitalet de Llobregat, Spain. E-mail: clazaro@iconcologia.net
‡Elisabet Munté and Carla Roca are both first authors (Joint authors), this should be properly shown.

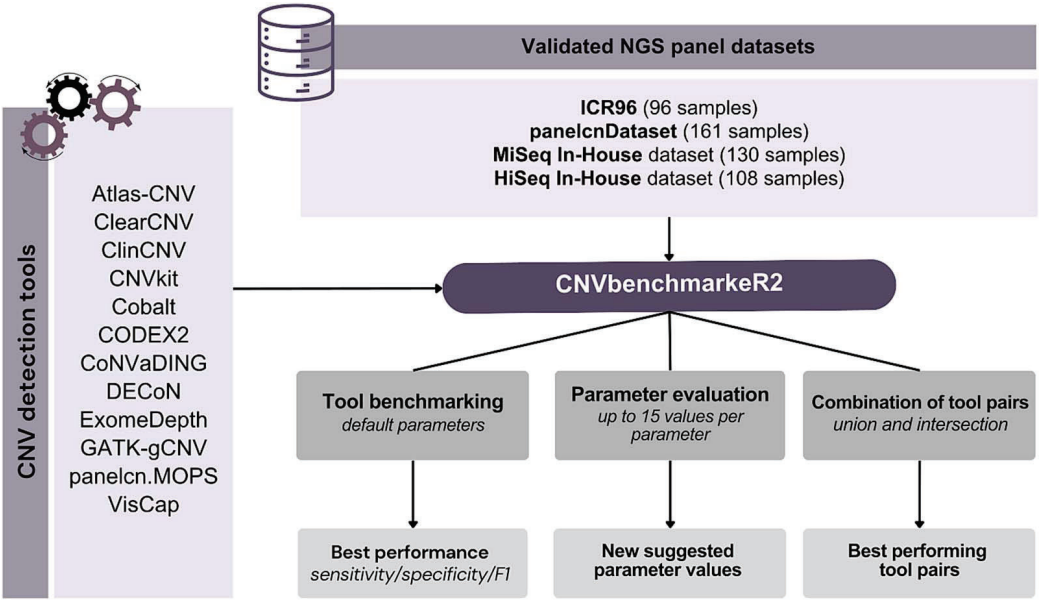
Abstract

Germline copy number variants (CNVs) play a significant role in hereditary diseases. However, the accurate detection of CNVs from targeted next-generation sequencing (NGS) gene panel data remains a challenging task. Several tools for calling CNVs within this context have been published to date, but the available benchmarks suffer from limitations, including testing on simulated data, testing on small datasets, and testing a small subset of published tools. In this work, we conducted a comprehensive benchmarking of 12 tools (Atlas-CNV, ClearCNV, ClinCNV, CNVkit, Cobalt, CODEX2, CoNVaDING, DECoN, ExomeDepth, GATK-gCNV, panelcn.MOPS, VisCap) on four validated gene panel datasets using their default parameters. We also assessed the impact of modifying 107 tool parameters and identified 13 parameter values that we suggest using to improve the tool F1 score. A total of 66 tool pair combinations were also evaluated to produce better meta-callers. Furthermore, we developed CNVbenchmarker2, a framework to help users perform their own evaluations. Our results indicated that in terms of F1 score, ClinCNV and GATK-gCNV were the best CNV callers. Regarding sensitivity, GATK-gCNV also exhibited particularly high performance. The results presented here provide an evaluation of the current state of the art in germline CNV detection from gene panel data and can be used as a reference resource when using any of the tools.

Received: June 11, 2024. **Revised:** November 4, 2024. **Accepted:** November 28, 2024
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

Graphical Abstract



Keywords: CNVs; benchmarking; gene panels; germline

Introduction

Copy number variants (CNVs) are structural genomic alterations that involve an abnormal number of copies of a DNA segment, resulting in both deletions and duplications. They are a type of structural variation caused by genomic rearrangements, varying in size from 50 bp to several megabases [1, 2]. CNVs are a major source of genomic variation in humans [3]. They affect various biological processes, including evolution, adaptation, and the development and predisposition to diseases such as autism, obesity, and cancer [4–6]. Once a CNV is identified, its clinical significance can be determined, and medical management and prevention measures can be implemented. Their detection is therefore crucial in clinical diagnostics [7].

Several methods to detect CNVs have been developed in recent decades. These methods include Polymerase chain reaction-based methods such as multiplex ligation–dependent probe amplification (MLPA), array-based technologies like microarray-based comparative genomic hybridization (aCGH) or SNP microarrays, massive parallel sequencing, fluorescence *in situ* hybridization (FISH), or Southern blotting. From the above list, MLPA, aCGH, and SNP microarrays are frequently used in diagnostic routines, being MLPA the most common approach for testing one or a few genes [8, 9]. However, these methods are still expensive, time-consuming, and have gene-specific limitations. For example, MLPA relies on single-gene approaches, while aCGH's sensitivity is restricted to sequences within the array's design assembly [2].

The arrival of next-generation sequencing (NGS) has transformed genetic testing by allowing millions of fragments to be sequenced simultaneously [10]. Diagnostic laboratories are using NGS methods to identify multiple types of variation, including CNVs. In diagnostic settings, where laboratories handle a large number of samples, targeted gene panels have emerged as a common and cost-effective approach.

Many bioinformatic tools have been published to identify germline CNVs from NGS data. While most of these tools are reliable for detecting large CNVs, they often struggle to detect small CNVs, especially those spanning a single exon. Furthermore, most tools are not optimized for calling CNVs from targeted gene panel data, as they were originally developed for use with whole-genome or whole-exome data. Beyond addressing these challenges, tools must demonstrate high sensitivity and specificity in diagnostic settings [9]. Therefore, it is crucial to accurately measure tool performance on gene panel data.

Previous studies have evaluated the performance of germline CNV callers on gene panel data. However, these studies suffer from some limitations. Most of them were performed by the tool authors, covered a small subset of currently available tools, and were evaluated on a single dataset [11–15]. To our knowledge, three benchmarks have been published to date by authors who did not evaluate their own tool [16–18]. However, two of them have similar limitations: Roca *et al*. evaluated mainly on simulated data with only a small number of validated CNVs, and Lepkes *et al*. benchmarked four tools on a single dataset where MLPA tests were performed only for CNV calling confirmation. The limitations were partially addressed in our previous work, which evaluated five tools on four real datasets with MLPA results available prior to tool execution [17]. However, our previous work only covered a subset of the tools published until 2018. Moreover, several new tools have been published since 2018, so our previous benchmark provides an incomplete assessment of the current state of the art. Here, we aim to provide a wider, comprehensive, and up-to-date evaluation of germline CNV detection tools on gene panel data by benchmarking 12 tools on four real and publicly available datasets, evaluating the impact of modifying 107 tool parameters and combining tool pairs.

Material and methods

Datasets

We defined the criteria for including datasets in the benchmark. These requirements comprised being obtained from gene panel sequencing, having MLPA results before *in silico* calling, including germline single-exon CNVs, and being publicly available. To the best of our knowledge, only four datasets met the requirements

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

as of April 2024, namely, the ICR96 exon CNV validation series (96 samples) [19], a subset of the data used in the panelcn.MOPS publication referred to as panelcnDataset (161 samples) [13], and two in-house datasets (130 and 108 samples sequenced in Illumina MiSeq and HiSeq platforms, respectively) [17]. Table 1 provides further details of the datasets used in this work. All datasets were obtained from hybridization-based capture panels designed for hereditary cancer diagnostics: the TruSight Cancer Panel (Illumina, San Diego, CA) and the ICO-IMPPC Hereditary Cancer Panel (I2HCP) [20]. The bed file defining the regions of interest (ROIs) for the ICR96 and panelcnDataset datasets can be found in Supplementary Table 1, whereas the one for the in-house datasets is in Supplementary Table 2. Datasets contain single and multi-exon CNVs detected in diagnostic routine through MLPA testing. Negative MLPA results, indicating unaffected genes, are also available. MLPA results for each dataset can be found in Supplementary Table 3.

Sample alignment was performed using Burrows-Wheeler Aligner (BWA) mem v0.7.17 to the GRCh37 human genome assembly [21, 22] We then used SAMtools v1.16.1 [21] to sort and index Binary Alignment Map (BAM) files and Picard v2.27.4 to include read group information. No further processing or filtering was applied to the BAM files.

Copy number variant detection tools

The selection of detection tools was based on multiple criteria. Specifically, they must be publicly available, capable of calling germline CNVs at the exon level, designed to work with gene panel data, and not purposely built only for amplicon-based sequencing data. Following the completion of the literature review in December 2023, 12 tools were selected according to these criteria (Table 2): clearCNV v0.306 [23], GATK-gCNV v4.5.0 [24], Atlas-CNV v.0 [15], Cobalt v0.8.0 [25], ClinCNV v1.18.3 [26], CNVkit v0.9.10 [27], VisCap v0.8 [28], DeCoN v2.0.1 [29], panelcn.MOPS v1.20.0 [13], ExomeDepth v1.1.16 [30], CoNVaDING v1.2.1 [11], and CODEX2 v1.3.0 [31]. The latter five were evaluated in our previous work, but we included them here to facilitate tool comparison and to evaluate the most updated versions. Nine germline CNV detection tools were considered for inclusion in this work but were later discarded for multiple reasons: SeqCNV, CNVPanelizer, CNV-Z, ifCNV, SavvyCNV, CCR-CNV, Hadoop-CNV-RF, PattRec, and the pipeline used by Singh *et al*. [9] Supplementary Table 4 provides a list of discarded callers and the reason for their exclusion.

Benchmark evaluation metrics

The performance of each tool was evaluated at two levels: per ROI and per gene. Detailed definitions of both levels can be found in Supplementary File 1.

For each tool, across all dataset and evaluation levels, a range of performance metrics were computed, including sensitivity, specificity, positive predictive value, negative predictive value, false negative rate, false positive rate, F1 score, accuracy, Matthews correlation coefficient, and Cohen's kappa coefficient (Supplementary File 1). We also measured tool run times on the ICR96 dataset using a workstation with 24 GB random access memory (RAM) and 1 central processing unit (CPU) per job.

Benchmark execution

We implemented CNVbenchmarkeR2, an R framework that enables the automatic and flexible benchmarking of CNV callers. Code and documentation are available at https://github.com/jpuntomarcos/CNVbenchmarkeR2, so other users can benefit

Table 1. Datasets used in the benchmark. List of datasets used in the benchmark including the number of samples, validated CNVs (single and multi-exon, deletions, and duplications), sequencing platform employed, availability of the datasets, and additional relevant details. Abbreviations: CNVs, copy number variants; ROI, region of interest; EGA, European Genome-phenome Archive

Dataset	Samples	Single-exon CNVs	Multi-exon CNVs	Deletion CNVs	Duplication CNVs	Validated genes with CNV	Validated ROIS with CNVs	Sequencing	Availability	Additional information
ICR96	96	25	43	51	17	68	296	TruSight Cancer Panel v2 (100 genes), HiSeq, 2 × 101 bp reads	EGA dataset ID: EGAD00001003335	Samples obtained from one run.
panelcnDataset	161	13	28	36	5	41	321	TruSight Cancer Panel (94 genes), MiSeq, 2 × 151 bp reads	EGA dataset ID: EGAS00001002481	The EGA dataset contains 170 samples, but 9 were excluded for this work (see Supplementary File 1)
In-house MiSeq	130	19	45	56	8	64	394	I2HCP Panel v2.0–v2.2 (122–135 genes), MiSeq, 2 × 300 bp reads	EGA dataset ID: EGAS00001004316	Samples obtained from 48 runs. Three samples had a mosaic CNV.
In-house HiSeq	108	18	40	49	9	58	525	I2HCP panel v2.0–v2.2 (122–135 genes), HiSeq, 2 × 251 bp reads	EGA dataset ID: EGAS00001004316	Samples obtained from 5 runs. Two samples had a mosaic CNV.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

Table 2. Benchmarked tools. List of benchmarked tools including version used in the benchmark, programming language, URL for accessing the tool's code, a brief summary of the CNV detection method, number of parameters examined in the parameter evaluation section, year of publication, number of citations, PubMed ID (PMID), and whether the tool was benchmarked in a previous work of Moreno-Cabrera *et al.* [17].

Tool	Version	Language	Availability	Methods	Number of evaluated parameters	Year (paper publication)	Citations[a]	PMID	Benchmarked in [14]
Atlas-CNV	0	R and Perl program	https://github.com/theodorc/Atlas-CNV	It normalizes individual read depth data to average read depth per target, converting it to reads per kilobase million (RPKM). It computes log2 scores for each sample/median ratio at every exon, assessing sample quality via SampleQC, checking StDev of log2 scores and analysis of variance (ANOVA) on mean RPKM coverage.	2	2019	14	30890783	No
ClearCNV	0.306	Python program	https://github.com/bihealth/clear-cnv	It utilizes match scores to group samples based on coverage patterns. It employs data normalization, scaled z-scores, and *r*-scores to identify copy number variations (CNVs) in both multi-exon and single-exon regions.	7	2022	1	35751599	No
ClinCNV	1.18.3	R, Java, Python program	https://github.com/imgag/ClinCNV	ClinCNV employs an algorithm that combines the strengths of circular binary segmentation and hidden Markov model–based techniques to perform multi-sample normalization and CNV calling.	2	2022[b]	6	–	No
CNVkit	0.9.10	Python program	https://github.com/etal/cnvkit	It uses targeted and the nonspecifically captured off-target reads to calculate log2 copy ratios across the genome.	18	2016	1212	27100738	No
Cobalt	0.8.0	Python program	https://github.com/ARUP-NGS/cobalt	It introduces two algorithmic adaptations to improve accuracy in a hidden Markov model. A method for computing target and copy number–specific emission distributions and they perform pointwise maximum *a posteriori* HMM decoding to improve sensitivity for small CNV.	8	2022	0	35854218	No
CODEX2	1.3.0	R package	https://github.com/yuchaojiang/CODEX2	Based on CODEX package, it models the GC content bias and normalizes the read depth data for CNV detection via a Poisson latent factor model.	8	2018	39	30477554	Yes (v.1.2.0)
CoNVaDING	1.2.1	Perl program	https://github.com/molgenis/CoNVaDING	Combination of ratio scores and *Z*-scores of the sample of interest compared to the selected normalized control samples.	7	2016	67	26864275	Yes (v.1.2.0)
DECoN	2.0.1	R program	https://github.com/RahmanTeam/DECoN	Modifies ExomeDepth package by altering the hidden Markov model probabilities to depend upon the distance between exons.	3	2016	59	28459104	Yes (v.1.0.1)
ExomeDepth	1.1.16	R package	https://github.com/vplagnol/ExomeDepth	Beta-binomial model with GC correction and hidden Markov model to combine likelihood across exons.	3	2012	516	22942019	Yes (v.1.1.10)
GATK-gCNV	4.5.0	Java, Python, R program	https://github.com/broadinstitute/gatk	It calculates read counts over specified genomic regions per sample; it clusters technically similar samples using principal component analysis to reduce biases and enhance efficiency. After estimating chromosomal ploidy, it denoises read depth, infers CNVs via a unified model using the Viterbi algorithm	35	2023	0	37604963	No
pan-elcn.MOPS	1.20.0	R package	https://github.com/bioinf-jku/panelcn.mops	Adaptation of cn.MOPS package, which decomposes variations in coverage across samples into integer copy numbers and noise by means of its mixture components and Poisson distributions.	13	2017	53	28449315	Yes (v.1.0.0)
VisCap	0.8	R program	https://github.com/pughlab/VisCap	It determines the portion of sequence coverage allocated to genomic intervals and calculates log2 ratios compared to the median of reference samples with a matching test setup. CNV candidates are identified when log2 ratios surpass thresholds set by the user.	2	2016	49	26681316	No

[a]Citations obtained on Nov 2023. [b]Preprint.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

from it to benchmark the tools against their own datasets. We used CNVbenchmarkeR2 to run each tool on each dataset using the default parameters specified in the tool documentation.

The CNVbenchmarkeR2 code shows the steps performed to run each tool. In the case of GATK-gCNV, we followed the guide published by GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035531152--How-to-Call-rare-germline-copy-number-variants) to call rare germline variants, including the AnnotateIntervals and FilterIntervals steps, both recommended in the guide. However, since these steps are described as optional in the guide and users may ignore their impact on performance, we benchmarked two additional workflows to compare them with the final one. Thus, we benchmarked: (i) the complete workflow (GATK-gCNV) including both AnnotateIntervals and FilterIntervals steps, (ii) a workflow excluding AnnotateIntervals and FilterIntervals steps (GATK-gCNV_no_AI_FI), and (iii) a workflow excluding the AnnotateIntervals step (GATK-gCNV_no_AI), which is the default approach in the GATK germline cohort workflow description language pipeline.

Parameter evaluation

All tools evaluated in this benchmark have adjustable parameters. However, in most cases, neither the tool documentation nor any other source is clear about the impact on tool performance when these parameters are changed. To address this issue, we systematically evaluated tool parameters by testing them over a wide range of values on all datasets. For numerical parameters, we tested 15 parameter values, including the default one. For categorical parameters, we tested all available options in the tool. We computed the metrics described in the Benchmark Evaluation Metrics section for all executions.

For numerical parameters, we also obtained the optimal range as follows: (i) for each dataset, we identified the parameter value that maximized the F1 score at the ROI level; (ii) The optimal range is defined by the lowest and highest parameter values obtained from the previous step. We used the optimal range to identify parameters where the optimal range is completely below or above the default parameter value. For these parameters, we also determined the suggested parameter value to use, which we defined as the value of the optimal range that is closest to the default value.

Combination of tool pairs

We assessed the impact of tool pair unions and intersections on performance and ascertained whether any pair was capable of detecting all CNVs. All 66 possible tool pairs were evaluated by combining the results obtained separately when using the default parameters on the four datasets included in this work. Both per ROI and per gene metrics were generated. The R package GenomicRanges v1.48.0 was used to calculate the union and intersection of tool calls.

Results

Benchmark with default parameters

The tools were run on every dataset using default parameters. Evaluation metrics were then calculated at two levels: per ROI and per gene (see Material and Methods section for details).

Per ROI metrics allow us to assess tool performance at single-exon resolution (Fig. 1, Supplementary Table 5). Regarding the F1 score, a common measure of binary classifier accuracy, tool performance varied widely across datasets, ranging from 0.42 (CNVkit in ICR96) to 0.98 (GATK-gCNV in panelcnDataset). Interestingly, GATK-gCNV and ClinCNV were the only tools to consistently score in the top five for each dataset, with values between 0.78 and 0.98. On the other hand, all tools were highly specific, achieving values over 0.94 in all tool–dataset runs. In terms of sensitivity, we observed more variability, with tools ranging from 0.43 to 0.99. For both the ICR96 and panelcnDataset datasets, all tools except Atlas-CNV and VisCap achieved a sensitivity >0.90. CNVkit exceeded this threshold as well, but only for the ICR96 dataset. However, in the in-house datasets, only ClinCNV, CODEX2, GATK-gCNV, and CoNVaDING achieved sensitivity values >0.90. Supplementary Figure 1 shows per ROI results sorted by sensitivity to facilitate the analysis.

Figure 2 and Supplementary Table 5 show benchmark results at the gene level, which are particularly relevant in diagnostic settings. Regarding sensitivity, a metric commonly used in diagnostics to assess the classifier's ability to detect positives, some tools demonstrated high performance. In particular, GATK, CoNVaDING, DECoN, and CODEX2 obtained values >0.93 for each dataset. CoNVaDING showed very high performance in detecting positives: it missed only 3 out of 231 positives across all datasets. GATK-gCNV, DECoN, and CODEX2 missed more positives in total (7, 11, and 12, respectively) but generated fewer false positives (FPs; 27, 62, and 93, respectively) compared to CoNVaDING (150 FPs). On the other hand, ClearCNV and Cobalt were the callers that missed most of the positives: 65 and 61, respectively. Supplementary Figure 2 shows the per gene results sorted by sensitivity. In terms of F1 score, tools showed again large differences with values ranging from 0.26 to 0.98. GATK-gCNV exhibited the highest performance based on this metric: only DECoN surpassed it in the InHouse HiSeq dataset.

Supplementary File 1 shows tool run times obtained in a workstation with 24 GB RAM and 1 CPU per job. The benchmarked tools required a median of 53 minutes to perform CNV calling on the ICR96 dataset. ClinCNV and CODEX2 were the fastest tools, completing the task in 13 and 14 minutes, respectively. In contrast, Atlas-CNV, CNVkit and VisCap were the most time-consuming tools, requiring 147, 148 and 192 minutes, respectively.

GATK-gCNV workflows

The GATK-gCNV results presented in this work were obtained including the AnnotateIntervals and FilterIntervals steps. However, to better understand the effect of including these steps, two additional workflows were benchmarked (see Methods). Supplementary File 1 shows that certain metrics exhibited considerable variability across workflows. With regard to per ROI metrics, the workflows including the FilterIntervals step demonstrated higher sensitivity across all datasets in comparison to the GATK-gCNV_no_AI_FI workflow, with increases ranging from 0.01 to 0.09. The enhancement was even more pronounced in per gene metrics. In particular, the workflows including the FilterIntervals step demonstrated superior performance compared to GATK-gCNV_no_AI_FI, with gains ranging from 0.03 to 0.30. Furthermore, these two workflows exhibited notable F1 score improvement at the gene level, with values increasing between 0.01 and 0.22.

When comparing both workflows that include the FilterInterval step, GATK-gCNV and GATK-gCNV_no_AI, neither tool demonstrated a clear advantage in terms of sensitivity, specificity, or F1 score across all datasets.

Parameter evaluation

We systematically varied each tool parameter over a broad range of values to assess its impact on tool performance (see

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

Figure 1. Benchmark results at the ROI level. The tools were run using default parameters and are listed in descending order based on their F1 score in each dataset. (FN, false negative; FP, false positive; F1, F1 score).

Material and Methods). A total of 6110 executions were conducted to evaluate 107 tool parameters across all datasets. All results are presented in Supplementary Table 6. Additionally, 436 figures containing sensitivity, specificity, and F1 score at the ROI and gene level were also generated and are available at https://doi.org/10.6084/m9.figshare.25930960. Tool users can utilize these results as a guide to understand the expected effect when modifying each parameter.

Modifying each parameter had a different effect on tool performance. We identified four main patterns of performance change: (i) no discernible effect on the tool performance, resulting in flat curves (e.g. DECoN mincorr parameter); (ii) Increase in sensitivity and decrease in specificity or vice versa (e.g. CODEX2 cn_del_threshold parameter); (iii) sensitivity or specificity exhibiting a bell-shaped behavior (e.g. ClearCNV zscale parameter); (iv) performance changes without a distinct pattern, often showing successive increases and decreases in sensitivity (e.g. ClearCNV sample_score_factor parameter).

At the ROI level, we also determined the optimal range for each numerical parameter and identified 13 parameters where the optimal range was completely below or above the default parameter value (Table 3). In such cases, adjusting the default parameter value in one direction, increasing or decreasing it, results in a higher F1 score at the ROI level across all datasets. We therefore identified the suggested parameter value to use as the one within the optimal range closest to the default value.

Combination of tool pairs

We evaluated all 66 combinations of tool pairs using their parameters set to default (Supplementary Table 7). The union of calls from tool pairs resulted in better sensitivity albeit at the expense of a lower specificity. From an ROI-level perspective, no combination of tools achieved a sensitivity of 1. As per gene level results, five tool pairs achieved the maximum sensitivity across all datasets (Supplementary File 1): Atlas-CNV/CoNVaDING, CODEX2/CoNVaDING, CNVkit/CoNVaDING, panelcn.MOPS/CoNVaDING, and DECoN/CODEX2. Among these pairs, the union of CNVkit and CoNVaDING yielded the lowest specificity in most datasets, with values between 0.65 and 0.84. The other pairs did not show large differences between them and achieved values ranging from 0.78 to 0.97 across datasets.

In contrast, intersecting tool calls increased specificity at the expense of a lower sensitivity. No tool pair achieved perfect specificity across all datasets. However, the intersection of CODEX2/GATK-gCNV and CODEX2/DECoN identified all true CNVs in the panelcnDataset dataset at the gene level, without generating FPs.

Discussion

The published benchmarks of germline CNV callers for gene panel data suffer from certain limitations. These limitations include evaluating mainly simulated data, evaluating small datasets, or testing only a small subset of published tools [16–18]. Here, we conducted a comprehensive benchmark of 12 tools on four publicly available datasets, using tool default parameters, assessing the impact of changing tool parameter values, and evaluating the combination of tool pairs to produce better meta-callers.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

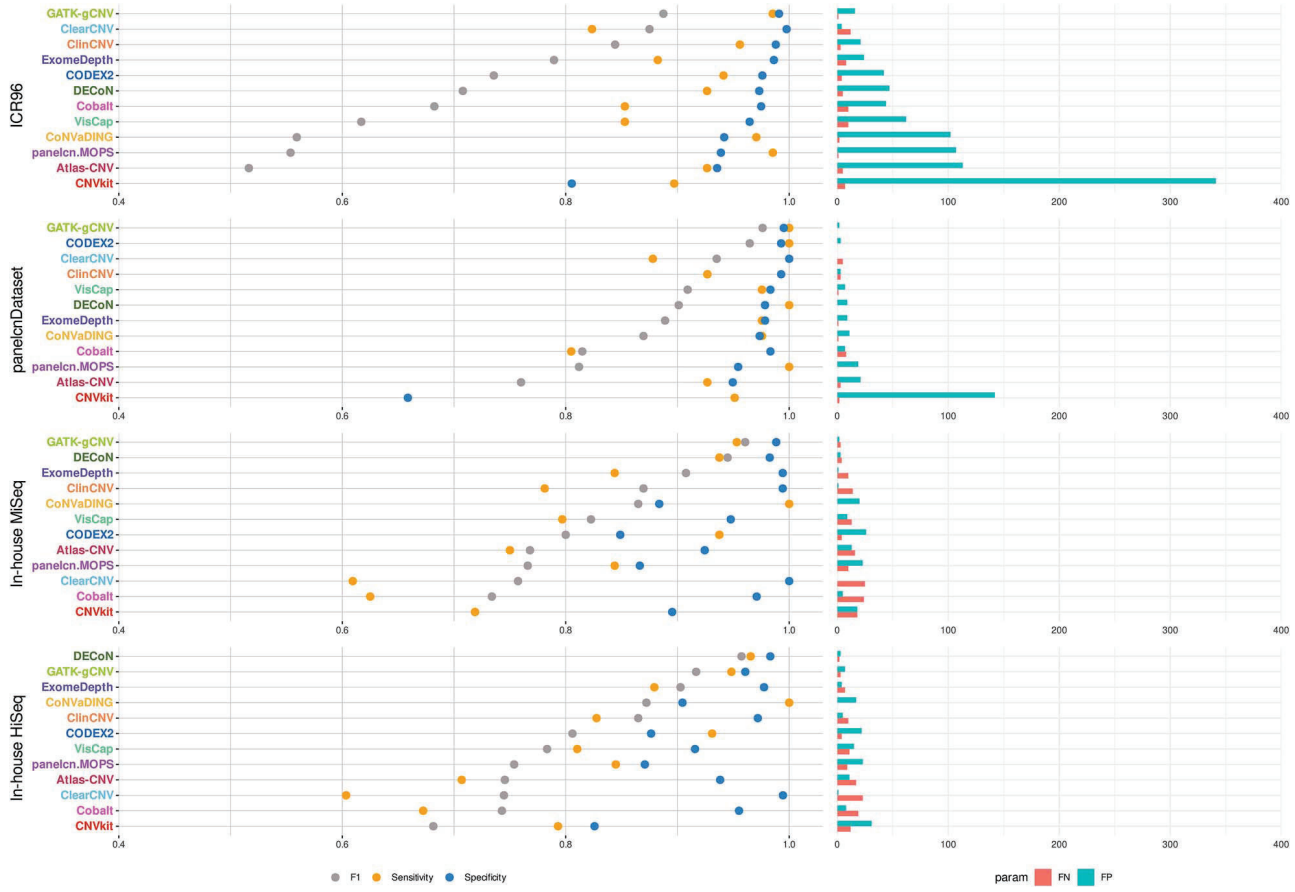


Figure 2. Benchmark results at the gene level. The tools were run using default parameters and are listed in descending order based on their F1 score in each dataset. The F1 scores for CNVkit on ICR96 and panelcnDataset, which were not included in the presented figure, are 0.26 and 0.35, respectively. (FN, false negative; FP, false positive; F1, F1 score).

Table 3. Suggested parameter values. Suggested parameter values to be used for improving the F1 score at the ROI level. The suggested value is the value within the optimal range closest to the default value. The mean F1 increase is calculated as the difference between the mean of the F1 scores obtained across datasets using the default value and the mean of the F1 scores obtained across datasets using the suggested value.

Tool	Parameter	Default value	Optimal range	Suggested value	Mean F1 increase
Atlas-CNV	threshold_dup	0.4	[0.44–0.6]	0.44	0.0040 (+0.76%)
ClearCNV	trans_prob	0.001	[0.0015–0.02]	0.0015	0.0006 (+0.07%)
ClinCNV	scoreG	20	[25–50]	25	0.0126 (+1.43%)
CNVkit	alpha (segmetrics)	0.05	[0.0001–0.04]	0.04	0.0006 (+0.09%)
CNVkit	drop-outliers	10	[1–4]	4	0.0080 (+1.26%)
Cobalt	high-depth-trim-frac	0.01	[0.025–0.1]	0.025	0.0033 (+0.50%)
Cobalt	var-cutoff	0.9	[0.91–0.99]	0.91	0.0053 (+0.81%)
CODEX2	cn_del_threshold	1.7	[1.3–1.67]	1.67	0.0172 (+2.14%)
CODEX2	cn_dup_threshold	2.3	[2.5–2.8]	2.5	0.0352 (+4.38%)
CODEX2	gc_thresh_down	20	[30–40]	30	0.0018 (+0.22%)
CoNVaDING	ratioCutOffLow	0.65	[0.5–0.6]	0.6	0.0168 (+2.28%)
panelcn.MOPS	CN3	1.46	[1.6–1.7]	1.6	0.0173 (+2,41%)
panelcn.MOPS	corrThresh	0.99	[0.5–0.985]	0.985	0.0017 (+0,24%)

Benchmark with default parameters

Several approaches can be used to measure tool performance. We have benchmarked tools using two levels of resolution, ROI and gene level, and several metrics such as F1 score, sensitivity, or specificity. These approaches facilitate the analysis of which tools are more suitable in each context. If we focus on the overall performance of the tool, the F1 score is a common metric used to evaluate the performance of binary classifiers. Based on this metric, we highlight GATK-gCNV and ClinCNV, which showed outstanding performance at the ROI level, with GATK-gCNV demonstrating superior performance at the gene level. Therefore, we suggest using ClinCNV and GATK-gCNV when the priority is to maximize the overall performance according to the F1 score and especially the latter when the focus is on the gene level. It is noteworthy that ClinCNV was also the fastest tool, requiring only 13 min to call CNVs. While run time is not typically the primary factor in

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

selecting a calling tool, it may be advantageous in settings where computational resources are limited or results must be delivered as quickly as possible.

On the other hand, sensitivity is a key metric in diagnostic settings, where the aim is usually to minimize the number of FNs. Also, in genetic diagnostics it is frequently useful to focus on the gene level because, if at least one exon from the CNV is detected, a subsequent MLPA test could be performed to confirm the CNV [17]. Focusing on the sensitivity and the per gene metrics, CoNVaDING, GATK-gCNV, CODEX2, and DECoN were among the best five in all datasets. Although we highlight the power of CoNVaDING to detect positives, with only three FNs across all datasets, it also produced a high number of FPs. In contrast, GATK-gCNV demonstrated high sensitivity and high specificity at the same time, which makes it a valuable candidate for use in genetic diagnostics. In any case, the aforementioned tools produced a relevant number of FPs. Since most diagnostic units validate CNV calls using orthogonal methods, the number of FPs should be taken into consideration to ensure the cost-effectiveness of diagnostic routines. On the opposite side, the highest rates of FNs were obtained by ClearCNV and Cobalt. This suggests that, when their default parameters are used, these tools may not be appropriate solutions for calling CNVs in genetic diagnostic settings from NGS panel data.

No previous work has evaluated the sensitivity and specificity of the two most highlighted tools discussed here, GATK-gCNV and ClinCNV, on gene panel data. Demidov *et al.* evaluated the performance of ClinCNV on WGS and WES data in their ClinCNV publication and only compared it with ExomeDepth and DELLY [26]. In the GATK-gCNV publication, the authors demonstrated that GATK-gCNV was capable of achieving 95% sensitivity in detecting CNVs of two or more exons [24]. However, GATK-gCNV was run on WES data, and the methodology differed from that used in the work presented here. Lepkes *et al.* included GATK-gCNV in their benchmark on gene panel data, but MLPA tests were performed after the benchmark execution, preventing the calculation of sensitivity and specificity [16].

GATK-gCNV workflows

The GATK guide to call rare germline CNVs includes two optional steps: AnnotateIntervals and FilterIntervals. To gain a deeper understanding of the effect of including them, we evaluated alternative GATK-gCNV workflows as detailed in Materials and Methods. Interestingly, the inclusion of the FilterIntervals step had a relevant impact on the performance: the workflows incorporating the FilterIntervals step clearly outperformed the one that excluded it. We consequently recommend including this step for the detection of rare germline CNVs from gene panel data. The GATK guide describes this step as optional but recommended, which may lead some users to overlook this step despite its impact on performance. We believe that the results observed here will encourage users to include the FilterIntervals step in their workflows. On the other hand, the inclusion of the AnnotateIntervals step, which entails explicit guanine and cytosine (GC)-content-based filtering, did not result in a discernible improvement in performance across all datasets. The GATK guide also described this step as optional but recommended, so we suggest users to further validate its effect on their own datasets.

Parameter evaluation

The available documentation on the effect of each parameter on tool performance is often scarce or nonexistent. Deciding which parameter to modify and how to tune it may be particularly challenging when a tool provides dozens of parameters, as is the case with GATK-gCNV. To address these issues, we repeated benchmark executions, modifying each parameter individually over a range of values. We generated 436 figures that can be used as reference guidance for research and diagnostic laboratories that are currently using or planning to use any of the tools in their settings. These results should help users to better understand the contribution of each parameter on tool performance, prevent them from inadvertently overlooking the most relevant parameters, and facilitate fine-tuning of tool parameters.

Parameters affected performance in multiple ways, and we grouped these changes into four main patterns. Some parameters had no effect on performance and should not be considered when trying to enhance tool performance. Other parameters resulted in the typical trade-off of binary classifiers: an increase in sensitivity leads to a decrease in specificity or vice versa. We recommend modifying this type of parameters to adjust the balance between sensitivity and specificity. Other parameters showed a bell-shaped behavior in sensitivity or specificity, suggesting that the tool performance could be optimized around a certain value. Finally, other parameters affected tool performance without a clear pattern, often with successive increases and decreases in sensitivity. We suggest modifying these parameters with caution as the observed variability makes it difficult to predict their impact on performance.

It is noteworthy that for certain parameters, the highest F1 scores across all datasets were observed on one side of the default parameter, either below or above it. Thus, we were able to identify the optimal range for 13 parameters, wherein the tools demonstrated superior performance compared to the default parameters. One possible explanation for this finding is that the tools were developed for specific datasets, which may restrict their applicability to other datasets, such as those used in this manuscript. Also, the authors may have optimized their tools based on performance metrics other than the F1 score. Anyway, while any value within this range could potentially improve the F1 score, the parameter value we suggested was the closest to the default value. Since it is the closest to the value set by the authors, we understand that this is the most conservative approach.

Using the suggested parameter values resulted in different F1 score increases. Although some yielded modest F1 score increments, such as the trans_prob and alpha parameters, others produced notable F1 score changes. Modifying the cn_dup_threshold, CN3, ratioCutOffLow, and cn_del_threshold parameters resulted in an average F1 increase of >2%. In any case, all suggested parameter values represent an opportunity to enhance the overall performance of the tools, and we recommend tool users to try them on their own datasets for further validation.

Combination of tool pairs

A common approach in bioinformatics is to join or intersect the results obtained by variant callers separately to produce new meta-callers [32–34]. In this work, no tool was capable of detecting all CNVs at the ROI or gene level with their default parameters. We therefore assessed the effect of tool unions and intersections on performance to determine if any meta-caller could achieve 100% sensitivity.

Although no union of tools detected all true positive ROIs in the per ROI results, five tool pairs did so at the gene level. These pairs may be employed in diagnostic scenarios where no true CNV should be overlooked. Indeed, if a CNV caller or meta-caller is capable of detecting all CNVs, it can be used as a screening step

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

prior to an orthogonal method validation, such as MLPA [35]. This approach has the potential to enhance the mutation detection yield and reduce costs in genetic testing for hereditary cancer [17]. In any case, we observed that the five tool pairs obtained largely different specificity values across the datasets. We hence encourage diagnostic units to conduct a thorough evaluation of their performance on their own in-house datasets.

Limitations

The results presented in this work have some limitations to note. First, the datasets used have an unusually high frequency of rare CNVs compared to the general population, where CNV frequency is expected to be considerably lower [36]. It would be of interest to assess the performance of the tools on datasets that would more accurately reflect the incidence of CNVs in the general population. Anyway, the datasets used in this benchmark provide a challenging scenario for the evaluation of the tools. Second, regarding the combination of tool pairs and the identification of suggested parameter values, we did not divide the datasets into training and validation subsets to assess whether the tool performance observed in a training subset was confirmed in a validation dataset. Hence, we recommend users to evaluate how tool pairs and suggested parameter values behave on their own datasets. Finally, we only evaluated combinations of tool pairs, leaving open the question of whether a combination of three or more tools could lead to a better meta-caller.

Conclusion

Here, we conducted a comprehensive evaluation of the current state of the art in germline CNV detection from gene panel data. Although the identification of CNVs remains challenging, our results indicate that certain tools can achieve very high performance. In terms of F1 score, ClinCNV and GATK-gCNV demonstrated superior calling performance compared to the other tools, with GATK-gCNV exhibiting high effectiveness in identifying true positives. The benchmark results, parameter evaluation, combination of tool pairs, and the CNVbenchmarkeR2 framework that we developed can serve as a valuable guide to research and diagnostic teams facing the task of detecting germline CNVs from gene panel data.

Key Points

- Comprehensive evaluation of 12 copy number variation callers on four real-validated datasets.
- ClinCNV and GATK-gCNV excelled, with GATK-gCNV achieving superior sensitivity.
- Assessment of the effect of modifying 107 tool parameters: 436 figures are publicly available.
- CNVbenchmarker2 enables users to conduct their own tool evaluations.

Acknowledgements

This study uses data from the ICR96 exon CNV validation series generated by Professor Nazneen Rahman's team at The Institute of Cancer Research, London as part of the TGMI. We are grateful to Katharina Wimmer's team at the Department of Human Genetics, Medical University of Innsbruck for providing access to the dataset deposited at EGA and hosted at EBI under accession number EGAS00001002481. We thank all patients and members of the Hereditary Cancer Program at ICO and the Clinical Genomics Unit at the HGTP. We also thank the IDIBELL IT unit and the ODAP-ICO IT team for their help.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare that they have no conflict of interest.

Funding

Research funded by the Instituto de Salud Carlos III FEDER – a way to build Europe – [PI23/00017, PI19/00553]; CIBERONC [CB16/12/00234]; by the Generalitat de Catalunya Pla estratègic de recerca i innovació en salut, PERIS, and the Agència de Gestió d'Ajust Universitaris i de Recerca, AGAUR, (2023SGR01112); and the Fundació La Marató de TV3 (202031-10). With the support of the 'Acció instrumental de formació de científics i tecnòlegs' (SLT017/20/000129) of the Departament de Salut de la Generalitat de Catalunya. We thank CERCA Programme/Generalitat de Catalunya for institutional support.

Data availability

The datasets underlying this article are available in the European Genome-Phenome Archive (EGA) and can be accessed with the following accession numbers: EGAD00001003335 for ICR96, EGAS00001002481 for panelcnDataset, and EGAS00001004316 for the In-house MiSeq / HiSeq datasets.

The CNVbenchmarkeR2 code is publicly available at https://github.com/jpuntomarcos/CNVbenchmarkeR2. Similarly, the complete set of 436 parameter evaluation figures is available for download at https://doi.org/10.6084/m9.figshare.25930960.

References

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**:363–76. https://doi.org/10.1038/nrg2958.
2. Pös O, Radvanszky J, Styk J. *et al.* Copy number variation: Methods and clinical applications. *Appl Sci* 2021;**11**:819.
3. Zarrei M, MacDonald JR, Merico D. *et al.* A copy number variation map of the human genome. *Nat Rev Genet* 2015;**16**:172–83. https://doi.org/10.1038/nrg3871.
4. Pinto D, Pagnamenta AT, Klei L. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010;**466**:368–72. https://doi.org/10.1038/nature09146.
5. Wheeler E, Huang N, Bochukova EG. *et al.* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet* 2013;**45**:513–7. https://doi.org/10.1038/ng.2607.
6. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009;**1**:1–9.
7. Valsesia A, Macé A, Jacquemont S. *et al.* The growing importance of CNVs: New insights for detection and clinical interpretation. *Front Genet* 2013;**4**:92.
8. Kerkhof J, Schenkel LC, Reilly J. *et al.* Clinical validation of copy number variant detection from targeted next-generation sequencing panels. *J Mol Diagn* 2017;**19**:905–20.
9. Singh AK, Olsen MF, Lavik LAS. *et al.* Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med Genet* 2021;**14**:1–12.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

10. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51. https://doi.org/10.1038/nrg.2016.49.
11. Johansson LF, van Dijk F, de Boer EN. *et al.* CoNVaDING: Single exon variation detection in targeted NGS data. *Hum Mutat* 2016;**37**:457–64. https://doi.org/10.1002/humu.22969.
12. Fowler A, Mahamdallie S, Ruark E. *et al.* Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Res* 2016;**1**:20.
13. Povysil G, Tzika A, Vogt J. *et al.* Panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Hum Mutat* 2017;**38**:889.
14. Kim HY, Choi JW, Lee JY. *et al.* Gene-based comparative analysis of tools for estimating copy number alterations using whole-exome sequencing data. *Oncotarget* 2017;**8**:27277.
15. Chiang T, Liu X, Wu TJ. *et al.* Atlas-CNV: A validated approach to call single-exon CNVs in the eMERGESeq gene panel. *Genet Med J Am Coll Med Genet* 2019;**21**:2135–44. https://doi.org/10.1038/s41436-019-0475-4.
16. Lepkes L, Kayali M, Blümcke B. *et al.* Performance of In silico prediction tools for the detection of germline copy number variations in cancer predisposition genes in 4208 female index patients with familial breast and ovarian cancer. *Cancers (Basel)* 2021;**13**:1–12.
17. Moreno-Cabrera JM, del Valle J, Castellanos E. *et al.* Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *Eur J Hum Genet* 2020;**28**:1645–55. https://doi.org/10.1038/s41431-020-0675-z.
18. Roca I, González-Castro L, Fernández H. *et al.* Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res Rev Mutat Res* 2019;**779**:114–25. https://doi.org/10.1016/j.mrrev.2019.02.005.
19. Mahamdallie S, Ruark E, Yost S. *et al.* The ICR96 exon CNV validation series: A resource for orthogonal assessment of exon CNV calling in NGS data. *Wellcome Open Res* 2017;**2**:35.
20. Castellanos E, Gel B, Rosas I. *et al.* A comprehensive custom panel design for routine hereditary cancer testing: Preserving control, improving diagnostics and revealing a complex variation landscape. *Sci Rep* 2017;**7**:1–12.
21. Li H, Durbin R. Fast and accurate short read alignment with burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754.
22. Li H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. arXiv preprint arXiv 2013;1303.3997v2.
23. May V, Koch L, Fischer-Zirnsak B. *et al.* ClearCNV: CNV calling from NGS panel data in the presence of ambiguity and noise. *Bioinformatics* 2022;**38**:3871–6. https://doi.org/10.1093/bioinformatics/btac418.
24. Babadi M, Fu JM, Lee SK. *et al.* GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat Genet* 2023;**55**:1589–97. https://doi.org/10.1038/s41588-023-01449-0.
25. O'Fallon B, Durtschi J, Kellogg A. *et al.* Algorithmic improvements for discovery of germline copy number variants in next-generation sequencing data. *BMC Bioinformatics* 2022;**23**:1–14.
26. Demidov G, Sturm M, Ossowski S. ClinCNV: Multi-sample germline CNV detection in NGS data. *bioRxiv* 2022. 2022.06.10.495642.
27. Talevich E, Shain AH, Botton T. *et al.* CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;**12**:e1004873.
28. Pugh TJ, Amr SS, Bowser MJ. *et al.* VisCap: Inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet Med* 2016;**18**:712–9. https://doi.org/10.1038/gim.2015.156.
29. Fowler A. DECoN: A detection and visualization tool for Exonic copy number variants. *Methods Mol Biol* 2022;**2493**:77–88. https://doi.org/10.1007/978-1-0716-2293-3_6.
30. Plagnol V, Curtis J, Epstein M. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics (Oxford, England)* 2012;**28**:2747–54. https://doi.org/10.1093/bioinformatics/bts526.
31. Jiang Y, Wang R, Urrutia E. *et al.* CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol* 2018;**19**:202.
32. Samarakoon PS, Sorte HS, Kristiansen BE. *et al.* Identification of copy number variants from exome sequence data. *BMC Genomics* 2014;**15**:1–11.
33. Liu S, Tsai WH, Ding Y. *et al.* Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res* 2016;**44**:e47–7.
34. Gabrielaite M, Torp MH, Rasmussen MS. *et al.* A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancer* 2021;**13**:6283.
35. Shen Y, Wu BL. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. *J Genet Genomics* 2009;**36**:257–65. https://doi.org/10.1016/S1673-8527(08)60113-7.
36. Ruderfer DM, Hamamsy T, Lek M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* 2016;**48**:1107–11. https://doi.org/10.1038/ng.3638.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae645/7922578 by guest on 25 December 2024

© The Author(s) 2024. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental File

Benchmark evaluation metrics

At the per ROI level, individual ROIs were assessed as standalone entities. They were categorized with correctness labels: True Positive (TP) or True Negative (TN) when the tool outcome was consistent with the MLPA result, False Negative (FN) when the tool missed a CNV identified by MLPA, and False Positive (FP) when the tool erroneously reported a CNV not detected by MLPA. This level of evaluation provides the most detailed assessment.

At the per gene level, since most MLPA kits cover entire genes, genuine CNVs would be confirmed by MLPA when any CNV call was verified within any ROI of the affected gene. Therefore, the per gene metrics assigned a correctness label to each gene, encompassing all its exons: TP if at least one of its ROIs proved to be a TP; FN if MLPA detected a CNV in at least one ROI and none were identified by the tool; FP if the tool indicated a CNV in at least one ROI and none were confirmed by MLPA; TN if neither MLPA nor the tool detected a CNV within any of its ROIs.

Multiple performance metrics were computed for each tool-dataset execution. Sensitivity was defined as

TP/(TP + FN), specificity was defined as TN/(TN + FP), positive predictive value (PPV) was defined as TP/(TP + FP), negative predictive value (NPV) was defined as TN/(TN + FN), false negative rate (FNR) was defined as FN/(FN + TP), false positive rate (FPR) was defined as FP/(FP + TN), F1 score (F1) was defined as 2TP/(2TP + FP + FN), accuracy was defined as (TP + TN) / (TP + FN + FP + TN), Matthews correlation coefficient (MCC) was defined as √(sensitivity × specificity × PPV × NPV) - √(FNR × FPR × (1 - NPV) × FDR) and Cohen's kappa coefficient was defined as (2 x (TP x TN – FN x FP)) / ((TP + FP) x (FP + TN) x (TP + FN) x (FN + TN)).

Tool and dataset selection

We followed two approaches for selecting tools for our analysis. First, we conducted searches on Google Scholar, using combinations of keywords including (CNV | CNA | copy number variants | copy number alterations) & (panel data | targeted panels | targeted gene panels) & germline. In addition to this search, we reviewed several articles that either cited benchmark or reviews for CNVs or referenced other CNV tools specifically used in panel data applications.

Dataset selection was performed through 1) review of the data used in germline CNV caller manuscripts, 2) search through Google Scholar and 3) search in the European Genome-phenome Archive (EGA) repository. Only four datasets met the criteria to be included in this manuscript.

Samples used in dataset panelcnDataset

The dataset EGAS00001002481 contains 170 samples. However, 9 samples were removed to form the final dataset panelcnDataset used in this benchmark. The excluded samples showed alterations out of the scope of this benchmark: 5 presented CNVs smaller than an exon (IBK9, IBK23, IBK67, IBK153, IBK166) and 4 contained ALUs insertions instead of CNVs (IBK141, IBK142, IBK143, IBK151). IBK141 ALU insertion was identified in our previous work (*Moreno-Cabrera et al.*, 2020).

Bed files generation

For TruSight-based datasets, ICR96 and panelcnDataset, we employed a modified version of the target bed file previously published (Fowler et al. 2016). The adjustments included the removal of the fourth column, addition of a gene column, join of overlapping regions, and ultimately sorting the bed file by chromosome and start position. For in-house datasets, we generated a target bed file encompassing all coding exons derived from protein-coding transcripts within the I2HCP panel. These genomic coordinates were extracted from Ensembl BioMart version 108.

ClinCNV used an annotated version of the bed files which included a GC content column. This bed files were generated using the BedAnnotateGC tool.

DECoN installation

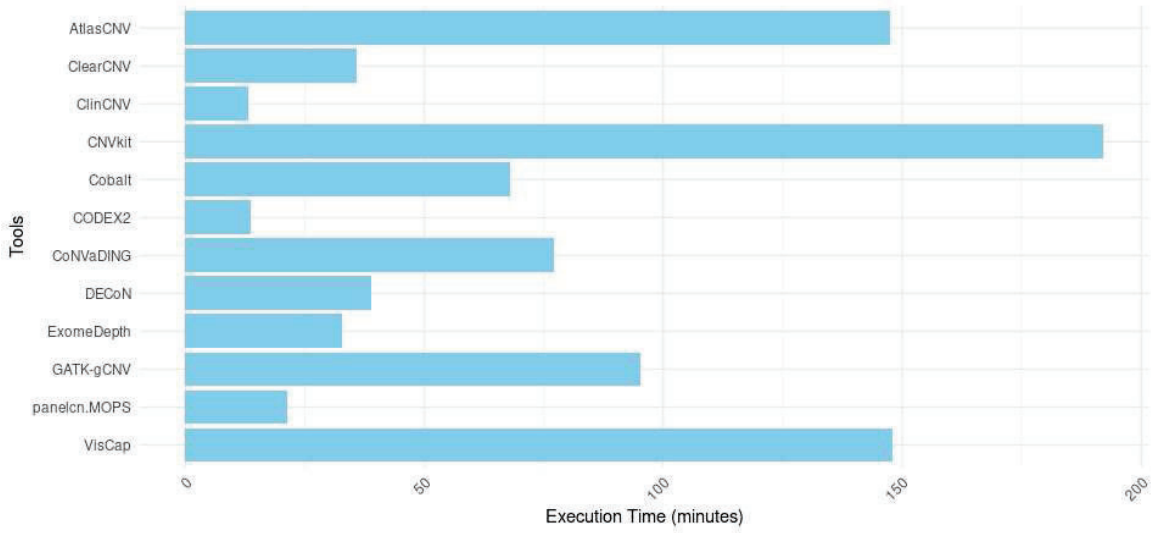
For the execution of DECoN, we modified the package environment's configuration due to compatibility issues encountered with ExomeDepth version 1.1.15. Specifically, we updated the renv.lock file to specify the use of ExomeDepth version 1.1.16. This adjustment addresses the problem reported in issue #45 on the DECoN GitHub repository (https://github.com/RahmanTeam/DECoN/issues/45).

CODEX2 code adaptation

The CODEX2 parameters were obtained from the CODEX2 targeted demo that the authors published. In order to further parameterize CODEX2, we also added the cn_del_threshold and cn_dup_threshold parameters, whose default values in the original demo were 1.7 and 2.3, respectively.

Run-times

We measured tool run-times on the ICR96 dataset using a workstation with 24 GB RAM and 1 CPU per job.



Tool run-times (minutes) using 24GB of memory and 1 core. The data was collected employing the default parameters on the ICR96 dataset. Values: Atlas-CNV 147,47; ClearCNV 35,92; ClinCNV 13,15; CNVkit 192,05; Cobalt 68,02; CODEX2 13,73; CoNVaDING 77,03; DECoN 38,93; ExomeDepth 32,72; GATK-gCNV 95,17; panelcn.MOPS 21,40; VisCap 147,90.

GATK-gCVN execution excluding the optional explicit GC-content-based filtering step

We followed the steps indicated in the GATK guide to call rare germline variants. In this guide, section 2, which includes AnnotateIntervals and FilterIntervals steps, is optional. Moreover, the AnnotateIntervals step, which involves GC-content-based filtering, is also described as optional within section 2. We hence decided to compare three workflows:

- GATK-gCNV. It is the final GATK-gCNV workflow used in all benchmark sections and includes AnnotateIntervals and FilterIntervals steps.
- GATK-gCNV_no_AI. It excludes the AnnotateInterval step, which is the default approach in the GATK WDL germline cohort workflow.
- GATK-gCNV_no_AI_FI. It excludes both optional AnnotateIntervals and FilterIntervals steps.

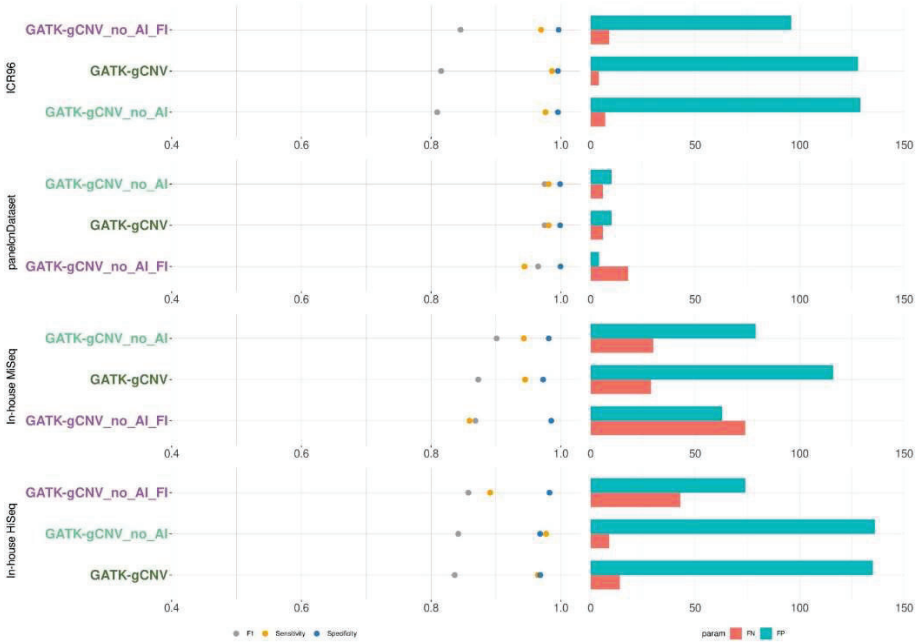
The following tables and figures show a performance comparison of three GATK-gCNV workflows.

Performance at ROI level

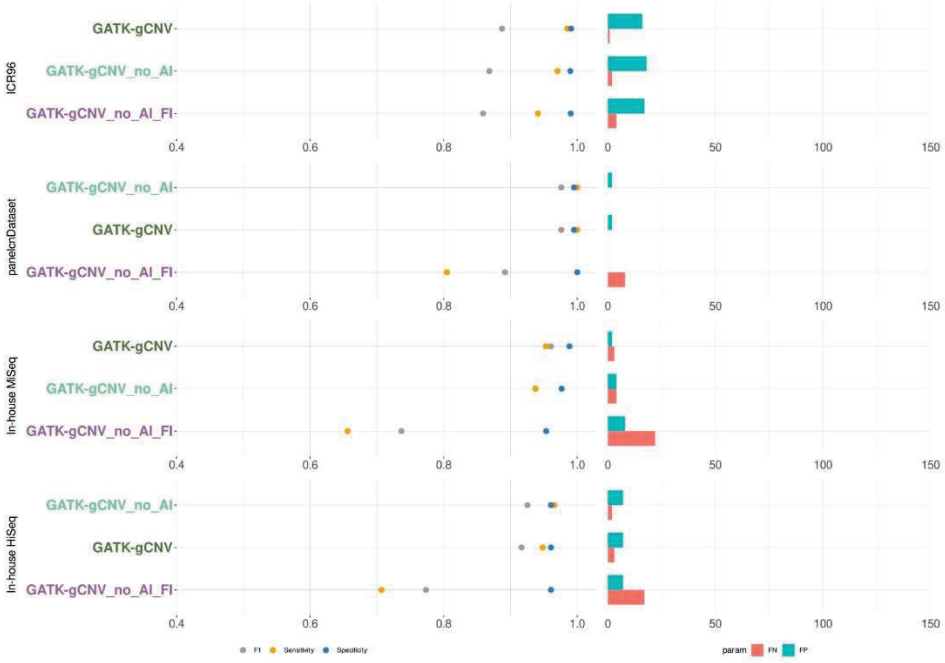
dataset	algorithm	TP	TN	FP	FN	total	sensitivity	specificity	F1
ICR96	GATK-gCNV	292	28382	128	4	28806	0.9865	0.9955	0.8156
	GATK-gCNV_no_AI	289	28381	129	7	28806	0.9764	0.9955	0.8095
	GATK-gCNV_no_AI_FI	287	28414	96	9	28806	0.9696	0.9966	0.8454
panelcnDataset	GATK-gCNV	315	9476	10	6	9807	0.9813	0.9989	0.9752
	GATK-gCNV_no_AI	315	9476	10	6	9807	0.9813	0.9989	0.9752
	GATK-gCNV_no_AI_FI	303	9482	4	18	9807	0.9439	0.9996	0.9650
inHouseMiSeq	GATK-gCNV	497	4128	116	29	4770	0.9449	0.9727	0.8727
	GATK-gCNV_no_AI	496	4165	79	30	4770	0.943	0.9814	0.901
	GATK-gCNV_no_AI_FI	452	4181	63	74	4770	0.8593	0.9852	0.8684
inHouseHiSeq	GATK-gCNV	381	4123	135	14	4653	0.9646	0.9683	0.8364
	GATK-gCNV_no_AI	386	4122	136	9	4653	0.9772	0.9681	0.8419
	GATK-gCNV_no_AI_FI	352	4184	74	43	4653	0.8911	0.9826	0.8575

Performance at gene level

dataset	algorithm	TP	TN	FP	FN	total	sensitivity	specificity	F1
ICR96	GATK-gCNV	67	1736	16	1	1820	0.9853	0.9909	0.8874
	GATK-gCNV_no_AI	66	1734	18	2	1820	0.9706	0.9897	0.8684
	GATK-gCNV_no_AI_FI	64	1735	17	4	1820	0.9412	0.9903	0.8591
panelcnDataset	GATK-gCNV	41	414	2	0	457	1.0000	0.9952	0.9762
	GATK-gCNV_no_AI	41	414	2	0	457	1.0000	0.9952	0.9762
	GATK-gCNV_no_AI_FI	33	416	0	8	457	0.8049	1.0000	0.8919
inHouseMiSeq	GATK-gCNV	61	170	2	3	236	0.9531	0.9884	0.9606
	GATK-gCNV_no_AI	60	168	4	4	236	0.9375	0.9767	0.9375
	GATK-gCNV_no_AI_FI	42	164	8	22	236	0.6562	0.9535	0.7368
inHouseHiSeq	GATK-gCNV	55	171	7	3	236	0.9483	0.9607	0.9167
	GATK-gCNV_no_AI	56	171	7	2	236	0.9655	0.9607	0.9256
	GATK-gCNV_noAI_FI	41	171	7	17	236	0.7069	0.9607	0.7736

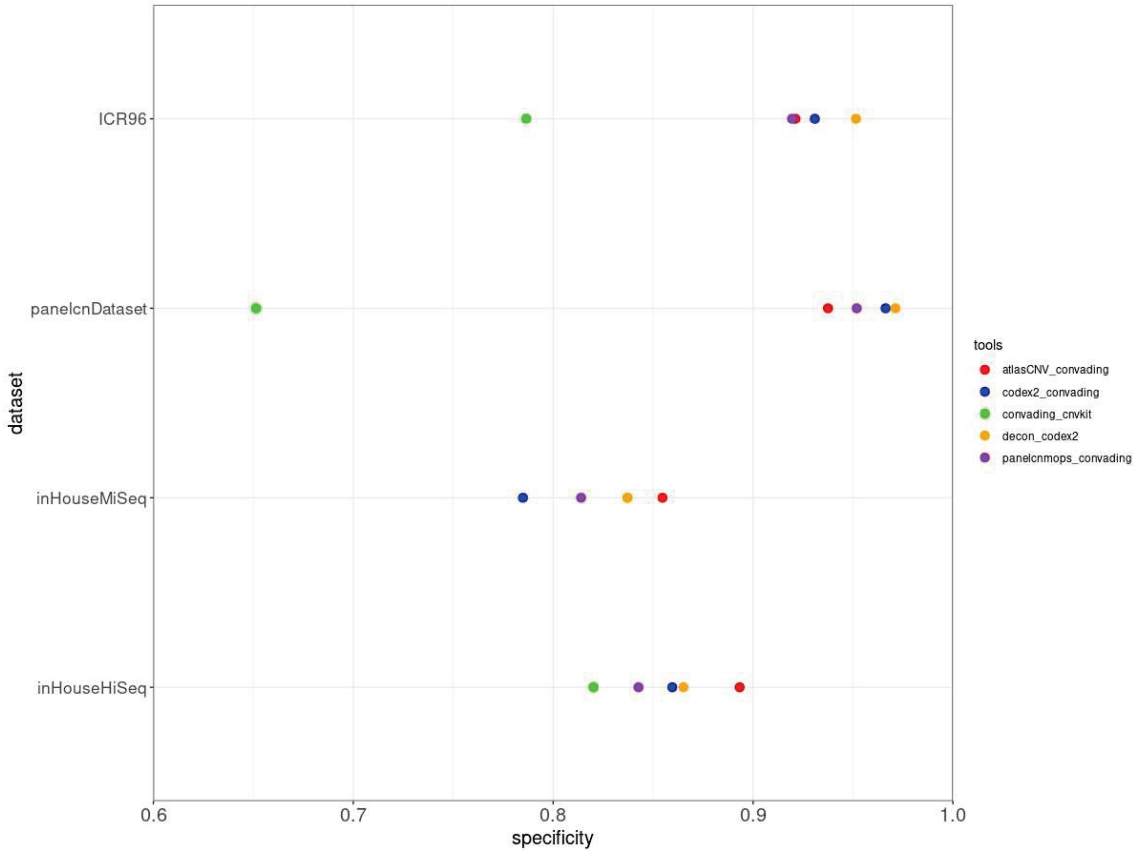


Comparison of three GATK-gCNV workflows at ROI level. GATK-gCNV is the final GATK-gCNV workflow used in all benchmark sections and includes AnnotateIntervals and FilterIntervals steps. GATK-gCNV_no_AI excludes the AnnotateInterval step, which is the default approach in the GATK WDL germline cohort workflow. GATK-gCNV_no_AI_FI excludes AnnotateIntervals and FilterIntervals steps. Results are ranked according to their F1 scores in each dataset. (FN false negative; FP false positive; F1 F1 score).



Comparison of three GATK-gCNV workflows at gene level. GATK-gCNV is the final GATK-gCNV workflow used in all benchmark sections and includes AnnotateIntervals and FilterIntervals steps. GATK-gCNV_no_AI excludes the AnnotateInterval step, which is the default approach in the GATK WDL germline cohort workflow. GATK-gCNV_no_AI_FI excludes AnnotateIntervals and FilterIntervals steps. Results are ranked according to their F1 scores in each dataset. (FN false negative; FP false positive; F1 F1 score).

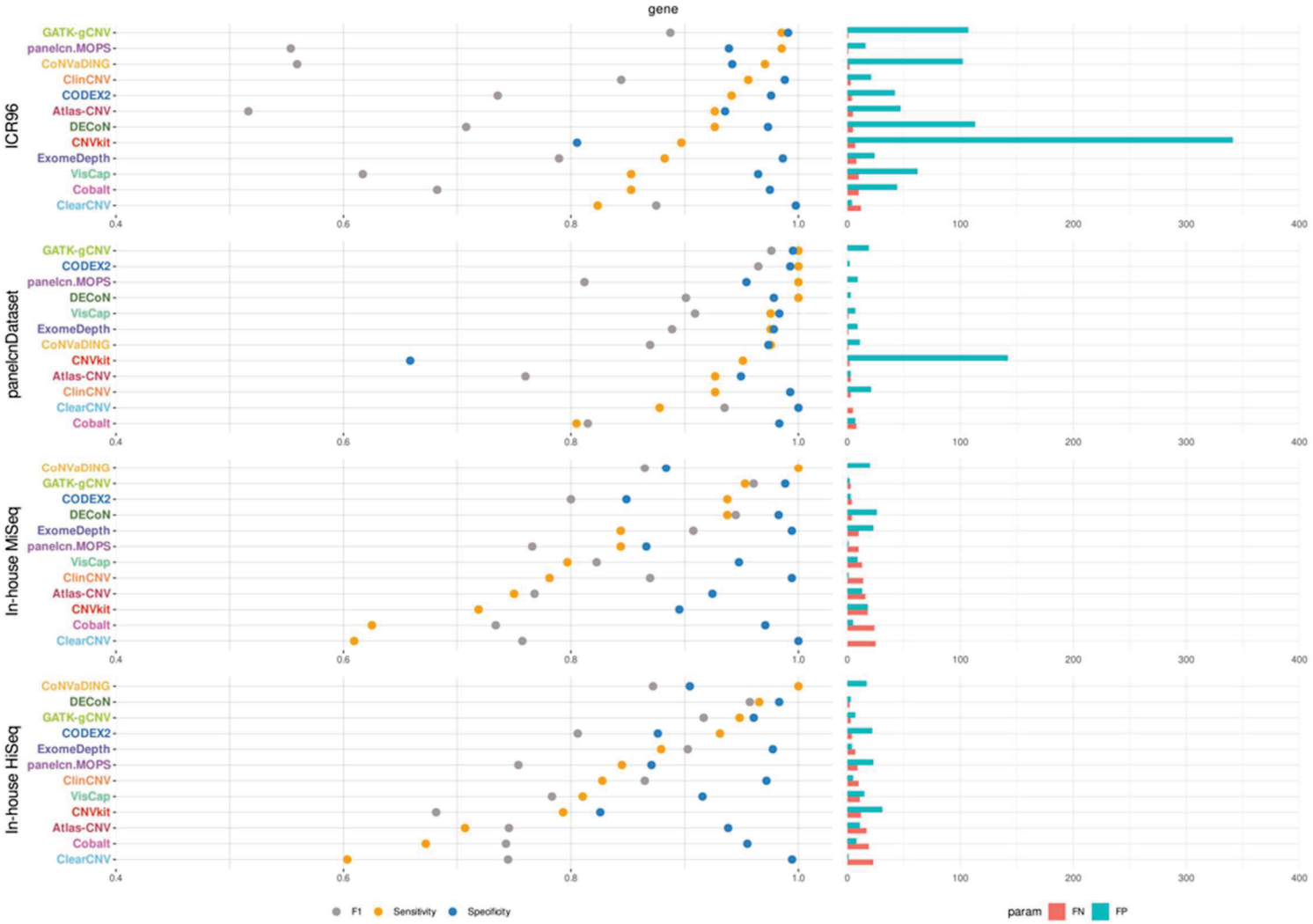
Tools achieving sensitivity 1 across all datasets at the gene level



Specificity of tool pairs achieving perfect sensitivity across all datasets at the gene level. In the InhouseMiSeq dataset, convading_cnvkit achieved the same specificity as decon_codex2, hence only 4 points are shown.

Supplementary Figure 1: Per ROI results sorted by sensitivity .

Supplementary Figure 2:Per gene results sorted by sensitivity .



Article publicat 3

vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines

Elisabet Munté, Lidia Feliubadaló, Marta Pineda, Eva Tornero, Maribel Gonzalez, José Marcos Moreno-Cabrera, Carla Roca, Joan Bales Rubio, Laura Arnaldo, Gabriel Capellá, Jose Luis Mosquera* i Conxi Lázaro*.

Bioinformatics. 2023 Mar 1;39(3) :btad128.

DOI: 10.1093/bioinformatics/btad128

Bioinformatics, 39(3), 2023, btad128
https://doi.org/10.1093/bioinformatics/btad128
Advance Access Publication Date: 14 March 2023
Original Paper

OXFORD

Sequence analysis

vaRHC: an R package for semi-automation of variant classification in hereditary cancer genes according to ACMG/AMP and gene-specific ClinGen guidelines

Elisabet Munté [1], Lidia Feliubadaló [1,2,*], Marta Pineda[1,2], Eva Tornero[1], Maribel Gonzalez[1], José Marcos Moreno-Cabrera [1], Carla Roca[1], Joan Bales Rubio[3], Laura Arnaldo[1], Gabriel Capellá[1,2], Jose Luis Mosquera [4,*], Conxi Lázaro[1,2,*]

[1]Hereditary Cancer Program, Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), IDIBELL, Catalan Institute of Oncology, L'Hospitalet de Llobregat 08908, Spain
[2]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain
[3]Department of Information Technologies, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat 08908, Spain
[4]Department of Bioinformatics, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat 08908, Spain

*Corresponding authors. Hereditary Cancer Program, Program in Molecular Mechanisms and Experimental Therapy in Oncology (Oncobell), IDIBELL, Catalan Institute of Oncology, L'Hospitalet de Llobregat 08908, Spain. E-mail: lfeliubadalo@iconcologia.net or clazaro@iconcologia.net and Department of Bioinformatics, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat 08908, Spain. E-mail: jmosquera@idibell.cat
Associate Editor: Can Alkan

Received on 7 November 2022; revised on 10 February 2023; accepted on 2 March 2023

Abstract

Motivation: Germline variant classification allows accurate genetic diagnosis and risk assessment. However, it is a tedious iterative process integrating information from several sources and types of evidence. It should follow gene-specific (if available) or general updated international guidelines. Thus, it is the main burden of the incorporation of next-generation sequencing into the clinical setting.

Results: We created the vaRiants in HC (vaRHC) R package to assist the process of variant classification in hereditary cancer by: (i) collecting information from diverse databases; (ii) assigning or denying different types of evidence according to updated American College of Molecular Genetics and Genomics/Association of Molecular Pathologist gene-specific criteria for *ATM*, *CDH1*, *CHEK2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *PTEN*, and *TP53* and general criteria for other genes; (iii) providing an automated classification of variants using a Bayesian metastructure and considering CanVIG-UK recommendations; and (iv) optionally printing the output to an .xlsx file. A validation using 659 classified variants demonstrated the robustness of vaRHC, presenting a better criteria assignment than Cancer SIGVAR, an available similar tool.

Availability and implementation: The source code can be consulted in the GitHub repository (https://github.com/emunte/vaRHC) Additionally, it will be submitted to CRAN soon.

1 Introduction

Cancer is a main public health problem and a leading cause of death (Siegel et al. 2022). Around 5%–10% of cancers worldwide are attributable to hereditary cancer (HC) syndromes (Nagy et al. 2004). HC patients harbour pathogenic germline variant(s) in cancer predisposition genes making them prone to develop multiple primary cancers at younger ages. Early identification of these individuals allows us to personalize their risk assessment, adapt their clinical follow-up, provide some targeted therapies, and offer cascade testing to relatives.

The use of next-generation sequencing (NGS) in diagnostics expands the number of genes analysed in a single test, increasing the diagnostic yield (Tung et al. 2016) but also the identification of variants of unknown significance (Lumish et al. 2017; Feliubadaló et al. 2019). Accurate variant classification is a huge challenge and a main burden of the incorporation of NGS into the clinical setting; only a correct classification allows proper genetic diagnosis and personalized risk assessment. Nowadays, variant classification is a time-consuming process that combines different type of evidence such as variant consequence, population frequencies, functional assays, *in silico* predictors,

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

and co-segregation studies. Moreover, it is iterative due to continuous information updates and guideline refinement, enforcing periodic revisions of variant classification.

In 2015, the American College of Molecular Genetics and Genomics (ACMG) with the Association of Molecular Pathologists (AMP) published generic guidelines to standardize and provide an objective framework for evaluating variant pathogenicity in Mendelian disorders (Richards et al. 2015). However, some criteria proposed were qualitative and indefinite, allowing discrepancies in variant interpretation between laboratories (Amendola et al. 2016). Later, specific guidelines were published for some genes by collaborative groups or expert panels like Clinical Genome Resource (ClinGen) to adjust variant classification to gene and disease particularities (https://clinicalgenome.org/; accessed November 2022). Moreover, it was demonstrated that the criteria combination in ACMG/AMP guidelines was compatible with a quantitative Bayesian formulation (Tavtigian et al. 2018), which was later abstracted to a naturally scaled point system (Tavtigian et al. 2020). Additionally, CanVIG-UK consensus recommendations proposed some limitations to overlapping criterion combinations to avoid double counting of evidence (Garrett et al. 2021, https://www.cangene-canvaruk.org/_files/ugd/ed948a_f64f11f58e644521bc88f0b4ef1f5d01.pdf; accessed November 2022).

Different programmes have been developed to automatize variant classification by integrating different types of information. Most tools are based on ACMG/AMP general rules, like InterVar (Li and Wang 2017), PathoMAN (Joseph et al. 2017; Ravichandran et al. 2019), ClinGen Pathogenicity Calculator (Patel et al. 2017), CharGer (Scott et al. 2019), Varsome (Kopanos et al. 2019), or Franklin (https://franklin.genoox.com). Other tools focus on a set of genes like CardioClassifier (Whiffin et al. 2018) and CardioVai (Nicora et al. 2018) for inherited cardiac conditions or Cancer Predisposition Sequencing Reporter (CPSR) (Nakken et al. 2021) and Cancer-SIGVAR (Li et al. 2021) for cancer predisposition genes. CPSR uses SherLoc algorithm (Nykamp et al. 2017) that provided several refinements to the original guidelines. Cancer-SIGVAR uses a ClinGen update of ACMG/AMP guidelines (Abou Tayoun et al. 2018), considers ClinGen's specific guidelines for *PTEN* (Mester et al. 2018), *CDH1* (Lee et al. 2018), RASopathies (Gelb et al. 2018), *RUNX1* (Luo et al. 2019) and hearing loss (Oza et al. 2018), and other sources (https://www.acgs.uk.com/media/11285/uk-practice-guidelines-for-variant-classification-2019-v1-0-3.pdf). However, specific guidelines or recommendations have been published for other HC genes like *ATM* (https://www.clinicalgenome.org/site/assets/files/7451/clingen_hbop_acmg_specifications_atm_v1_1.pdf; accessed November 2022), *CHEK2* (Vargas-Parra et al. 2020), *TP53* (Fortuno et al. 2021), and mismatch repair genes (MMRs) (https://www.insight-group.org/content/uploads/2021/11/DRAFT_Nov_2021_TEMPLATE_SVI.ACMG_Specifications_InSiGHT_MMR_V1.pdf; accessed November 2022). A deeper comparative of the above-mentioned tools can be found in Supplementary Table S1. Although automated tools aid the variant interpretation journey, the curator is still needed for proper integration of some clinical, genetic, functional, and literature information. Therefore, not all criteria can be fully automated.

Here, we introduce vaRiants in HC (varHC), an R package developed to automate, as far as possible, the variant classification process for HC genes. The Catalan Institute of Oncology is a monographic cancer centre, and our diagnostics laboratory is dedicated to HC testing. Accordingly, we aimed to increase accuracy in variant classification by automated collection and combination of data, and assignation of several criteria according to gene-specific guidelines for *ATM*, *CDH1*, *CHEK2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *PTEN* and *TP53*, and the updated general ACMG/AMP rules for other cancer susceptibility genes.

2 Materials and methods

2.1 vaRHC package

The vaRHC package was conceived for the statistical computing environment R (v4.1.2), using functions from both R/Bioconductor and CRAN packages (more details in https://github.com/emunte/vaRHC).

2.2 Criteria analysis

ACMG/AMP original guidelines proposed 28 different criteria to evaluate pathogenicity or benignity. Not all criteria have the same strength: pathogenic criteria can be very strong (PVS1), strong (PS1–PS4), moderate (PM1–PM6), or supporting (PP1–PP5), whereas benign criteria can be standalone (BA1), strong (BS1–BS4), or supporting (BP1–BP7) (Richards et al. 2015). ClinGen later further specified these criteria. For instance, a decision-tree algorithm was developed to adapt PVS1 strength according to type and position of loss of function (LoF) variants (Abou Tayoun et al. 2018). PP5 and BP6 criteria, relying on reputable source classification without access to primary data, could lead to errors and double counting and were removed (Biesecker and Harrison 2018). Hence, many of the 28 criteria were changed, deleted, or extended to different weights depending on the gene.

Individual analysis of the criteria was performed to determine which could be fully automated, partially automated (requiring manual curation), not automatable, and which do not apply to some or all HC genes. General and gene-specific criteria used by vaRHC and their combinations are detailed in Supplementary Tables S2 and S3, respectively.

2.3 Criteria combination for variant classification

Tavtigian et al.'s (2020) naturally scaled point system was used to calculate the final classification of variants. In this approach, each pathogenic criterion achieving supporting, moderate, strong, or very strong strength sums 1, 2, 4, or 8 points, respectively, to the final score, and each criterion in favour of benignity subtracts these same values. The final score determines the classification of the variant ($\geq$10: pathogenic; 6–9: likely pathogenic; 5–0: unknown significance; (-1)–(-5): likely benign; and ≤ -6: benign).

2.4 Information retrieval and database

vaRHC leverages several existing databases and programmes. Due to their different conception and implementation, the information is retrieved from those sources in real time or queried in a local relational database implemented with MySQL (8.0.28-0ubuntu0.20.04.3 for Linux on x86_64).

2.4.1 Local queries

Some source databases display stable information, with a renewal frequency of over 1 year. Hence, they are not queried real-time to decrease execution time and avoid possible issues due to their webpage maintenance. Instead, vaRHC is fed with a relational database (MySQL) gathering the information. The database is maintained at the Institut de Recerca Biomèdica de Bellvitge. Since SpliceAI and Provean predictors have recently retired their web-based or API version, the only way to obtain their scores is by downloading the software. To mitigate the difficulties that may cause installing extra software, several scores have been precalculated. The types of variants with precalculated scores and the complete information source list are found in Supplementary Methods.

2.4.2 Real-time queries

Other databases, like ClinVar (Landrum et al. 2014, 2018) and InSiGHT (https://www.insight-group.org/variants/databases/; accessed November 2022) are updated weekly or monthly. To access the latest information, vaRHC queries them real-time via web scrapping, which it also uses to interrogate web interfaces from databases without download options. Online programmes without pre-computed databases are queried via REST API. The complete list of databases queried real-time is in Supplementary Methods.

2.4.3 Customizable parameters

For easy customization, our local database contains a table with general and gene-specific cut-offs used for several criteria (BA1, BS1, PM2, PP3, BP4, and BP7). The user can provide vaRHC a txt file with custom values to change the default ones. See https://

github.com/emunte/vaRHC/blob/main/data/gene_specific.txt to download a template.

2.5 Performance assessment

Classified variants from *ATM* ($n = 32$), *CDH1* ($n = 279$), *PTEN* ($n = 139$), and *TP53* ($n = 118$) were downloaded from the ClinGen evidence repository (https://erepo.genome.network/evrepo/, March 2022). For *CHEK2*, 13 variants classified in Vargas-Parra (2020) were used. An in-house dataset of 78 variants classified according to 'ClinGen InSiGHT Hereditary Colorectal Cancer/Polyposis Variant Curation Expert Panel Specifications to the ACMG/AMP Variant Interpretation Guidelines Version DRAFT 1' was used for MMR genes.

The comparison between vaRHC results and manual classification of the aforementioned variants was performed criterion by criterion, instead of only considering the final variant classification. Outcomes were grouped into nine scenarios for each criterion taken (Table 1).

2.6 Benchmark dataset

CDH1 and *PTEN* variant datasets were also analysed with CancerSIGVAR (Li et al. 2021) using default settings. The number of differences between both software was statistically evaluated using the Kappa test from vcd (v 1.4-10) CRAN package; *P*-value was adjusted using Benjamini–Hochberg correction for multiple comparisons (Benjamini and Hochberg 1995). Significance for the adjusted *P*-value was 0.05.

3 Results

3.1 vaRHC package

The vaRHC package classifies single-nucleotide substitutions, deletions, and insertions up to 25-bp, intronic variants, and untranslated region variants.

Our package has a main function *vaR()* acting as a wrapper for three functions: *vaRinfo*, *vaRclass*, and vaRreport (Fig. 1). From the input of a gene, a transcript (RefSeq ID), and a variant name (in coding DNA nomenclature; http://varnomen.hgvs.org/), vaRinfo gathers relevant information from diverse sources. The second function vaRclass() uses the output of vaRinfo() to apply updated ACMG/AMP criteria considering gene specificities to calculate the different criteria met by the variant and explains the assignment or rejection of each criterion. Furthermore, it returns a final variant classification using Tavtigian's Bayesian metastructure and most CanVIG-UK recommendations (Supplementary Table S3). Lastly, vaRreport() generates a user-friendly .xlsx file to examine and store results, allowing non-bioinformatic users to work with them and modify the file, adding their considerations or information regarding non-automatable criteria.

For users wanting to classify variants in batch, vaRbatch() function has been created to interrogate vaR() function sequentially. The input can be either a dataframe with variants in coding DNA nomenclature or a variant call format file. The latter can be based on the GRCh37 or GRCh38 genome assemblies and will be annotated in coding DNA, considering MANE select transcript or Locus Reference Genomic (LRG) t1 transcripts depending on user specifications. Moreover, each time it is executed a new log file is created. This provides an accurate per variant time-execution registry and collects all possible errors encountered during the process.

The current package works properly for the main HC genes ($n = 53$). It was also tested for all genes with a LRG entry ($n = 1325$) and it works for most of them (listed in Supplementary Table S4). It should also work for most genes with a MANE select transcript, although not all of them have been tested. However, variants located at positions where the reference allele differs between GRCh37 and GRCh38 cannot be computed by vaRHC. Nevertheless, these variants are usually polymorphisms that can be classified as benign based only on their high frequency in population datasets. The package relies on Mutalyzer v3 for variant nomenclature correction. Since not all transcript versions are supported by Mutalyzer, the tool searches for an available version and returns a warning to inform the user.

3.2 Performance assessment

To evaluate vaRHC's performance, 659 variants previously manually classified using specific guidelines were selected, of which 20 (3%) were not supported by vaRHC due to their nature (complex deletion–insertions, deletions/duplications >25 bp, inversions); thus, 639 variants were finally assessed: 29 *ATM*, 274 *CDH1*, 13 *CHEK2*, 13 *MLH1*, 22 *MSH2*, 33 *MSH6*, 10 *PMS2*, 128 *PTEN*, and 117 *TP53*.

3.2.1 Criteria assignation performance

The tool's performance was compared with the previous classification, showing that each fully automated criterion is correctly assigned in at least 97.7% of the variants. Figure 2 shows the performance broken down by gene and criteria (details in Supplementary Table S6). Supplementary Tables S7–S12 include all variants used for validation and a detailed explanation of discordant criteria (labelled according to Table 1).

Below, the validation results are depicted according to the different criteria groups.

3.2.1.1 Population data (criteria codes BA1, BS1, BS2, PS4, and PM2). The gnomAD v2.1.1 dataset was chosen to assess the variant allele frequency in control populations (Karczewski et al. 2020). Although v2 gathers fewer genome sequences than v3, it contains many more exome sequences (the main regions of interest for

Table 1. Labels assigned to the nine possible scenarios to quantitatively evaluate the performance of the programme.

Label	Description/scenario
Positive agreement	The criterion is assigned by both manual classification and vaRHC.
Negative agreement	The criterion is denied by both.
Previous version	The criterion used in the manual classification did not follow the most up to date guidelines for the gene.
Manual error	The manual choice clearly differs from the criterion statement in the guidelines.
Refined criterion	There is a discrepancy between the manual judgement and our software but in our view vaRHC's output is more accurate.
Partially automated	There is a discrepancy due to the inability to fully automate the criteria.
Not assessed	When BA1 is assigned by manual classification, other criteria are sometimes not evaluated by manual classification, thus the performance of the additional criteria for that variant cannot be compared.
Not automated	The criterion has not been automated.
Not applicable	According to the guidelines, the criterion should not be applied.

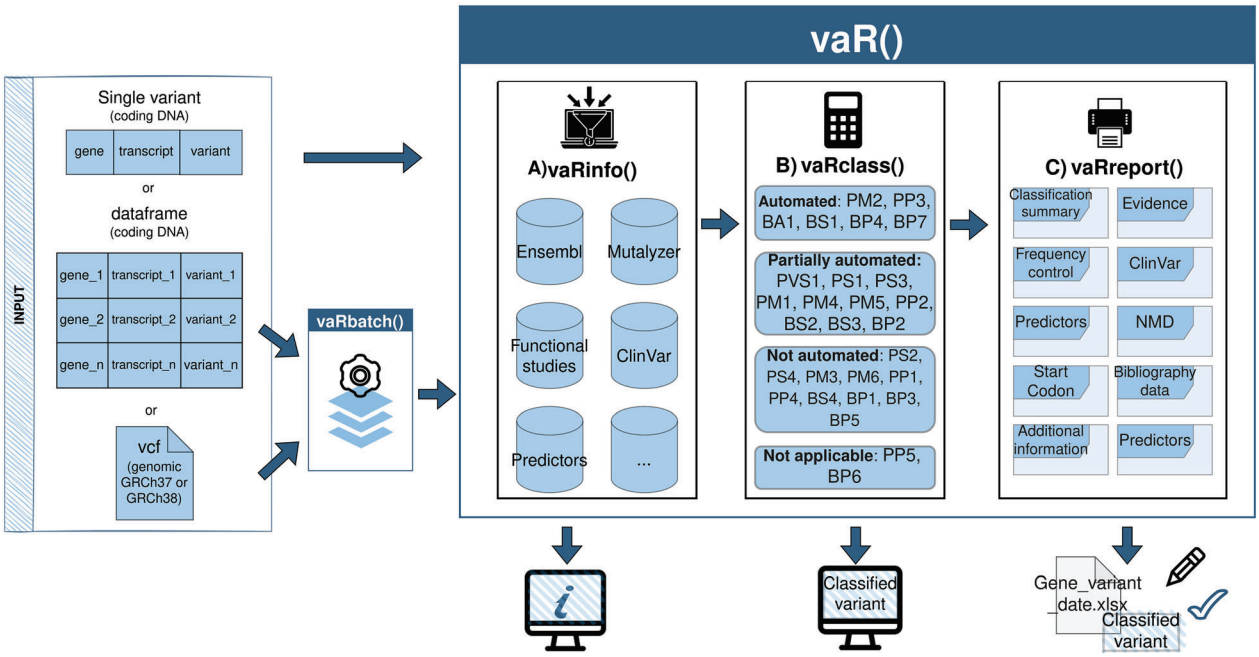


Figure 1 varRHC package: main functions and workflow. varR() contains three functions as follows: (**A**) vaRinfo: retrieves variant information from distinct databases; (**B**) vaRclass: combines information to assign or deny ACMG criteria returning a final classification of the variant; and (**C**) vaRreport: prints the results in a spreadsheet (.xlsx) file. varRbatch() allows to do the process sequentially.

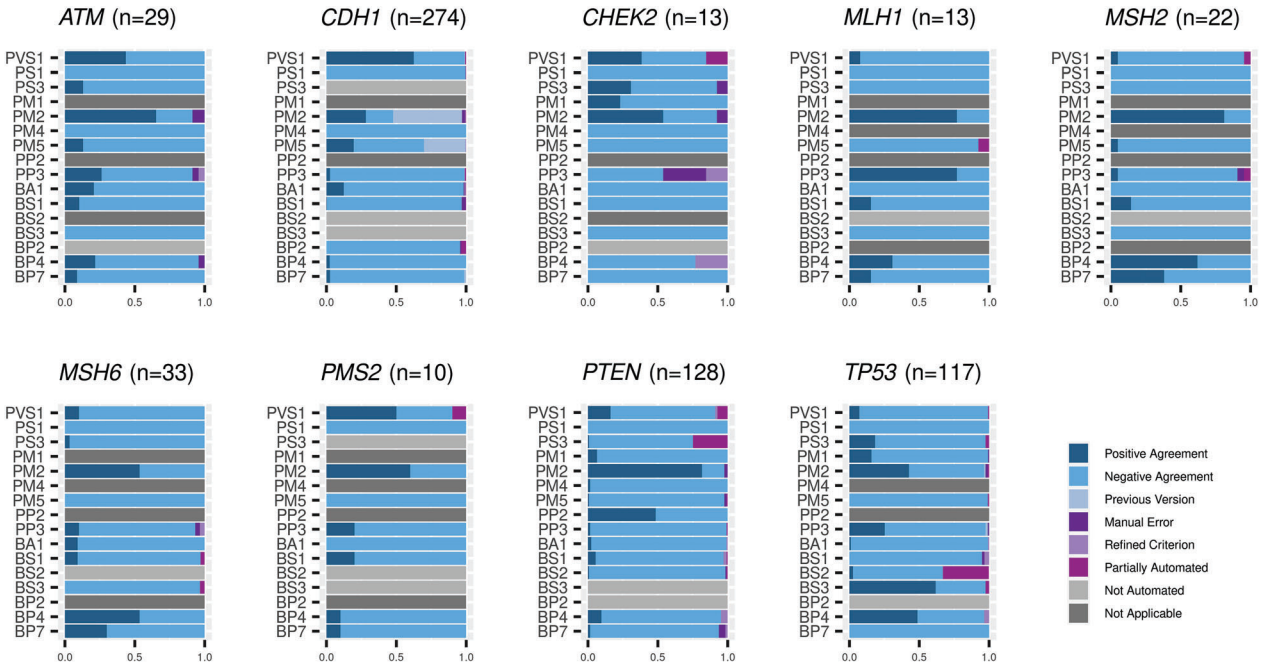


Figure 2 Tool performance assessed by gene and criteria. Stacked bar diagrams show the proportion of variants falling within each scenario (Table 1), labelled with a different colour (see legend in the figure). Each gene and criterion is evaluated separately (only fully or partially automated ones). Maximum number of variants analysed per gene is in parenthesis and corresponds to variants evaluated for BA1 and BS1. Since BA1 is a standalone criterion (incompatible with BS1), when BA1 was assigned by manual classification, the remaining criteria were not assessed in most variants and omitted from the comparison.

diagnostics), reaching a much higher allele number (±250 000) and increasing its statistical power. These databases were not aggregated as they overlap and it would distort allele frequencies (Gudmundsson et al. 2022). Specifically, the non-cancer dataset (minimum coverage of 20×) was selected (details in Supplementary Methods).

Furthermore, following ENIGMA recommendations (https://enigmaconsortium.org/library/general-documents/enigma-classification-criteria/; accessed November 2022), for BA1 and BS1, gnomAD founder populations, like Ashkenazi Jewish and Finnish, were omitted as bottleneck effects could mask natural negative selection of pathogenic variants. This exception was not considered in ClinGen; thus, disagreements were labelled as 'refined criteria'.

Positive and negative agreements were found in 98.9% of variants for BA1 and in 96.9% for BS1. The refined criterion accounted

for 0.63% of variants for BA1 and 1.10% for BS1. When manual ClinGen classifications used gnomAD v3, Exome Aggregation Consortium, or 1000 Genomes as population datasets to assign BA1 or BS1, these were conservatively labelled as 'partially automated' and accounted for 0.16% and 0.47% of variants, respectively.

Some gene-specific guidelines require the use of the lower confidence interval (CI) limit of allele frequencies from population datasets to compare them with maximum credible allele frequencies for a pathogenic variant to assign BA1 or BS1. Nevertheless, the corresponding manual classification dataset has not always applied it (see Supplementary Table S8). These manual errors explain the remaining 0.31% and 1.57% for BA1 and BS1, respectively.

Regarding the PM2 criterion, positive and negative agreements were found in 77.2% of variants. The main reason for disagreements was that some ClinGen pilot variants were classified using previous versions of gene-specific guidelines, lacking recent specifications, as vaRHC does. For example, for *CDH1*, PM2 was downgraded in version 3 to supporting strength, but many variants in the repository remain as moderate. These variants account for 20.0% of the total and were labelled as 'previous version'.

Only two variants (0.34%) were not assigned because they were only present in the gnomAD v3 non-cancer dataset (partially automated). Manual errors account for 2.2% of cases, many consisting of input variant nomenclature mistakes (e.g. using information from another variant or not finding the variant in the gnomAD database, although it was there). For HC susceptibility genes, gnomAD non-cancer is a more accurate population dataset, thus 0.2% of the variants were labelled as refined criterion. The BS2 criterion does not apply to *ATM* and *CHEK2* guidelines (6.2%) and is not automated for *CDH1* and MMR genes (52.9%). The lack of databases sharing this information impedes full automation of this criterion. Consequently, positive agreements only accounted for 0.69% of cases and were based on information from the FLOSSIES database and, for some genes, homozygote status from the gnomAD v2.1.1 non-cancer dataset. Negative agreements represented a 33.4%. ClinGen expert panels assign BS2 to 6.7% of variants thanks to in-house datasets (not shared), the literature, or ClinVar comments, which are not easily automatable. Only one variant (0.17%) was labelled as manual error (see Supplementary Table S11).

The PS4 criterion is not applicable to MMR genes or relies on information not stored in public databases, thus is not automated for other genes.

3.2.1.2 Computational and predictive criteria (PVS1, PS1, PM4, PM5, PP3, BP4, and BP7). For PVS1, the programme follows an algorithm based on Tayoun's decision tree (Abou Tayoun et al. 2018) for general classification and incorporates gene specificities where there are specific guidelines (Fig. 3). As shown in the figure, it integrates the splicing prediction, as a splicing alteration would affect the variant consequence.

Positive and negative agreements represented 96.7% of variants for PVS1 and 2.7% were labelled as 'partially automated'. The limitation in automation was mostly due to difficulty in determining splicing outcome when spliceAI predicts a splice site gain. Thus, vaRHC is conservative, assigning a supporting strength and returning a warning message suggesting an RNA test before assigning a higher strength. Likewise, when exon skipping is predicted for canonical splice variants located at the first or last exon, vaRHC returns a warning with no PVS1 strength assigned.

PTEN guidelines do not incorporate Tayoun's algorithm, only assigning a very strong strength. As a refined criterion, vaRHC considers modified Tayoun workflow for this gene (but incorporates the specific guideline consideration that truncating variants 5′ to c.1121 must be assigned as very strong); this changes the PVS1 strength assigned to two *PTEN* variants (0.34%).

Furthermore, variant c.1137 + 1delG in *CDH1* was assigned as very strong by ClinGen. However, per site-specific recommendations in the splicing table in version 3.1, it should be downgraded to strong. Thus, the variant was classified as 'previous version' (0.17%).

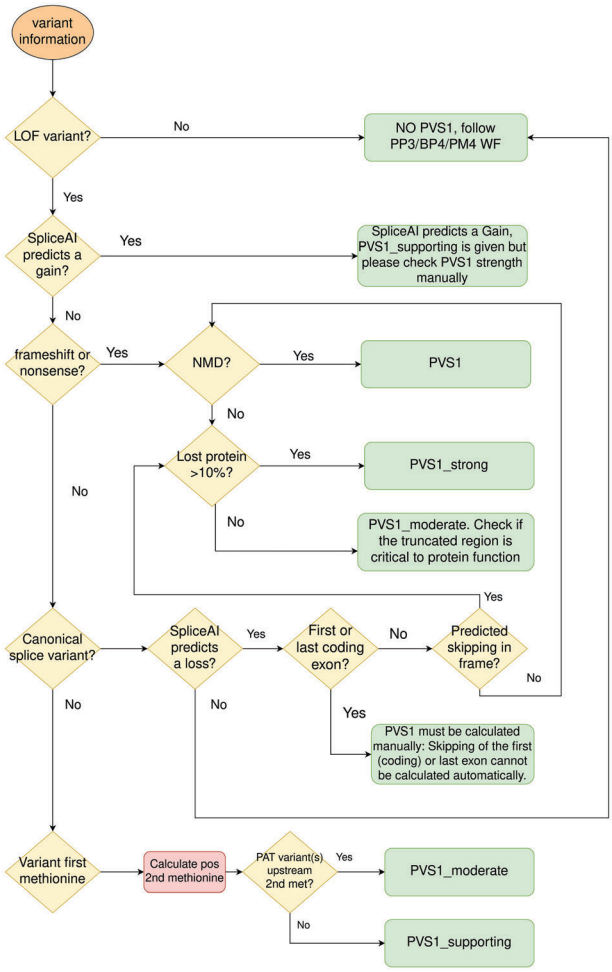


Figure 3 Flowchart showing the algorithm implemented in the vaRHC package to assign different PVS1 criterion strengths to LoF variants, based on ClinGen recommendations (Abou Tayoun et al. 2018).

Regarding prediction criteria (PP3 and BP4), only some guidelines specify which to use and their cut-offs. For the other genes, as general guidelines, REVEL metapredictor was adopted to predict protein impact, with optimized cut-offs proposed by Cubuk (2021). SpliceAI (Jaganathan et al. 2019) was selected as the main splice predictor. Since no cut-off was previously established, 518 RNA-tested variants were analysed to set PP3 and BP4 thresholds (Supplementary Methods and Supplementary Table S13). The final cut-offs for spliceAI were $\geq$0.5 for PP3 and $\leq$0.15 for BP4, giving pathogenicity odds ratio values of 25.3 and 0.037 for PP3 and BP4, respectively. Per Tavtigian's Bayesian framework, these could account for PP3_Strong and BP4_Strong, but a supporting strength was conservatively maintained in the tool.

Only MMR gene splice defect prediction combines SpliceAI with other algorithms from http://priors.hci.utah.edu/PRIORS, as specified in their guidelines. The retrieved predictors and cut-offs for each gene are in Supplementary Table S14.

From the above data, positive and negative agreement represented 96.7% of variants for PP3 and 97.5% in for BP4. For PP3, two *TP53* variants were affected by changes in predictor cut-offs between specific guideline versions and thus classified as 'previous version' (0.34%). Manual errors in PP3 were mostly due to applying PP3 when PVS1 or PM1 (by functional domain) was also assigned. According to CanVIG-UK, these criteria should not be combined as the same information is used to calculate them. Another manual error arises from mistakes in variant information queries. Manual errors represented 1.37% in PP3 and 0.18% in BP4. Discrepancies due to lack of detail in specific guidelines, which led us to choose

REVEL and SpliceAI, were considered 'refined criterion' and accounted for 0.86% in PP3 and 2.29% in BP4. Four variants (0.69%) were categorized as partially automated for PP3 (see Supplementary Tables S8, S10, and S11).

The BP7 criterion was classified as positive or negative agreement in 98.1% of variants. In *CDH1*, three variants were labelled as 'previous version', since manual classification did not assign BP7 to them, for being in a highly conserved nucleotide, a condition no longer applied. Manual errors correspond to six variants in *PTEN* (1.03%) where the nucleotide is predicted to be strongly conserved by Phastcons (value = 1) but manual classification assigned BP7 anyway. Moreover, Phylop was also assessed to determine nucleotide conservation (when required). Some scores obtained with the tool did not match those from manual classification. Three variants in *CDH1* were manually classified according to previous versions (0.51%) and two *PTEN* variants as refined criteria (0.34%) (see Supplementary Table S11).

Concerning PS1 and PM5, variants in the same codon as the test variant must be also classified according to gene-specific guidelines. Moreover, some guidelines, such as *TP53*, require that the variant be specifically classified by the ClinGen expert panel. To ensure this, vaRHC uses information from ClinVar variants classified by an expert panel (i.e. ClinGen for HC genes). Furthermore, these criteria were modified over time in *CDH1* guidelines: PS1 no longer applies and PM5 uses a new criterion applying to non-sense and frameshift variants predicted or proven to undergo NMD and to some canonical splicing variants.

It was difficult to validate vaRHC for PS1 and PM5 since the criteria dictates to compare test variants with expert panel classified variants. This is ClinGen dataset, the same that we are using to do the performance assessment. As expected, there were few pairs of variants at the same codon, both classified by the expert panel. Therefore, no variant was assigned PS1 by the software, and 116 of 121 variants assigned PM5 were due to the new *CDH1* criterion for truncating variants rather than the classic variant comparison criterion. To reliably validate the tool's performance, a list of 37 hypothetical variants located at the same codons as the variants used in the primary validation was created (see Supplementary Methods and Supplementary Table S15). The programme assigned PS1 or PM5 to 19 variants; PS1 and PM5 were denied in the remaining 18 variants as some criteria requirements were not met.

3.2.1.3 Functional data (PS3, PM1, PP2, BS3, and BP3). Functional data were generally refractory to automatic extraction. However, publication of some mid-to-high throughput clinically calibrated functional assays allowed their incorporation as an innovative feature of vaRHC. Most articles listed in gene-specific guidelines and some accomplishing the experimental conditions demanded in Brnich et al. (2019) to assign PS3 or BS3 were collected in the database (see Supplementary Methods, Section 1.1). Consequently, 5.1% of variants were assigned PS3 and 12.2% BS3. No literature information was used to assign PS3/BS3 to *CDH1* and *PMS2* genes (42.0%), or BS3 to *PTEN* (63.2%). Criteria were assigned only by manual classification in 5.8% of variants for PS3 and 0.69% for BS3.

Additionally, a search string combining different names for the variant was provided to be used in Internet search engines; this can help users find the most relevant articles for functional, allelic, and clinical criteria.

The PM1 criterion does not apply *to ATM*, *CDH1*, and MMR genes (56.8%). Only one variant (c.892G>T in *TP53*, 0.17%) was labelled as 'manual error' since manual classification assigned it PVS1 and PM1. Per CanVIG-UK, they should not be combined because PM1 can only be used for missense variants and small in-frame deletions and insertions. The remaining 43.0% correspond to positive and negative agreements.

Per specific guidelines, the PP2 criterion should only be used for *PTEN*, being correctly assigned in all cases.

3.2.1.4 Allelic data (PM3 and BP2). PM3 was not automated as allelic data in patients are seldom collected in databases. The BP2 criterion does not apply to MMR genes (12.7%) and can only be partially automated for *CDH1*. Specifically, only supporting strength can be assigned when the variant is in homozygosity in gnomAD since no public database of individuals without personal and/or family history of associated tumours was found. Manual classification assigned BP2_Strong to some variants as they were homozygous in a control cohort (gnomAD v2.1.1). However, according to Harrison et al. (2019), individuals in gnomAD should be cautiously considered as general population instead of healthy individuals for adult-onset conditions. Thus, 0.3% of variants were categorized as 'manual errors'.

3.2.1.5 Other databases (PP5 and BP6). According to ClinGen, reputable source not linked to the supporting evidence should not be used as criteria thus PP5 and BP6 should not be applied (Biesecker and Harrison 2018).

3.2.1.6 Segregation data (PP1 and BS4), de novo data (PS2 and PM6), and other data (PP4 and BP5). Due to the current lack of segregation information in databases, none of these criteria could be automated.

3.2.2 Automated classification performance

vaRHC classified each variant in 15–120 s. Automated classification concorded with manual classification in 63.4% of variants with five-tier classification, increasing to 74.0% with three-tier classification (Supplementary Table S5). Users would be expected to add non-automatable criteria to reach a final classification.

3.3 Benchmark

Cancer-SIGVAR (Li et al. 2021) is a free web tool (http://cancersigvar.bgi.com) based on ACMG/AMP rules and focused on interpreting HC variants. We analysed, criteria by criteria, the performance of Cancer-SIGVAR and vaRHC for *CDH1* and *PTEN* variants against the ClinGen repository (the same as in the validation dataset). However, in the vaRHC previous performance assessment we had identified manual errors and a proportion of variants classified without updated guidelines (see Fig. 2 and Supplementary Tables S7–S12). To address this, some ClinGen criteria assignments were modified, correcting variants labelled as 'manual errors' and 'previous version'. Conservatively, variants categorized as 'refined criterion' were not altered. The performance of both tools against this modified benchmark dataset was compared in (Fig. 4).

As it can be seen in the figure, vaRHC does not use PP5 and BP6 criteria, as dictated by ClinGen (Biesecker and Harrison 2018), it incorporates the updated v3 CDH1 guidelines (September 2021) and also data from functional assays. Due to these and other assets, vaRHC improves the performance of Cancer-SIGVAR.

Cohen's Kappa test comparing cancer-SIGVAR and vaRHC revealed significant differences for PM2 (kappa = 0.06, *P*-value = 2.98*E*−06), PM5 (kappa = 0.33, *P*-value = 1.34*E*−101), PP3 (kappa = 0.32, *P*-value = 9.91*E*−05), and BP4 (kappa = 0.47, *P*-value = 1.72*E*−07) criteria in *CDH1* and PM1 (kappa = 0.48, *P*-value = 0), PM4 (kappa = 0.42, *P*-value = 3.08*E*−02), PM5 (kappa = 0.30, *P*-value = 2.63*E*−02), BP4 (kappa = 0.42, *P*-value = 4.49*E*−10), and BP7 (kappa = 0.35, *P*-value = 6.58*E*−03) in *PTEN*. In contrast, comparing vaRHC with the modified benchmark dataset, all these criteria obtained Kappa values >0.7 (see Supplementary Table S16).

4 Discussion

Variant classification is a main challenge and bottleneck in the genetic testing process using NGS. ClinGen and expert panels work on adapting generic ACMG/AMP guidelines has led to several gene-specific recommendations, adding difficulty to the already long and complex manual classification process. Thus, the development of automated tools could assist this process, accelerating it and

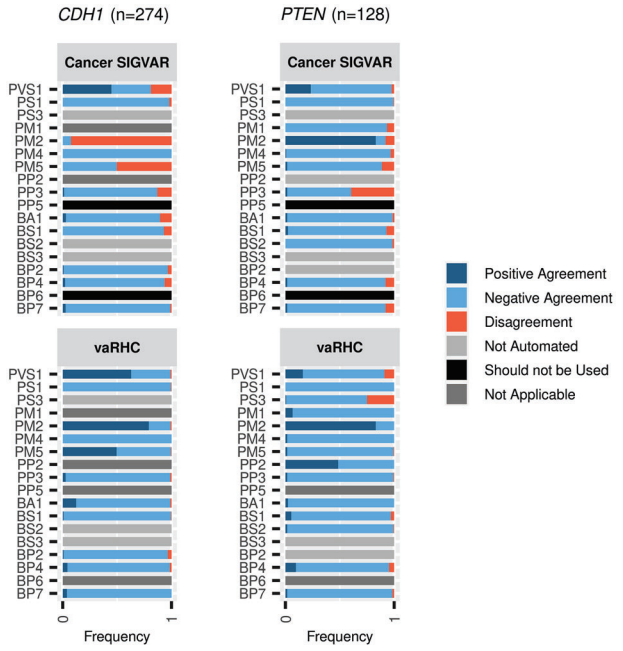


Figure 4 Performance of Cancer-SIGVAR and vaRHC in comparison with the modified benchmark dataset. Stacked bar diagrams show the proportion of variants falling within each scenario (see Sections 3, 3.3, and Table 1), labelled with a different colour (see legend in the figure), for CDH1 and PTEN genes. The maximum number of variants analysed per gene is in parenthesis. This corresponds to variants evaluated for BA1 and BS1. Since BA1 is a standalone criterion (incompatible with BS1), when BA1 was assigned by manual classification, the remaining criteria were not assessed in most variants and omitted from the comparison.

reducing manual error. However, most tools available use original ACMG/AMP general recommendations and few are adjusted to gene-specific ones.

Here, we present vaRHC, an R package developed to cover these needs. It automates as much as possible the variant classification process: it collects pieces of information from several databases and it combines them to assign or deny criteria according to the most up-to-date guidelines. To support classification in a more meaningful clinical class, vaRHC leverages Tavtigian's Bayesian approach and it also integrates most CanVIG-UK recommendations to finally assign or deny criteria, thus avoiding combining overlapping criteria, evading double counting. Moreover, vaRHC can be easily incorporated into bioinformatic pipelines, and adds a downloadable output as an editable, user-friendly spreadsheet (.xlsx) file. This function allows variant curators unfamiliar with the R environment to work with the data and add extra considerations, like the refinement of automated criteria or information derived from non-automatable criteria. Our validation demonstrates the robustness of the software developed in assigning automated criteria supporting the notion that automation tools are valuable in the variant classification process. However, they should not substitute the crucial role of the variant curator in supervising automated criteria, reviewing and incorporating data from the literature and in-house databases to assign semi-automatable and non-automatable criteria.

Despite ACMG/AMP and ClinGen's efforts to standardize guideline criteria to provide an objective framework, some ambiguous criteria remain that could cause discrepancies between laboratories. When identified we have labelled them as 'refined criteria'. Examples are the use of the non-cancer dataset and outbred subpopulations to calculate criteria related to population data or the choice of predictors and/or thresholds when the criteria do not establish them. Particularly, SpliceAI was selected as the unique splicing predictor, except for MMR genes, where Prior (UTAH) was also used according to gene-specific guidelines. Some guidelines suggest variant scoring using a consensus of two or three predictors. However, it was recently demonstrated that this provides little benefit (Wai et al. 2020). Moreover, a benchmark recommended the use of SpliceAI because it has the highest area under the curve (Riepe et al. 2021). To establish cut-offs for benignity and pathogenicity, a specific assessment was designed with 518 RNA-tested variants. Of note, vaRHC uses all proposed thresholds as defaults but is easily modifiable with a txt file.

Previous tools have almost exclusively focused on the final classification of the variant instead of analysing criterion by criterion. As automated tools lack some information derived from non-automatable criteria, many variants are classified as 'unknown significance'. To palliate this, some programmes are less strict in assigning some criteria (e.g. permissive population frequency thresholds for BA1, BS1, and PM2 by software like Varsome) or use criteria that currently do not apply (e.g. PP5 and BP6 by SIGVAR, InterVar, Varsome, Franklin, Pathoman, and CharGer) to leverage ClinVar classification information and artificially approach its result. We recommend to use these tools in a research context to prioritize variants rather than directly employ them in diagnostics. One of vaRHC's strengths is that it is strict in assigning criteria. Moreover, for each denied or assigned criteria, it always returns an explanation and, when appropriate, suggests extra considerations that should be noted for manual curation of criteria. This makes vaRHC suitable for use in molecular diagnostics units. In fact, the package is currently used by the Catalan Institute of Oncology (ICO) Molecular Diagnostics Service.

Although vaRHC was developed to answer to HC genetic testing needs, its approach to classify variants in HC genes without gene-specific guidelines is generalizable to most genes where loss-of-function variants cause heritable diseases. Furthermore, the customizable nature of most parameters allows users to adapt the tool to their needs. The use of the gnomAD v2.1. non-cancer dataset for variant population frequency assessment is compatible with other diseases, this dataset has only subtracted a few samples belonging to cancer patients, that could have invalidated conclusions for variants involved in these conditions and would not have increased substantially the power of the datasets.

The performance of our tool was compared with the ClinGen datasets for each HC gene with specific ClinGen guidelines. These datasets have a limited size compared with ClinVar, but they are manually curated by experts and inform the assignation of each criterion, which cannot be matched by the ClinVar dataset. Although a more homogeneous dataset would be desirable, each ClinGen Variant Curation Expert Panel chose the number of variants for its curated set as appropriate to exemplify the use of its criteria. We have expanded this collection with manually curated *CHEK2* variants from an article that 'proposes' *CHEK2*-specific rules and with MMR gene variants from the in-house ICO diagnostics laboratory DB. Our tool and Cancer-SIGVAR were compared with the ClinGen dataset only for *CDH1* and *PTEN*, since those are the genes where Cancer-SIGVAR follows gene-specific guidelines. Cancer-SIGVAR incorporates specific guidelines for some genes but does not cover recent ClinGen guidelines for relevant cancer genes. For impartiality, variants labelled as 'manual errors' and 'previous version' in ClinGen were corrected. Besides the aforementioned differences in using non-applicable criteria, the Kappa Test analysis revealed significant differences in other criteria assigned, favouring vaRHC.

A limitation of vaRHC is that, currently, it does not work for all variant types or lengths. Its execution time is around 15–30 s per variant, but can reach 2 min for insertions and deletions, where SpliceAI cannot be precomputed. Furthermore, its connection to ClinVar and other databases relies on the good performance and connectivity of their websites. Moreover, perpetual maintenance is planned and needed because any change in the html structure or content of queried websites via web scrapping can lead to different types of errors. Also, the database contains information on some published functional assays, but this is not due to a self-renewing ability to mine the literature, but to a manual effort. Nevertheless, the current vaRHC release plan includes the addition of recently

published functional assays. Also, updates will include multiple features such as using gnomAD v3 information and downloading the report in other file extensions.

In summary, the performance assessment and benchmark carried out in the present work corroborates the robustness and excellent performance of vaRHC to assist variant classification. To our knowledge, our package outperforms tools available for several reasons: (i) it is the first freely available R package that semi-automates the process; (ii) it uses Tavtigian's Bayesian metastructure nuanced by CanVIG-UK criterion combination rules, and (iii) it includes gene-specific guidelines for several commonly studied cancer genes, like *ATM*, *CDH1*, *CHEK2*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *PTEN*, and *TP53*. Altogether, we expect that vaRHC will facilitate the task of variant curators in clinical settings by reducing time for variant classification, limiting manual errors, and allowing the personalization of some parameters according to clinical and laboratory data.

Acknowledgements

We thank the CERCA Program/Generalitat de Catalunya for institutional support. We also thank all the members of the ICO HC Program.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the Carlos III National Health Institute and Ministerio de Ciencia e Innovación, funded by FEDER funds—a way to build Europe—[PI19/00553] and CIBERONC [CB16/12/00234]; the Government of Catalonia [Pla estratègic de recerca i innovació en salut (PERIS_MedPerCan and URDCat projects), 2017SGR1282, 2017SGR496, and 2021SGR01112]; and 'Acció instrumental de formació de científics i tecnòlegs' [SLT017/20/000129] of the Departament de Salut de la Generalitat de Catalunya.

References

Abou Tayoun AN, Pesaran T, DiStefano MT *et al.*; ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* 2018;**39**:1517–24.

Amendola LM, Jarvik GP, Leo MC *et al.* Performance of ACMG–AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am J Hum Genet* 2016;**98**: 1067–76.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;**57**: 289–300.

Biesecker LG, Harrison SM; ClinGen Sequence Variant Interpretation Working Group. The ACMG/AMP reputable source criteria for the interpretation ofsequence variants. *Genet Med* 2018;**20**:1687–8.

Brnich SE, Abou Tayoun AN, Couch FJ *et al.*; Clinical Genome Resource Sequence Variant Interpretation Working Group. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med* 2019;**12**: 3–12.

Cubuk C, Garrett A, Choi S *et al.* Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genet Med* 2021;**23**:2096–104.

Feliubadaló L, López-Fernández A, Pineda M *et al.*; Catalan Hereditary Cancer Group. Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. *Int J Cancer* 2019;**145**:2682–91.

Fortuno C, Lee K, Olivier M *et al.*; ClinGen TP53 Variant Curation Expert Panel. Specifications of the ACMG/AMP variant interpretation guidelines for germline *TP53* variants. *Hum Mutat* 2021;**42**: 223–36.

Garrett A, Durkie M, Callaway A *et al.*; CanVIG-UK. Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: canVIG-UK consensus recommendations. *J Med Genet* 2021;**58**: 297–304.

Gelb BD, Cavé H, Dillon MW *et al.*; ClinGen RASopathy Working Group. ClinGen's RASopathy expert panel consensus methods for variant interpretation. *Genet Med* 2018;**20**:1334–45.

Gudmundsson S, Singer-Berk M, Watts NA *et al.* Variant interpretation using population databases: Lessons from gnomad. *Hum Mutat* 2022;**43**: 1012–30.

Harrison SM, Biesecker LG, Rehm HL. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Curr Protoc Hum Genet* 2019;**103**:e93.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF *et al.* Predicting splicing from primary sequence with deep learning. *Cell* 2019;**176**: 535–48.e24.

Joseph V, Ravichandran V, Offit K *et al.* Pathogenicity of mutation analyzer (PathoMAN): a fast automation of germline genomic variant curation in clinical sequencing. *J Clin Oncol* 2017;**35**:1529.

Karczewski KJ, Francioli LC, Tiao G *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**: 434–43.

Kopanos C, Tsiolkas V, Kouris A *et al.* VarSome: the human genomic variant search engine. *Bioinformatics* 2019;**35**:1978–80.

Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**:D980–5.

Landrum MJ, Lee JM, Benson M *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**: D1062–7.

Lee K, Krempely K, Roberts ME *et al.* Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline *CDH1* sequence variants. *Hum Mutat* 2018;**39**:1553–68.

Li H, Liu S, Wang S *et al.* Cancer SIGVAR: a semiautomated interpretation tool for germline variants of hereditary cancer-related genes. *Hum Mutat* 2021;**42**:359–72.

Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *Am J Hum Genet* 2017;**100**:267–80.

Lumish HS, Steinfeld H, Koval C *et al.* Impact of panel gene testing for hereditary breast and ovarian cancer on patients. *J Genet Couns* 2017;**26**: 1116–29.

Luo X, Feurstein S, Mohan S *et al.* ClinGen myeloid malignancy variant curation expert panel recommendations for germline RUNX1 variants. *Blood Adv* 2019;**3**:2962–79.

Mester JL, Ghosh R, Pesaran T *et al.* Gene-specific criteria for *PTEN* variant curation: recommendations from the ClinGen *PTEN* expert panel. *Hum Mutat* 2018;**39**:1581–92.

Nagy R, Sweet K, Eng C *et al.* Highly penetrant hereditary cancer syndromes. *Oncogene* 2004;**23**:6445–70.

Nakken S, Saveliev V, Hofmann O *et al.* Cancer predisposition sequencing reporter (CPSR): a flexible variant report engine for germline screening in cancer. *Int J Cancer* 2021;**149**:1955–60.

Nicora G, Limongelli I, Gambelli P *et al.* CardioVAI: an automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. *Hum Mutat* 2018;**39**:1835–46.

Nykamp K, Anderson M, Powers M *et al.*; Invitae Clinical Genomics Group. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med* 2017;**19**:1105–17.

Oza AM, DiStefano MT, Hemphill SE *et al.*; ClinGen Hearing Loss Clinical Domain Working Group. Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. *Hum Mutat* 2018;**39**: 1593–613.

Patel RY, Shah N, Jackson AR *et al.*; on behalf of the ClinGen Resource. ClinGen pathogenicity calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Med* 2017;**9**:1–9.

Ravichandran V, Shameer Z, Kemel Y *et al.*; Toward automation of germl variant curation in clinical cancer genetics. *Genet Med* 2019;**21**: 2116–25.

Richards S, Aziz N, Bale S *et al.*; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet Med* 2015;**17**:405–24.

Riepe TV, Khan M, Roosing S *et al.* Benchmarking deep learning splice prediction tools using functional splice assays. *Hum Mutat* 2021;**42**:799–810.

Scott AD, Huang K-L, Weerasinghe A *et al.* CharGer: clinical characterization of germline variants. *Bioinformatics* 2019;**35**:865–7.

Siegel RL, Miller KD, Fuchs HE *et al.* Cancer statistics, 2022. *CA Cancer J Clin* 2022;**72**:7–33.

Tavtigian SV, Greenblatt MS, Harrison SM *et al.*; ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* 2018;**20**:1054–60.

Tavtigian SV, Harrison SM, Boucher KM *et al.* Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat* 2020;**41**:1734–7.

Tung N, Lin NU, Kidd J *et al.* Frequency of germline mutations in 25 cancer susceptibility genes in a sequential series of patients With breast cancer. *J Clin Oncol* 2016;**34**:1460–8.

Vargas-Parra G, Del Valle J, Rofes P *et al.* Comprehensive analysis and ACMG-based classification of *CHEK2* variants in hereditary cancer patients. *Hum Mutat* 2020;**41**:2128–42.

Wai HA, Lord J, Lyon M *et al.* Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med* 2020; **22**:1005–14.

Whiffin N, Walsh R, Govind R *et al.* CardioClassifier: disease- AND gene-specific computational decision support for clinical genome interpretation. *Genet Med* 2018;**20**:1246–54.

Supplementary Methods

1. Databases

1.1 Local queries

The database created contains Locus Reference Genomic (LRG) stable reference sequences for clinical reporting (https://www.lrg-sequence.org/; accessed March 2022) from a set of cancer genes listed in Supplementary Table 3. It also stores their transcript and genomic information with Ensembl transcript ID, NCBI RefSeqs, and the corresponding consensus coding sequence (CCDS). The user can execute vaRHC for variants in other cancer and non-cancer genes and different transcripts, but those stored in the database have been tested for bugs.

The database also contains information on variant frequency in the large population datasets gnomAD (v2.1.1) non-cancer and non-neuro (Karczewski et al., 2020), large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants (Parsons et al., 2019), and functional studies for ATM (Scott et al., 2002; Barone et al., 2009; Mitui et al., 2009), BRCA1 (Lyra et al., 2020), mismatch repair genes (MMR) (Drost et al., 2019, 2020; Jia et al., 2021; Rayner et al., 2022), TP53 (Kato et al., 2003; Giacomelli et al., 2018; Kotler, Segal and Oren, 2018), PTEN (Trojan et al., 2001; Agrawal, Pilarski, and Eng, 2005; Chen et al., 2017) and CHEK2 (Wu, Webster and Chen, 2001; Wu et al., 2006; Schwarz, Lovly, and Piwnica-Worms, 2003; Sodha et al., 2006).

Moreover, it stores in-silico predictor databases such as the HCI Database of Prior Probabilities of Pathogenicity for Single Nucleotide Substitutions protein level information (http://priors.hci.utah.edu/PRIORS/; accessed November 2022) and dbnsfp, a database developed for functional prediction and annotation of all potential non-synonymous, single-nucleotide substitutions in the human genome. From dbnsdfp, varHC obtains predictions of REVEL, VEST4, PROVEAN, BayesDel (Liu et al., 2020), Align_gvgd Zebrafish, and Provean no AF for TP53.

Finally, it also includes valuable site-specific information extracted from gene-specific guidelines. For instance, the splicing effects table for canonical variants in CDH1, sequence of the first six bases of the intron to assign or deny PVS1 to G>non-G for last base of exon variants for MMR and ATM.

1.2 Real-time queries

The following databases update information periodically: ClinVar aggregates information about human variations and their relationship to phenotypes (Landrum et al., 2014, 2018) and InSiGHT collects information about DNA variants from genes related to gastrointestinal cancer (https://www.insight-group.org/variants/databases/)(InSiGHT variants databases - InSiGHT, 2022). These databases are queried via web scrapping so as to have their latest version.

vaRHC queries some software that do not have a pre-computed database for all variants, such as Mutalyzer (Lefter et al., 2021) or Variant Effect Predictor (VEP) (McLaren et al., 2016) and web interfaces that do not provide the option to download data, such as the Fabulous Ladies Over Seventy (FLOSSIES) database (https://whi.color.com/). Online tools without pre-computed databases are queried via REST API. For example, base-wise conservation in the vertebrate Multiz Alignment & conservation (100 species), Phastcons, and Uniprot values/data are obtained from UCSC API (https://genome.ucsc.edu/goldenPath/help/api.html).

Although SpliceAI provides a pre-computed database, it only considers the author's default parameters (Jaganathan et al., 2019). However, increasing the "Max distance" window allows assessing the effect of the variant on the score of more distant positions. Calculating SpliceAI Δscores for a variant is time-consuming using REST API (https://spliceailookup-api.broadinstitute.org/). To

reduce the number of queries and thus time spent, the masked Δscores using a window of 1000 nucleotides have been previously calculated and stored in our database for all possible exonic substitutions in ATM, CHEK2, MLH1, MSH2, MSH6, PMS2, PTEN, and TP53.

2. GnomAD dataset pre-processing

The GnomAD non-cancer dataset was selected as a general population control with a minimum coverage of 20x. Sequenced exomes and genomes in v2.1.1 belong to different individuals. Thanks to this, the total allele number and allele count is considered jointly (if both datasets reach the minimum coverage). When a variant is not found in exomes or genomes, GnomAD does not provide its allele number (total number of called high genotypes at its position). The omission of this information can lead to a false allele frequency calculation for the whole dataset. To mitigate this issue, vaRHC adopts the following approach: it uses the allele number of the nearest upstream and downstream variant in a window of up to 50 bp and adopts its allele number (if it finds both, it calculates the mean). When no variant is found, the program does not calculate criteria related to population frequencies.

3. Splicing predictor assessment and cut-off selection

3.1. Variant selection

A dataset of 518 RNA-tested variants with unequivocal splicing results was used to establish cut-offs for benignity (BP4) and pathogenicity (PP3). Variants were obtained from the literature (Menéndez et al., 2012; Thomassen et al., 2012; Colombo et al., 2014; Whiley et al., 2014; Quiles et al., 2016; Rofes et al., 2020) and belonged to genes (number of variants in parenthesis): APC-NM_001354896.1 (2), ATM-NM_000051.3 (9), BRCA1-NM_007294.3 (167), BRCA2-NM_000059.3 (161), BRIP1-NM_032043.2 (3), CDH1-NM_004360.3 (12), CHEK2-NM_007194.3 (3), DKC1-NM_001363.3 (1), FBN1-NM_000138.4 (9), FGFR1-NM_001174067.1 (1), FLNB-NM_001457.3 (1), KANSL1-NM_015443.3 (1), MED13L-NM_015335.4 (1), MEF2C-NM_002397.4 (1), MLH1-NM_0002492 (74), MSH2-NM_000251.2 (25), MSH6-NM_000179.2 (9), MUTYH-NM_001128425.1 (1), MYBPC3-NM_000256.3 (1), NGLY1-NM_018297.3 (1), PAFAH1B1-NM_000430.3 (1), PALB2-NM_024675.3 (8), PIGB-NM_004855.4 (1), PMS2-NM_000535.5 (2), PNKP-NM_007254.3 (1), POLD1-NM_001308632.1 (1), PTCH1-NM_000264.4 (1), PTEN-NM_000314.4 (6), RAD51D-NM_002878.3 (1), SF3B4-NM_005850.4 (1), SKI-NM_003036.3 (1), SMAD4 - NM_005359.5 (1), STK11-NM_000455.4 (2), TP53-NM_000546.5 (7), TSC-NM_000368.4 (1), and TSC2-NM_000548.3 (1).

Variants affecting splicing (n=317) were divided into four groups depending on their consequence: 21 caused an acceptor gain (AG), 25 a donor gain (DG), 94 an acceptor loss (AL), and 177 a donor loss (DL). Variants not affecting splicing (n=202) where divided into two groups: 89 next to a donor site and 113 next to an acceptor site.

3.2 Cut-off selection for SpliceAI

Firstly, the SpliceAI neural network algorithm was run for the 518 variants using the Broad Institute API (https://spliceailookup-api.broadinstitute.org/; accessed March 2022). It delivers four Δscores, whose interpretation is explained in the SpliceAI flagship article (Jaganathan et al., 2019). A variant is considered to alter splicing when any of the Δscores exceeds the set cut-off; thus, the four Δscores were plotted simultaneously for each variant (Supplementary Figure 2). Secondly, the number of variants reaching PP3 or BP4 was calculated for each category using different cut-offs (Supplementary table 12). Since variants not affecting splicing are more frequently detected than

variants affecting splicing, our dataset was biased and the results could be biased too. To avoid this, we simulated variants not affecting splicing to be nine times the 317 variants affecting splicing. This proportion would match a prior proportion of 0.1, as the one suggested in Tavtigian's proposal for ACMG variant classification (Tavtigian et al., 2018). By doing so, the final number of variants not affecting splicing was 2835, thus more weight is assigned to benign variants (Supplementary table 12).

The odds in favour of pathogenicity for every cut-off were calculated as published in Easton (2007). The cut-offs were chosen to ensure high odd ratios of pathogenicity and benignity while limiting the grey area of variants without predictive evidence.

4. Evaluation of PS1 and PM5 criteria

4.1 Variant generation

PS1 and PM5 criteria compare the novel missense change with other previously classified variants at the same amino acid residue. The previous variant must be classified using gene-specific guidelines, sometimes by the corresponding expert panel. To ensure this, the program only uses as previous variants those classified in ClinVar by Expert Panel (which for HC genes usually is ClinGen). Since our repository is downloaded from ClinGen, the comparing variants are limited. To validate the performance of PS1 and PM5, missense variants classified as pathogenic or likely pathogenic in ClinVar by Expert Panel were selected for the ATM, CDH1, PTEN, TP53, and MMR genes (https://www.ncbi.nlm.nih.gov/clinvar/; downloaded on May 2022).

The PS1 criterion is assigned when the novel missense variant generates the same amino acid change as a previously established pathogenic variant. From the list of missense variants, 11 new variants could be generated harbouring the same amino acid change as the original ClinGen or ClinVar Variant (Supplementary Table 14).

However, PM5 is assigned when the novel missense variant at the same codon generates a different amino acid change as a previously established pathogenic variant. There are many hypothetical variant candidates to accomplish PM5, thus, for the purpose of validation, 26 variants were chosen randomly.

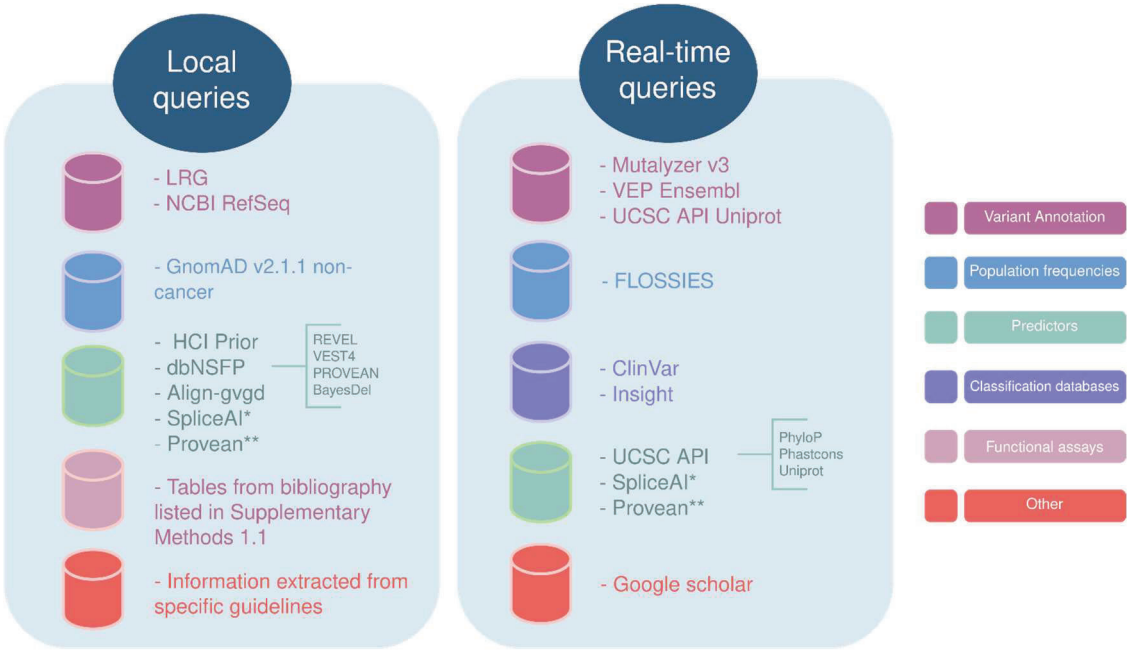
4.2 Validation

Regarding PS1 results, 8 out of 11 variants were assigned PS1 or PS1_moderate. Three variants were denied PS1: 1) a CDH1 variant, since PS1 do not apply to CDH1; 2) a TP53 variant because the variant classified by Expert Panel was likely pathogenic and according to TP53 guidelines only pathogenic variants should be considered; 3) a PMS2 variant because SpliceAI prediction suggested a splicing alteration.

As for PM5, 11 of 26 variants were assigned moderate or strong strength. The cases where PM5 was denied are detailed in Supplementary table 14 and depend on gene-specific guidelines, but most were due to the criterion not being applicable to the gene, the test variant having a higher BLOSUM62 value or a higher Grantham distance score than the previous variant, or the variant having a prior Utah value score below 0.68 for MMR genes

Supplementary Figure 1

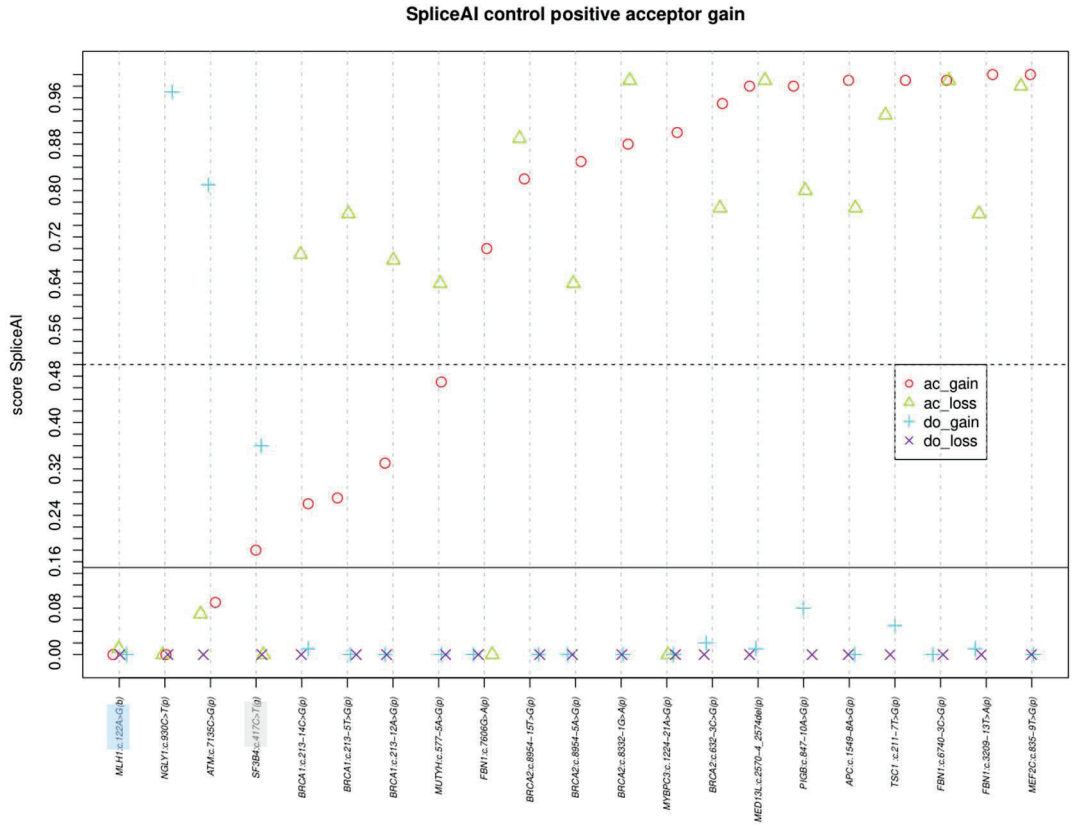
List of sources queried by vaRHC. They are separeted by the type of query (local vs real-time) and the type of information ectracted from the database.



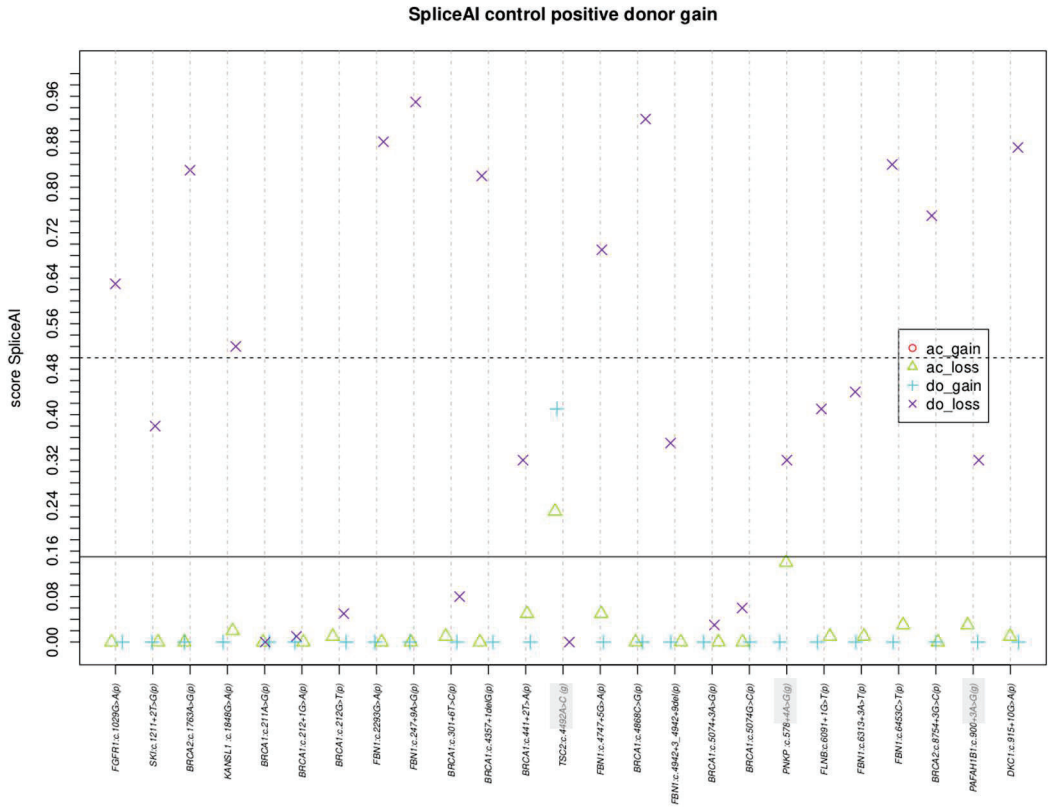
Supplementary Figure 2

SpliceAI 4 Δscores plotted for each variant separated by RNA consequence. A) Variants causing an acceptor gain; B) Variants causing a donor gain; C) Variants causing an acceptor loss; D) Variants causing a donor loss; E) Variants not altering splicing, next to an acceptor site; F) Variants not altering splicing, next to a donor site. Variants not altering splicing, next to a donor site. Variants marked with colours were erroneously categorized by SpliceAI: red= altering splicing, grey= not classified, blue= not altering splicing, all according to the proposed SpliceAI thresholds.

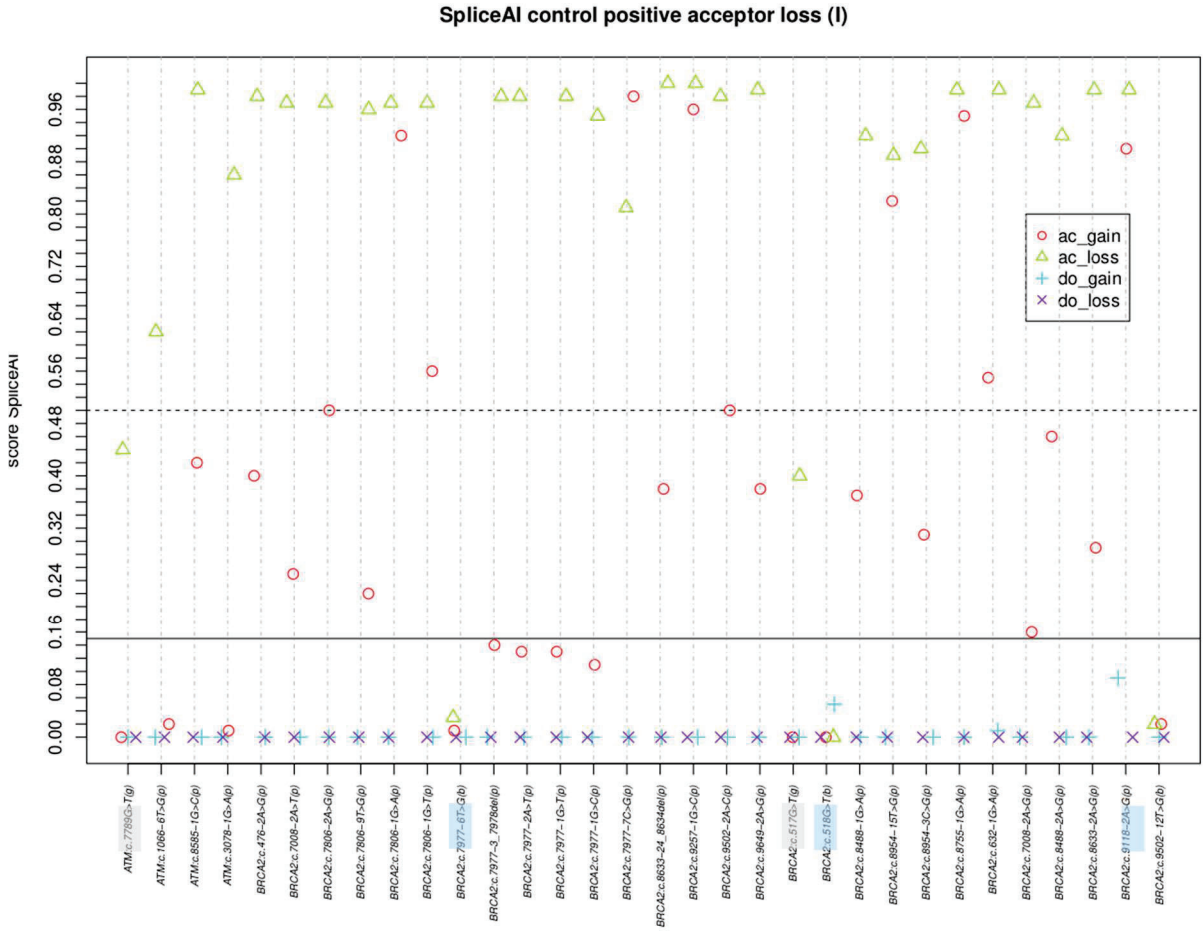
A)

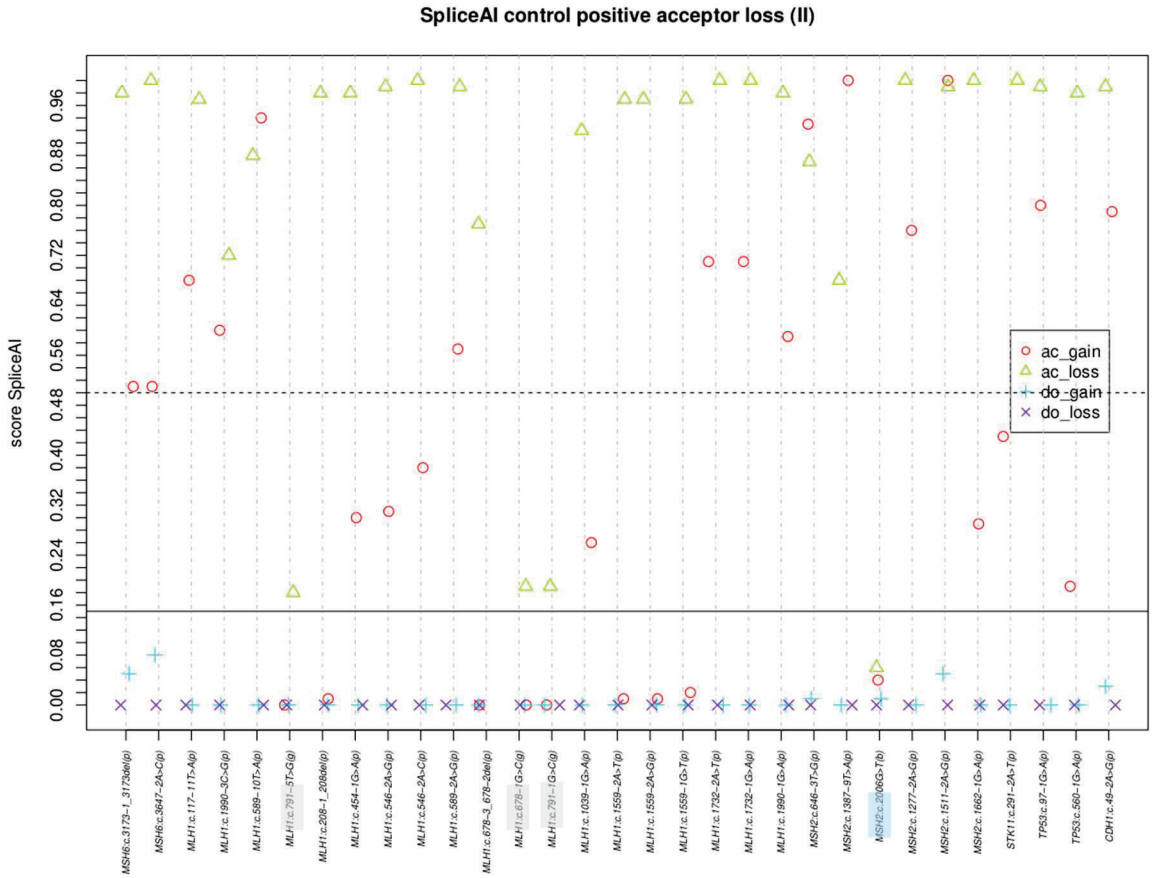

SpliceAI control positive acceptor gain
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss
MLH1:c.122A>G(b)
NGLY1:c.830C>T(p)
ATM:c.7136C>G(p)
SF3B4:c.417C>T(p)
BRCA1:c.213-14C>G(p)
BRCA1:c.213-57>G(p)
BRCA1:c.213-12A>G(p)
MUTYH:c.577-5A>G(p)
FBN1:c.7696G>A(p)
BRCA2:c.8954-157>G(p)
BRCA2:c.8954-5A>G(p)
BRCA2:c.8332-1G>A(p)
MYBPC3:c.1224-21A>G(p)
BRCA2:c.632-3C>G(p)
MED13L:c.2570-4_2574del(p)
PIGB:c.847-10A>G(p)
APC:c.1549-8A>G(p)
TSC1:c.211-7T>G(p)
FBN1:c.6740-3C>G(p)
FBN1:c.3209-13T>A(p)
MEF2C:c.835-97>G(p)

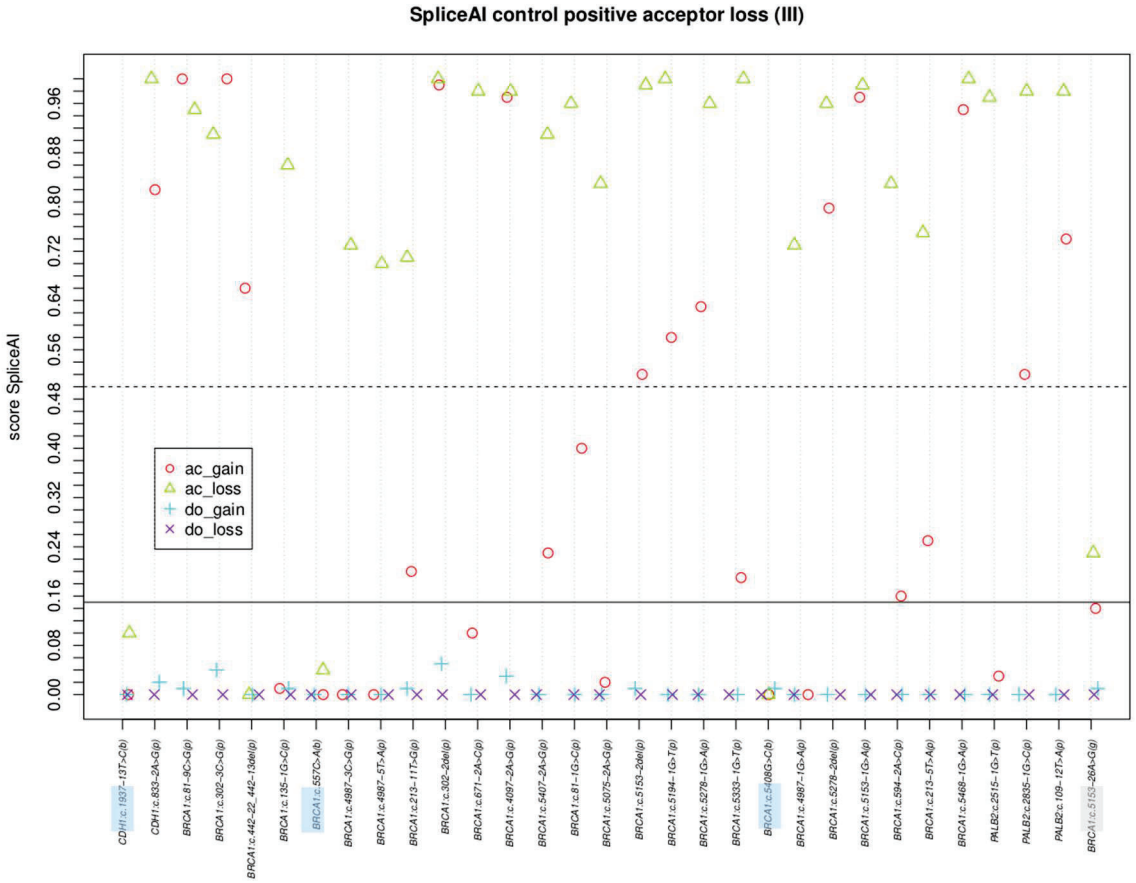
B)


SpliceAI control positive donor gain
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

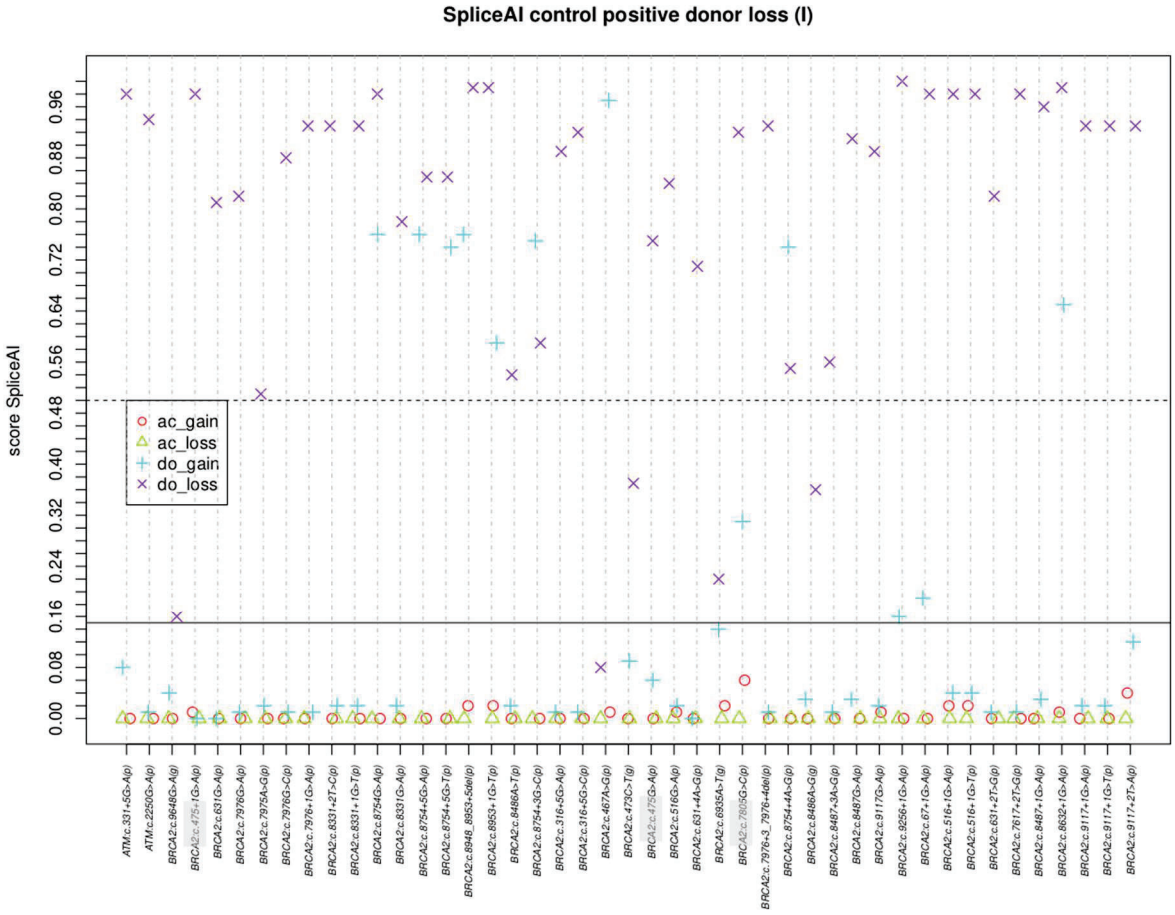
C)

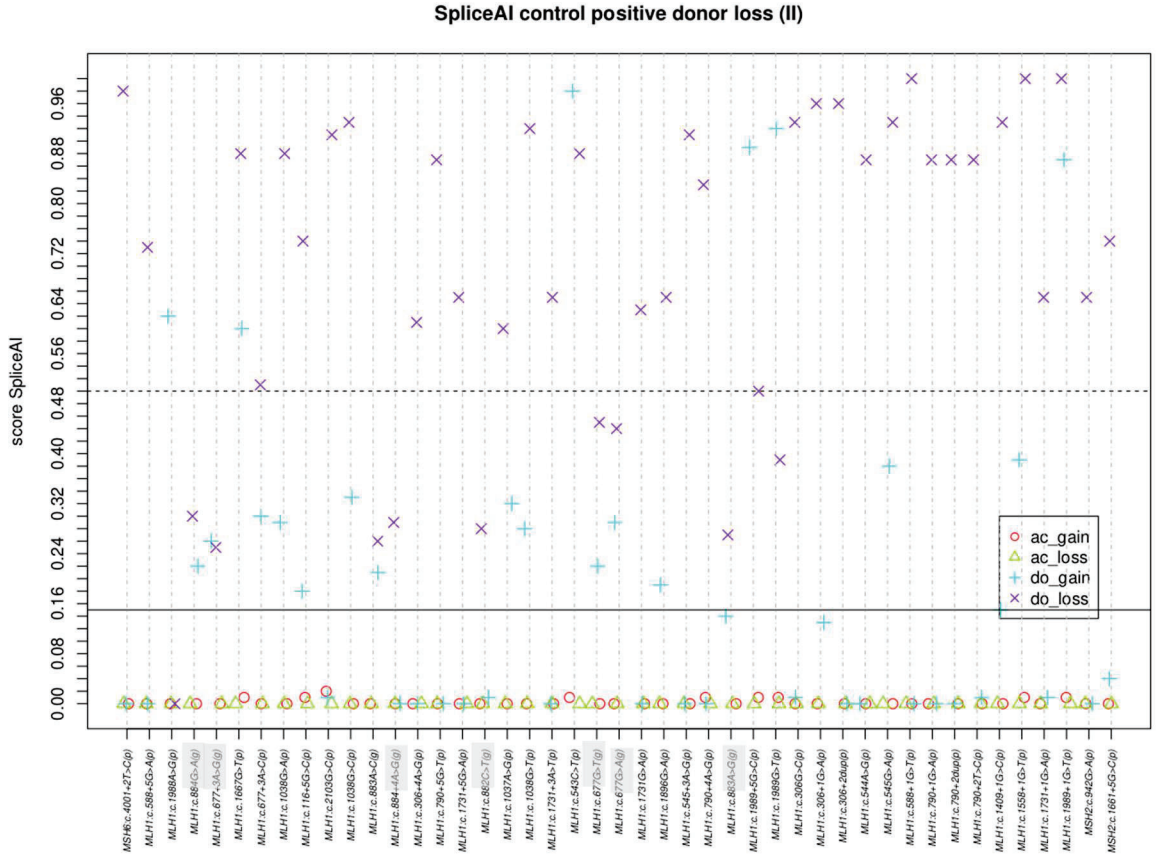

SpliceAI control positive acceptor loss (I)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

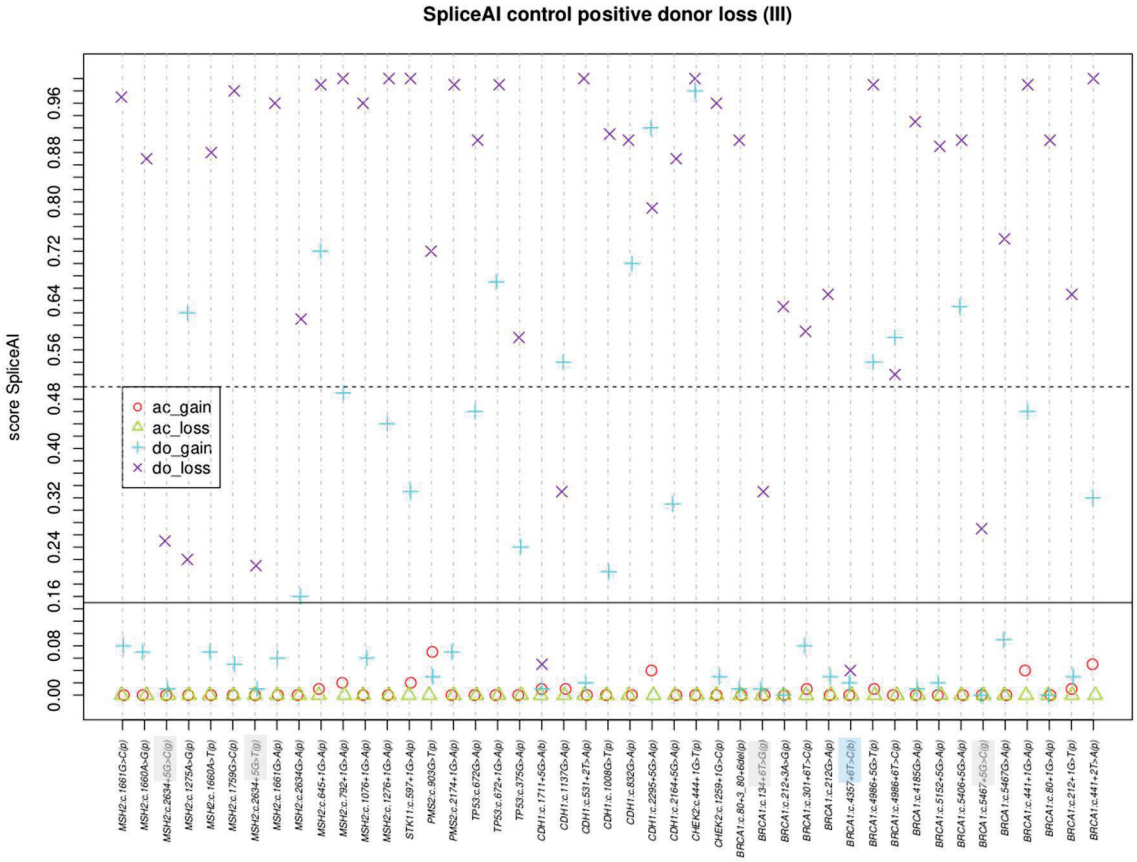

SpliceAI control positive acceptor loss (II)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

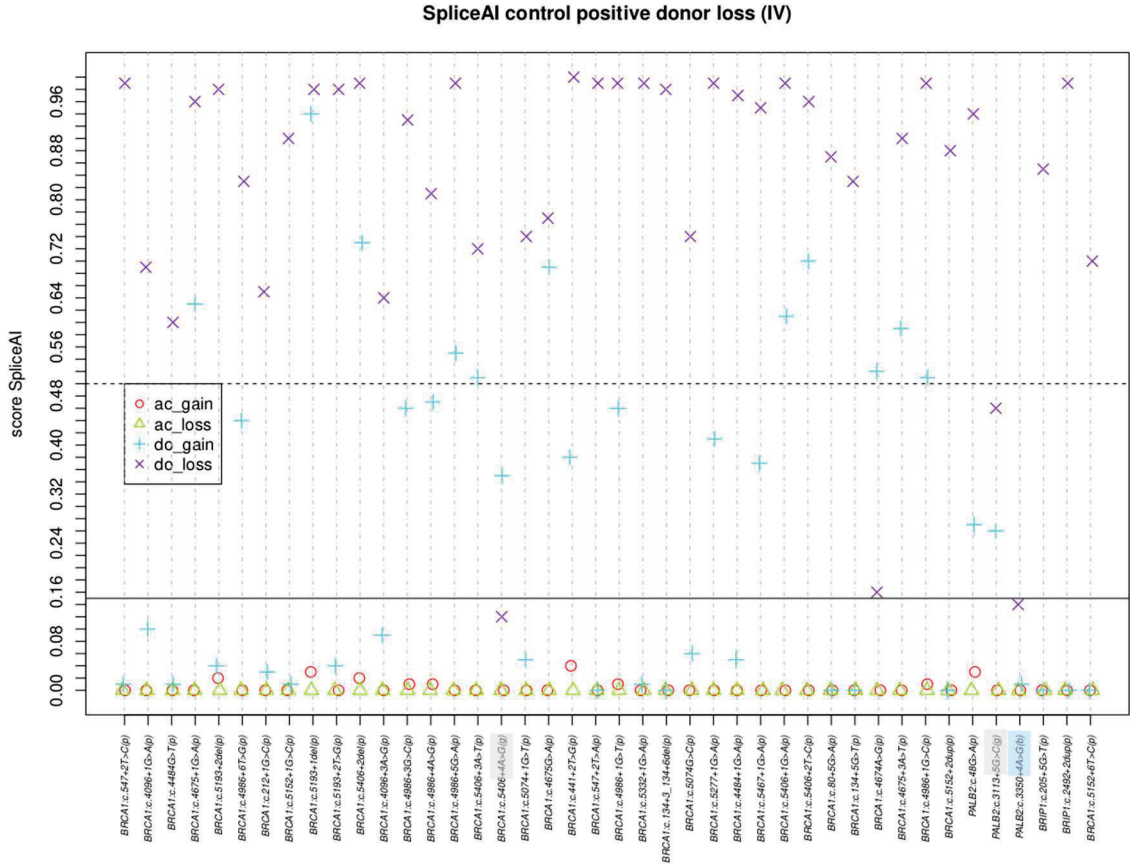

SpliceAI control positive acceptor loss (III)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

D)

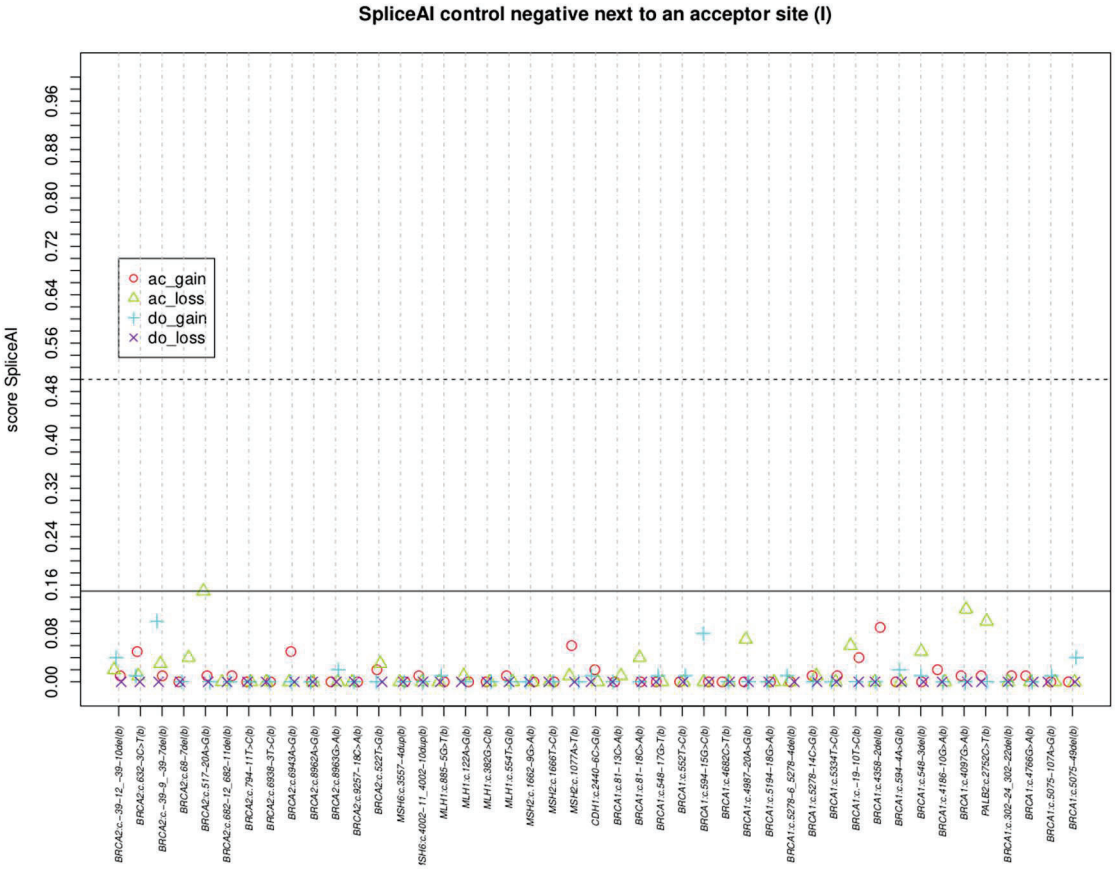

SpliceAI control positive donor loss (I)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

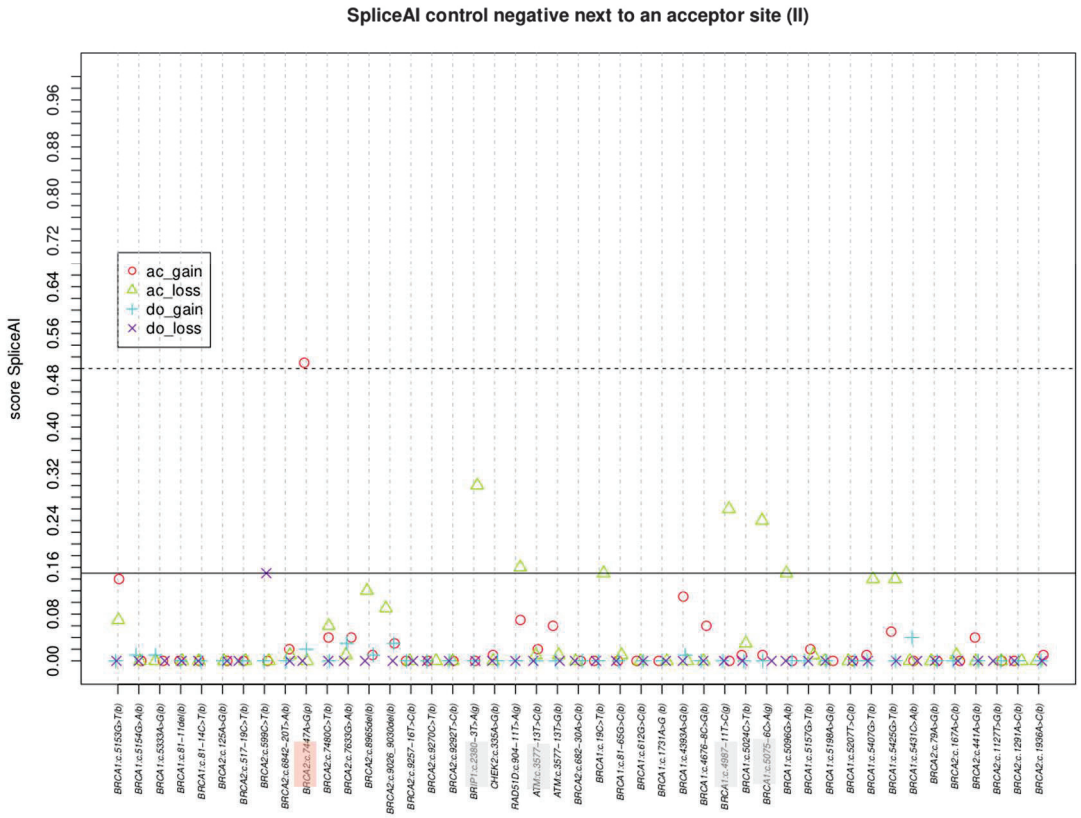

SpliceAI control positive donor loss (II)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

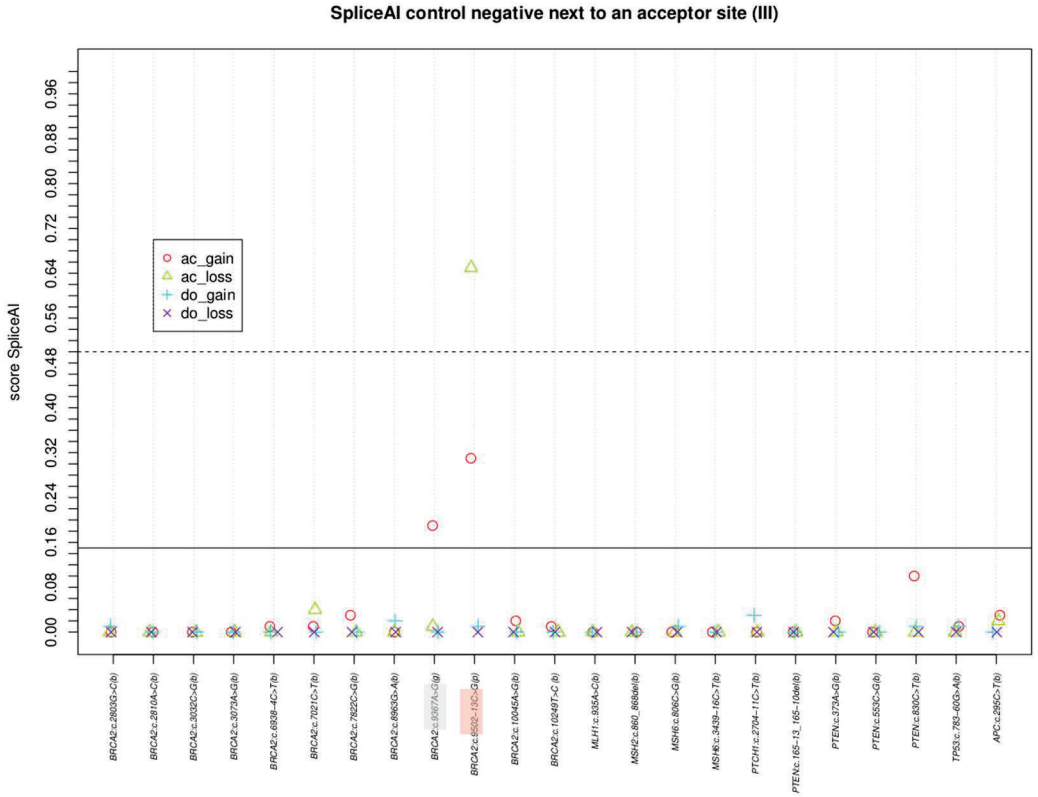

SpliceAI control positive donor loss (III)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss


SpliceAI control positive donor loss (IV)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

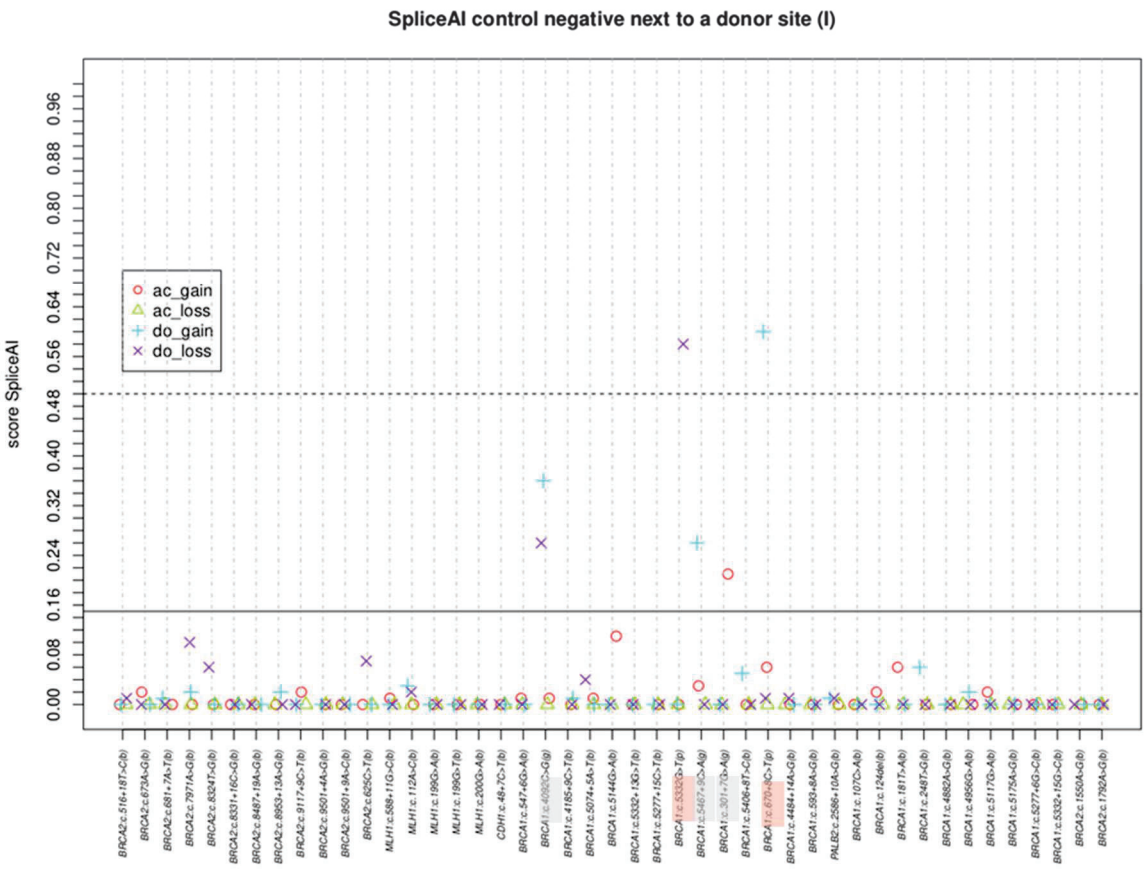
E)

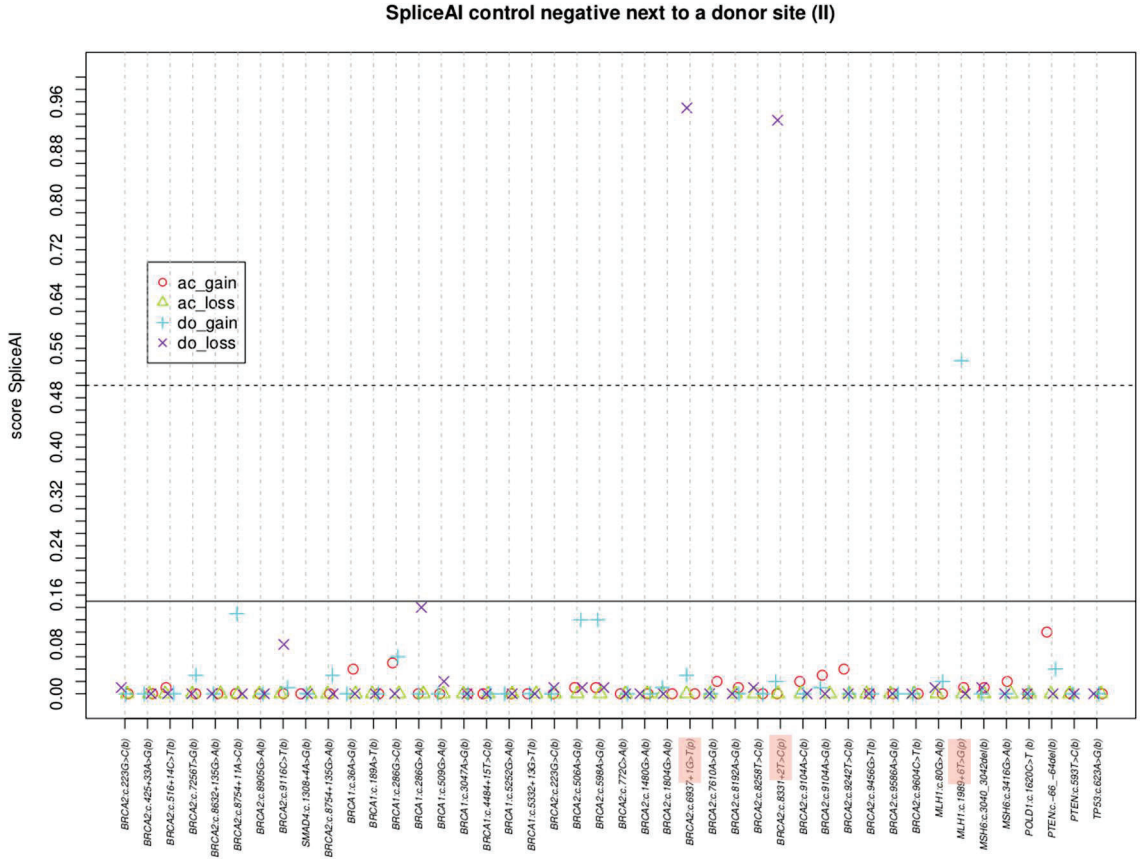

SpliceAI control negative next to an acceptor site (I)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss


SpliceAI control negative next to an acceptor site (II)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss


SpliceAI control negative next to an acceptor site (III)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

F)


SpliceAI control negative next to a donor site (I)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss


SpliceAI control negative next to a donor site (II)
score SpliceAI
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
ac_gain
ac_loss
do_gain
do_loss

Supplementary references

Agrawal, S. et al. (2005) Different splicing defects lead to differential effects downstream of the lipid and protein phosphatase activities of PTEN. Hum. Mol. Genet., 14(16), pp. 2459–2468.

Barone, G. et al. (2009) Modeling ATM mutant proteins from missense changes confirms retained kinase activity. Hum. Mutat., 30(8), pp. 1222–1230.

Chen, H.J. et al. (2017) Characterization of cryptic splicing in germline PTEN intronic variants in Cowden syndrome. Hum. Mutat., 38(10), p. 1372.

Colombo, M. et al. (2014) Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Hum. Mol. Genet., 23(14), pp. 3666–3680.

Drost, M. et al. (2019) A functional assay-based procedure to classify mismatch repair gene variants in Lynch syndrome. Genet. Med., 21(7), pp. 1486–1496.

Drost, M. et al. (2020) Two integrated and highly predictive functional analysis-based procedures for the classification of MSH6 variants in Lynch syndrome. Genet. Med. 2020 225, 22(5), pp. 847–856.

Easton, D.F. et al. (2007) A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am. J. Hum. Genet., 81(5), pp. 873–883.

Giacomelli, A.O. et al. (2018) Mutational processes shape the landscape of TP53 mutations in human cancer. Nat. Genet., 50(10), pp. 1381–1387.

InSiGHT variants databases - InSiGHT (2022).

Jaganathan, K. et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), pp. 535-548.e24.

Jia, X. et al. (2021) Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. Am. J. Hum. Genet., 108(1), pp. 163–175.

Karczewski, K.J. et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nat. 2020 5817809, 581(7809), pp. 434–443.

Kato, S. et al. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc. Natl. Acad. Sci. U. S. A., 100(14), pp. 8424–8429.

Kotler, E. et al. (2018) Functional characterization of the p53 'mutome'. Mol. Cell. Oncol., 5(6).

Landrum, M.J. et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res., 42(D1), pp. D980–D985.

Landrum, M.J. et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res., 46(D1), pp. D1062–D1067.

Lefter, M. et al. (2021) Next Generation HGVS Nomenclature Checker. Bioinformatics, 37(18), pp. 2811–2817.

Liu, X. et al. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med., 12(1), pp. 1–8.

Lyra, P.C.M. et al. (2020) Integration of functional assay data results provides strong evidence for classification of hundreds of BRCA1 variants of uncertain significance. Genet. Med. 2020 232, 23(2), pp. 306–315.

McLaren, W. et al. (2016) The Ensembl Variant Effect Predictor. Genome Biol., 17(1), pp. 1–14.

Menéndez, M. et al. (2012) Assessing the RNA effect of 26 DNA variants in the BRCA1 and BRCA2 genes. Breast Cancer Res. Treat., 132(3), pp. 979–992.

Mitui, M. et al. (2009) Functional and computational assessment of missense variants in the ataxia-telangiectasia mutated (ATM) gene: mutations with increased cancer risk. Hum. Mutat., 30(1), pp. 12–21.

Parsons, M.T. et al. (2019) Large scale multifactorial likelihood quantitative analysis of BRCA1 and BRCA2 variants: An ENIGMA resource to support clinical variant classification. Hum. Mutat., 40(9), pp. 1557–1578.

Quiles, F. et al. (2016) Investigating the effect of 28 BRCA1 and BRCA2 mutations on their related transcribed mRNA. Breast Cancer Res. Treat., 155(2), pp. 253–260.

Rayner, E. et al. (2022) Predictive functional assay-based classification of PMS2 variants in Lynch syndrome. Hum. Mutat. [Preprint].

Rofes, P. et al. (2020) Improving Genetic Testing in Hereditary Cancer by RNA Analysis: Tools to Prioritize Splicing Studies and Challenges in Applying American College of Medical Genetics and Genomics Guidelines. J. Mol. Diagn., 22(12), pp. 1453–1468.

Schwarz, J.K. et al. (2003) Regulation of the Chk2 Protein Kinase by Oligomerization-Mediated cis-and trans-Phosphorylation.

Scott, S.P. et al. (2002) Missense mutations but not allelic variants alter the function of ATM by dominant interference in patients with breast cancer. Proc. Natl. Acad. Sci. U. S. A., 99(2), pp. 925–930.

Sodha, N. et al. (2006) Rare Germ Line CHEK2 Variants Identified in Breast Cancer Families Encode Proteins That Show Impaired Activation. Cancer Res., 66(18), pp. 8966–8970.

Thomassen, M. et al. (2012) Characterization of BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. Breast Cancer Res. Treat., 132(3), pp. 1009–1023.

Trojan, J. et al. (2001) Activation of a cryptic splice site of PTEN and loss of heterozygosity in benign skin lesions in Cowden disease. J. Invest. Dermatol., 117(6), pp. 1650–1653.

Whiley, P.J. et al. (2014) Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. Clin. Chem., 60(2), pp. 341–352.

Wu, X. et al. (2006) Characterization of CHEK2 mutations in prostate cancer. Hum. Mutat., 27(8), pp. 742–747.

Wu, X. et al. (2001) Characterization of tumor-associated Chk2 mutations. J. Biol. Chem., 276(4), pp. 2971–2974.

Resultat no publicat 1

Optimizing GRIDSS for clinical use: a targeted NGS filtering strategy for germline structural variant detectiu

Elisabet Munté[¥], Paula Rofes[¥], Míriam Millán-Castillo, Ares Solanes, Xavier Muñoz, Olga Campos, Ania Alay, Mònica Salinas, Raquel Cuesta, Maria Ajenjo, Lídia Feliubadaló, Jesús del Valle* i Conxi Lázaro*.

Optimizing GRIDSS for Clinical Use: A Targeted NGS Filtering Strategy for Germline Structural Variant Detection

Elisabet Munté*[1-3], Paula Rofes*[1,2,4], Miriam Millán-Castillo[1,3] Ares Solanes [1-2], Xavier Muñoz[1-2], Olga Campos[1-2], Ania Alay[5], Mónica Salinas[1-2], Raquel Cuesta[1-2], Maria Ajenjo[5], Marta Pineda, Lidia Feliubadaló [‡1,2,4], Jesús del Valle [‡1,2,4] and Conxi Lázaro [‡1,2,4].

*Elisabet Munté and Paula Rofes contributed equally and share first authorship.

‡Lidia Feliubadaló, Jesús del Valle and Conxi Lázaro should be considered senior co-authors.

[1]Hereditary Cancer Group, Oncobell Program, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet del Llobregat, Spain

[2]Hereditary Cancer Program, Catalan Institute of Oncology (ICO), L'Hospitalet del Llobregat, Spain

[3]Doctoral Programme of Genetics, University of Barcelona, Barcelona, Spain

[4]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain

[5]Unit of Bioinformatics for Precision Oncology (UBOP), Catalan Institute of Oncology, Avinguda de la Granvia de l'Hospitalet, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain.

ABSTRACT

Structural variants (SVs) represent a significant source of genomic variation, playing a key role in the etiology of numerous genetic diseases, including cancer. Current diagnostic algorithms, primarily based on short-read NGS data and depth-of-coverage analyses, often fail to detect certain SVs. GRIDSS software integrates multiple strategies to improve SV detection, but its lack of specificity and annotation strategy complicates its clinical implementation. This study aimed to develop a pipeline incorporating filters to prioritize germline SVs in a diagnostic setting.

We analyzed 9,750 samples from patients with suspected hereditary cancer using GRIDSS. The custom R pipeline was designed to exclude recurrent variants identified in our dataset, focus on clinically relevant regions of actionable genes, and remove alignment artifacts. Candidate SVs were visually inspected in IGV, and those with potential clinical relevance were experimentally validated.

Initial analyses identified over 1.3 million candidate variants. After applying filters and performing visual inspection, 24 variants were prioritized. Of these, 15 were excluded due to limited clinical relevance, and the remaining nine were validated, confirming all as true positives. These included two pathogenic variants in *MSH6*, one likely pathogenic variant in BARD1, one likely pathogenic LINE insertion in *APC*, four likely pathogenic Alu insertions (two in *BRCA2*, one in *PALB2*, and one in *ATM*), and one variant of uncertain significance in *PALB2*. All findings were consistent with patients' phenotypes.

The restrictive filtering strategy we employed has proven applicable in a diagnostic setting, providing a reliable method to enhance the detection of clinically relevant SVs and improve hereditary cancer diagnostics.

INTRODUCTION

Targeted short-read NGS panels are widely used in routine diagnostics due to their optimal balance between cost and efficiency (Rehm, 2013). Despite their numerous advantages, targeted approaches may miss certain variants located outside the regions of interest. Moreover, short-read data may challenge the detection of certain types of variants, such as those located in homologous, low-complexity, highly variable or highly repetitive regions, as well as structural variants of

intermediate size (Lincoln et al., 2021). Although there is no consensus definition of 'intermediate size', it typically refers to variants ranging from 50 bp to 1 kb, which are particularly difficult to detect as they may partially span or even exceed the read length (Mahmoud *et al.*, 2019).

Structural variants represent a significant source of genomic variation that may play a role in numerous diseases with a genetic etiology, including cancer. Therefore, improving detection methods is crucial to uncover these variants and fully understand their role in human diseases. There are four main strategies to call structural variants from short-read data, which can be based on read depth, paired-end mapping, split reads or assembling (Escaramís et al. 2015). The read depth method relies on statistical comparisons of coverage changes across specific genomic windows, determining whether the observed coverage matches the expected levels. Paired-end mapping analyzes the distribution of paired-end reads to identify discrepancies in the expected distance, orientation, or order. The split-read strategy, as the name itself suggests, detects structural variants by identifying those reads that map to at least two genomic locations. These misaligned bases are called soft-clipped bases. Lastly, *de novo* assembly methods reconstruct genomic sequences without a reference genome. Reads are assembled into longer contiguous sequences known as contigs, which are then compared to the reference genome or another assembly.

Each method has its own strengths and limitations, thus relying on a single approach can lead to incomplete detection of some variant types. For instance, many molecular diagnostics laboratories rely exclusively on read depth-based methods, which are effective for identifying large copy-number variants but often overlook other structural variants. To address this issue, recent efforts have focused on developing tools that combine multiple strategies to improve detection capabilities. One such tool is GRIDSS (Genome Rearrangement Identification Software Suite),(Cameron *et al.*, 2017) which integrates three of the four detection methods (paired-end mapping, split reads and *de novo* assembly), offering a more comprehensive approach to increase variant identification.

However, implementing GRIDSS in routine diagnostics poses certain challenges. Originally designed for Illumina sequencing data in a whole-genome context, GRIDSS has not yet been validated for exon-targeted data. Additionally, it reports variants in a breakend notation, which, although comprehensive in describing structural variations, can be difficult to interpret. While GRIDSS offers high sensitivity, it lacks specificity, often returning many variants. Even though the authors have developed a code to filter somatic variants, no specific guidance or tools are available for filtering germline variants, further complicating its use in a diagnostic setting.

This study aims to determine the prevalence of structural variants that were previously overlooked following the read-depth approach, by adapting GRIDSS for use with gene panel data. In addition, the study seeks to establish practical filters to identify germline structural variants, making the process suitable in a routine diagnostic practice context.

METHODS

Studied Cohort

A total of 9,750 NGS panel results from patients with suspected hereditary cancer were included in this study. All of them were referred to the Molecular Diagnostics Service at the Catalan Institute of Oncology (ICO) and provided informed written consent for both diagnostic and research purposes. The study protocol was approved by the Ethics Committee of the Catalan Institute of Oncology–Bellvitge University Hospital (PR278/19).

Routine diagnostics genetic testing

DNA was extracted from peripheral blood leukocytes and genetic testing was performed using the custom NGS ICO-IMPPC Hereditary Cancer Panel (I2HCP), which includes between 122 and 165 hereditary cancer-associated genes, depending on the panel version (Castellanos *et al.*, 2017).

Samples were sequenced using three different platforms: 5,598 samples were sequenced on a NextSeq platform (average coverage = 595x, average read length = 150 bp), 2,707 on a HiSeq platform (average coverage = 868x, average read length = 250 bp), and 1,445 on a MiSeq platform (average coverage = 494x, average read length = 300 bp) (Illumina, San Diego, CA, US). Two callers were used for variant detection: VarsCan for single-nucleotide variants and short indels and DECoN for large copy-number variants (Castellanos *et al.*, 2017; Moreno-Cabrera *et al.*, 2020, 2022). Gene selection for the analysis was phenotype-driven, based on each patient's clinical presentations, and typically ranged from 6 to 20 genes, in compliance with the Catalan Health Service guidelines (Feliubadaló *et al.*, 2019).

Bioinformatic pipeline

GRIDSS

GRIDSS (v.2.13.2) software was installed, and the core script was executed on each BAM file using the default parameters to generate a GRIDSS SV VCF file. Additionally, Repeat Masker (v.4.1.5) was configured to enable the use of the gridss_annotate_vcf_repeatmasker script, which annotates the VCF file with breakpoints and single breakend inserted sequences based on RepeatMasker classification.

Variant filtering pipeline

The filter_gridss pipeline was developed in the R statistical computing environment (R-4.2.1), using functions from both R/Bioconductor and CRAN packages.

For each VCF file, breakpoints with a mate ID matching another event were paired to ensure that each variant was represented by a single row. These variants composed a dataset dedicated to two-breakend variants. Variants without a matching mate ID were stored separately in a single-breakend dataset. This separation allowed for the application of distinct filtering criteria to each dataset. Variants involving breakpoints on different chromosomes were excluded from the analysis.

The following filters were applied uniformly to both datasets: 1) exclusion of variants with identical breakpoints detected in more than ten samples; 2) exclusion of variants located outside coding regions or beyond ±150 bp from intron boundaries in genes related to the patient's phenotype; 3) exclusion of variants with a variant allele frequency (VAF) lower than 10%; 4) exclusion of variants called in genomic regions where more than ten highly similar variants were called across different samples, as these likely represented the same underlying event inaccurately called; 5) exclusion of regions classified as simple repeats or low-complexity region by RepeatMasker.

Additional dataset-specific filters were applied. For the two-breakend dataset, exclusion of deletions shorter than 20 bp. For the single-breakend dataset, the analysis was restricted to variants identified as transposable elements by RepeatMasker.

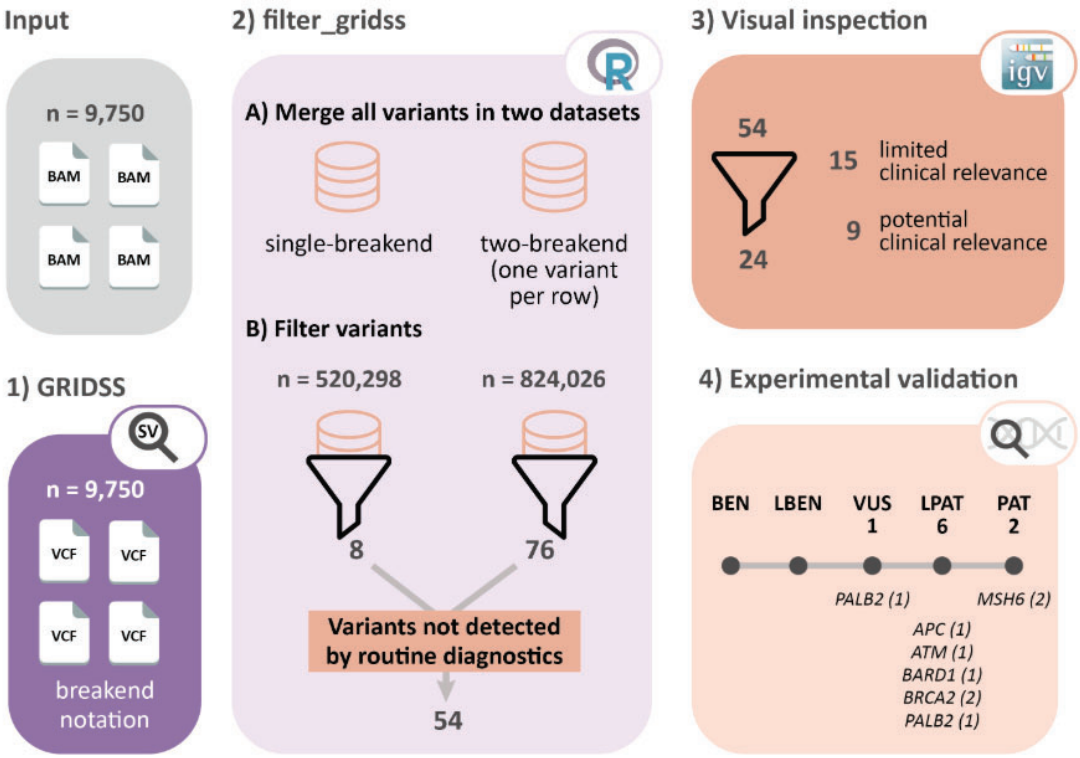
Selection of candidates by visual inspection

Following the pipeline, structural variants were visually inspected in Integrative Genomics Viewer (IGV) to identify likely true positive calls. Coverage patterns, read pair orientation and soft-clipped bases were examined, and the exact breakpoints were extracted for subsequent experimental validation.

Experimental validation

Variants selected after visual inspection were validated in genomic DNA by Sanger sequencing. PCR reactions were performed using primers flanking the expected breakpoints, with DreamTaq DNA Polymerase or Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Waltham, MA, US) according to the manufacturers' protocols. PCR products were purified with ExoSAP-IT and sequenced on an AB3500 Genetic Analyzer using BigDye™ Terminator v3.1 kit (Thermo Fisher Scientific). Primer sequences and PCR conditions are available upon request. A Long Interspersed Nuclear Element (LINE) insertion in the *APC* gene

Figure 1. Overview of the GRIDSS analysis and filtering workflow for germline structural variant detection.



was validated through long-read sequencing. The analysis followed the Comprehensive Germline Cancer Panel Workflow by Oxford Nanopore Technologies. Briefly, genomic DNA libraries were prepared using the Native Barcoding Kit 24 V14 (SQK-NBD114.24) and sequenced on a PromethION instrument with an R10.4.1 flow cell (Oxford Nanopore Technologies, Oxford, UK). *In silico* enrichment of a panel targeting 241 hereditary cancer genes was performed via adaptive sampling, and the analysis was run in EPI2ME with the workflow wf-hereditary-cancer.

An mRNA assay was performed in those patients harboring structural variants predicted to disrupt splicing, as previously described (Rofes et al., 2020) . Briefly, total RNA was isolated using TRIzol reagent from cultured peripheral blood lymphocytes treated with and without puromycin, and reverse transcribed with iScript cDNA Synthesis kit (Bio-Rad Laboratories, Hercules, CA, US). cDNA amplification was performed using exonic primers that encompassed the region of interest with DreamTaq DNA Polymerase (Thermo Fisher Scientific), and PCR products were purified and sequenced on an AB3500 Genetic Analyzer (Thermo Fisher Scientific).

Variant classification

Variants were classified following the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines (Richards *et al.*, 2015). Gene-specific guidelines developed by ClinGen's Sequence Variant Interpretation Working Group (SVI WG) were used whenever possible (accessible at: https://cspec.genome.network/cspec/ui/svi/). Variants with splicing evidence were classified according to the recommendations from ClinGen SVI Splicing Subgroup (Walker *et al.*, 2023)

RESULTS

Variant filtering strategy

The filter_gridss script was developed to prioritize germline variants from the VCF files generated by the gridss_annotate_vcf_repeatmasker script. BAM files of 9,750 samples (5,598 sequenced on a

NextSeq platform, 2,707 on a HiSeq and 1,445 on a MiSeq) were analyzed using GRIDSS and subsequently filtered using the filter_gridss script. The initial two-breakend dataset contained 824,026 variants (317,587 from NextSeq, 404,841 from HiSeq and 101,598 from MiSeq), while the single-breakend dataset included 520,298 variants (453,660 from NextSeq, 51,114 from HiSeq and 15,524 from MiSeq). After applying the filters described in the Methods section, 76 two-breakend variants and 8 single-breakend variants passed the filtering criteria. For a detailed breakdown of the number of variants filtered at each step, refer to Supplementary Figure 1. Of the remaining 84 variants, 30 had been previously detected by routine diagnostic callers (13 short indels detected by VarScan and 17 copy-number variants detected by DECoN) and were therefore disregarded. A total of 54 variants were retained for visual inspection (Figure 1).

Visual inspection of candidate variants

Visual inspection involved analyzing coverage patterns to detect regions with abnormal read depth and identifying the exact positions of breakpoints by examining soft-clipped bases. Additionally, evaluating pair orientation using the read pair option, along with aligning soft-clipped bases through IGV-assisted BLAT, provided a clearer understanding of the nature of many underlying variants. Of the 54 variants visually inspected, 30 showed no evidence of structural variation and were discarded. Of the 24 remaining variants, eight involved the *POLE or POLD1* genes, in which loss of proofreading function is associated with missense pathogenic variants (Mur *et al.*, 2020); two were located within exon 1 or intron 3 of *EPCAM* gene, regions unrelated to *MSH2* inactivation (Ligtenberg *et al.*, 2013); and five were detected within deep intronic or untranslated regions. These 15 variants were not experimentally validated due to their unexpected clinical relevance (Table 1; Figure1).

Experimental validation of candidate variants and their clinical relevance

Nine variants located within coding regions were experimentally validated and confirmed (four two-breakpoint and five one-breakpoint) (Table 1; Figure 1). Among two-breakpoint variants, two pathogenic frameshift duplications in the *MSH6* gene were identified: c.3834_3862dup (p.Lys1288Thrfs*49) and c.3922_3979dup (p.Asn1327Thrfs*11). These findings were consistent with the loss of MSH6 expression in the tumors of both patients, which helped in their classification as pathogenic, leading to a diagnosis of Lynch syndrome Additionally, the *BARD1* c.1865_1903+274del variant was identified in a breast cancer patient. Since this deletion encompassed the canonical donor site of exon 9, an mRNA was performed to assess its effect on splicing. Two alternative transcripts were detected: (1) the predominant transcript caused the skipping of exons 8 and 9 (r.1678_1903del), resulting in a frameshift predicted to trigger nonsense-mediated decay (p.Met560Glyfs*2); (2) the minor transcript resulted in the skipping of exon 9 (r.783_806del), an in-frame alteration that removed a central region within the BRCT1 domain (p.Val604_Trp635delinsGly). Consequently, this variant was classified as likely pathogenic. Lastly, an in-frame duplication in the *PALB2* gene was identified in a breast cancer patient (c.739_891dup; p.Thr247_Thr297dup). However, with the current information, it was classified as a variant of uncertain significance (VUS) (Table 2).

Among the five single-breakend variants, four were Alu insertions: one in the *PALB2* gene identified in a breast cancer patient diagnosed at age 38, one in the *ATM* gene found in a prostate cancer patient diagnosed at age 53, and two in the *BRCA2* gene, both identified in breast cancer patients with two tumor diagnoses each (ages 33 and 57 in patient 23, and ages 46 and 51 in patient 24). The fifth case involved a LINE1 element insertion in the *APC* gene, found in a patient diagnosed with adenomatous polyposis at age 14. His family history included his mother'sdiagnosis of colorectal polyposis at age 38 and his brother's diagnosis at age 18 (Table 2).

Table 1: List of structural variants identified by GRIDSS and filtered-in using our custom pipeline.

ID	Gene	Predicted breakpoint(s)	Sequencing instrument	Experim. vali? (Method)	Variant type (Length*)	Variant location
1	*ATM*	11:108190859-108190860	NextSeq	No	Duplication	Intron 44
2	*BARD1*	2:215609517-215609831	NextSeq	Yes (Sanger)	Deletion (313 bp)	Exon 9
3	*EPCAM*	2:47596709-47601826	NextSeq	No	Deletion	Exon 1
4	*EPCAM*	2:47602010-47602368	HiSeq	No	Deletion	Intron 3
5	*MSH6*	2:48033622-48033623	HiSeq	Yes (Sanger)	Duplication (29 bp)	Exon 9
6	*MSH6*	2:48033711-48033768	NextSeq	Yes (Sanger)	Duplication (58 bp)	Exon 9
7	*PALB2*	16:23646978-23647130	NextSeq	Yes (Sanger)	Duplication (153 bp)	Exon 4
8	*POLD1*	19:50902799-50902845	HiSeq	No	Deletion	Intron 3
9	*POLD1*	19:50902799-50902845	NextSeq	No	Deletion	Intron 3
10	*POLE*	12:133235843-133235844	MiSeq	No	Duplication	Intron 26
11	*POLE*	12:133235853-133235880	NextSeq	No	Deletion	Intron 26
12	*POLE*	12:133235865-133235866	MiSeq	No	Duplication	Intron 26
13	*POLE*	12:133240781-133240821	HiSeq	No	Deletion	Intron 22
14	*POLE*	12:133250054-133250126	NextSeq	No	Deletion	Intron 13
15	*POLE*	12:133250054-133250126	NextSeq	No	Deletion	Intron 13
16	*RAD51D*	17:33446835-33446874	NextSeq	No	Deletion	5' UTR
17	*SMAD4*	18:48556699-48556700	NextSeq	No	Duplication	5' UTR
18	*APC*	5:112174697	HiSeq	Yes (LRS)	LINE1 insertion (L1Ta1d)	Exon 16
19	*ATM*	11:108204717	NextSeq	No	Alu insertion (AluYa5)	Intron 54
20	*ATM*	11:108204717	NextSeq	No	Alu insertion (AluYa5)	Intron 54
21	*ATM*	11:108106407	NextSeq	Yes (Sanger)	Alu insertion (AluYa5)	Exon 5
22	*BRCA2*	13:32893302	NextSeq	Yes (Sanger)	Alu insertion (AluYa5)	Exon 3
23	*BRCA2*	13:32910689	NextSeq	Yes (Sanger)	Alu insertion (AluYb8)	Exon 11
24	*PALB2*	16:23614840	HiSeq	Yes (Sanger)	Alu insertion (AluYb8)	Exon 13

*Lenght only specified if experimental validation was performed. Abbreviations: bp: base pairs; LINE: long interspersed nuclear element; LRS: long-read sequencing; UTR: untranslated region.

Table 2: Experimentally validated structural variants and clinical information of patients and relatives.										
	Structural variant information						Clinical Information			
	Gene	Variant nomenclature (c.)[1]	Variant nomenclature (r.)[1]	Variant nomenclature (p.)[1]	Variant classification (Score[1,2])	ACMG/AMP Criteria	Proband phenotype (age at diagnosis)	Family history of FDR (age at diagnosis)	Family history of SDR or TDR (age at diagnosis)	(Likely) pathogenic variants in other cancer-susceptibility genes
2	*BARD1*	c.1865_1903+274del	r.[1678_1903del,1811_1903del]	p.[Met560Glyfs*2, Val604_Trp635delinsGly]	LP (9)	PVS1 (RNA) + PM2_sup	BR (35)	- Mother: MEL (28) - Father: MEL (55)	- Maternal grandfather: CRC (U) - Paternal aunt: BR_nc (50) - Paternal uncle: LK_nc (55)	Not identified
5	*MSH6*	c.3834_3862dup	-	p.Lys1288Thrfs*49	P (10)	PVS1 + PP4 + PM2_sup	CRC (55)	- Mother: PAN_nc (83) - Father: BL (65), CRC (77), PELV (82)	- Maternal aunt: STO_nc (75) - Maternal grandfather: LV_nc (U)	Not identified
6	*MSH6*	c.3922_3979dup	-	p.Asn1327Thrfs*11	P (10)	PVS1 + PP4 + PM2_sup	CRC (57), ENDO (62)	- Mother: BR (63)	- Maternal uncle: PENIS (65), PAN (68) - Maternal uncle: PR_nc (U) - Maternal aunt: OV_nc (U)	*ATM* c.6289G>T; p.Glu2097* (P)
7	*PALB2*	c.739_891dup	-	p.Thr247_Thr297dup	VUS (2)	PVS1_Sup + PM2_Sup	BR (48, 59)	- Father: BL_nc (U), LG_nc (U)	- Maternal aunt: BR_nc (58) - Paternal aunt: BR_nc (53), ENDO_nc (U) - Paternal aunt: BR (65) - Paternal aunt: ENDO_nc (U) - Paternal uncle: LG_nc (60)	Not identified
18	*APC*	c.3406_3407insLINE1	-	p.(Glu1136Glyfs*9)	LP	pending	CR polyp (14)	- Mother: CR polyp (38), CRC (63, 63) - Brother: CR polyp (18)	- CRC_nc (80)	Not identified
21	*ATM*	c.342_343insAluYa5		p.(Leu115Glyfs*40)	LP	pending	PR (53)	- Mother: MM (74)	- Paternal uncle: PR (86) - Paternal cousin: LG (54)	Not identified
22	*BRCA2*	c.156_157insAluYa5	-	p.(Lys53Alafs*9)	LP	pending	BR (33, 57)	- Father: THY_nc (60) - Brother: HN (61)	- Maternal uncle: LG_nc (U) - Paternal cousin: BR_nc (50)	Not identified
23	*BRCA2*	c.2197_2198insAluYb8		p.(Val733Glyfs*32)	LP	pending	BR (46, 51), HN (54)	- Mother: BR (68, 78) - Sister: BR (36)	- Maternal cousin: BR (37)	Not identified
24	*PALB2*	c.3501_3502insAluYb8	-	p.(Asp1168Trpfs*32)	LP	pending	BR (38)			Not identified

1. Lenght, variant nomenclature and variant classification only specified if experimental validation was performed. For RNA variant nomenclature, an mRNA assay was required. 2. Scored ACMG/AMP classification reference: Tavtigian et al., 2020 (PMID: 32720330). Abbreviations: LP: likely pathogenic variant; nc: not confirmed cancer diagnosis; P: pathogenic variant; U: unknown age at diagnosis; VUS: variant of uncertain significance. Cancer abbreviations: BL: bladder; BR: breast; CRC: colorectal; ENDO: endometrial; HN: head and neck; LG: lung; LK: leukemia; LV: liver; MEL: melanoma; MM: multiple myeloma; OV: ovarian; PAN: pancreatic; PELV: renal pelvic; PR: prostate; STO: gastric; THY: thyroid...

DISCUSSION

Our study aimed to adapt GRIDSS for detecting germline structural variants of intermediate size from targeted NGS data, addressing a gap in routine diagnostics. By implementing a customized filtering pipeline, we identified eight (likely) pathogenic variants, increasing the diagnostic yield by 0.6%. In terms of colorectal cancer susceptibility, we diagnosed two patients with Lynch syndrome and one patient with familial adenomatous polyposis. Additionally, we identified four (likely) pathogenic variants in breast cancer susceptibility genes and one likely pathogenic variant in a prostate cancer patient, highlighting the role of structural variants in the missing heritability of cancer. These diagnoses are of high clinical value, involving high- to moderate-risk genes with well-established management, surveillance, and treatment protocols. Furthermore, other family members may benefit from predictive testing, allowing for personalized risk management and prevention strategies.

Notably, five of the eight (likely) pathogenic variants from our dataset were mobile element insertions, including *Alu* and LINE elements. While their detection is challenging due to their repetitive nature and ubiquitous presence in the genome, mobile elements are estimated to account for up to 0.3% of all disease-causing variants (Qian et al., 2017). Therefore, our findings further reinforce the importance of incorporating mobile element detection strategies into routine diagnostic pipelines.

When developing a new tool for clinical practice, achieving an optimal balance between sensitivity and specificity is crucial. Our filtering strategy was designed to reduce variant burden and focus on clinically relevant findings. Variants frequently detected in our dataset (≥10 samples) were disregarded. Although this cutoff was arbitrary, it was intended to reduce false positive calls, sequencing artifacts and polymorphisms. However, it could hinder the detection of recurrent or founder disease-causing variants that other tools might miss. Likewise, we implemented a filter to discard highly similar variants detected in more than 10 samples, assuming these would likely represent the same underlying event miscalled multiple times. Recognizing that this threshold might exclude frequent structural variants, we adjusted it to 15 samples, thus identifying an AluYa5 insertion in intron 54 of the *ATM* gene (c.8010+30_8010+31insAluYa5; p.(Val2671Serfs*17)). This insertion was found in five breast cancer patients and six additional individuals with no clinical suspicion of *ATM*-related conditions (including melanoma, polyposis, ovarian, and gastric cancers). Previously reported (Klein *et al.*, 2023), this variant is known to cause exon 54 skipping in 38% of total *ATM* transcripts in heterozygous carriers, with incomplete expressivity and probably reduced penetrance due to the leaky splicing effect. While our filtering pipeline would not have detected this variant, its apparently reduced lifetime cancer risk supports our strategy of prioritizing variants with clearer pathogenic impact.

We also acknowledge that setting a 10% VAF threshold may be too restrictive, particularly for mosaic variants and those located in challenging regions, where reads may be sparse or poorly aligned. Additionally, excluding repetitive regions to minimize sequencing artifacts could result in the loss of clinically relevant variants located within these regions.

While GRIDSS is a powerful tool, its performance can be influenced by sequencing quality. In our dataset, an AluYb8 insertion in *BRCA2* was identified in patient 23. Although typically one proband per family is studied, two family members were also included in the study: her sister, diagnosed with breast cancer at age 36, and a distant cousin, diagnosed with breast cancer at age 37. Initially, the *Alu* insertion was not detected in either relative. However, upon further inspection of the VCF files, the variant was found in the sister's data but had failed to meet the GRIDSS quality threshold of 1,500 (quality score = 918) and was subsequently discarded. Additionally, there were few reads supporting this variant (VAF = 6.4%). Visual inspection using IGV, however, suggested that the variant was present. It is plausible that other structural variants with low quality parameters may have remained undetected due to this same issue.

It is important to note that GRIDSS cannot detect structural variants with breakpoints located outside covered regions. While this is usually not an issue for whole-genome sequencing data, it certainly hinders the detection of structural variants in targeted approaches. In contrast, read-depth methods could identify some of these variants if the affected region includes at least one exon. Given the strengths and limitations of each tool, combining methods based on different strategies is a more suitable approach to optimize the detection of all variant types and sizes.

Despite these challenges, our approach has advanced the detection of clinically relevant structural variants. The restrictive filtering strategy we employed has shown practical applicability in a diagnostic setting, offering a reliable method for identifying high-impact variants in clinical practice.

REFERENCES

Cameron, D.L. *et al.* (2017) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res, 27(12), pp. 2050–2060.

Castellanos, E. *et al.* (2017) A comprehensive custom panel design for routine hereditary cancer testing: preserving control, improving diagnostics and revealing a complex variation landscape. Scientific Reports 2017 7:1, 7(1), pp. 1–12.

Escaramís, G. *et al.* (2015) A decade of structural variants: description, history and methods to detect structural variation.

Feliubadaló, L. *et al.* (2019) Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. Int. J. Cancer, 145, pp. 2682–2691.

Klein, J. *et al.* (2023) A Novel Alu Element Insertion in ATM Induces Exon Skipping in Suspected HBOC Patients. Hum Mutat, 2023.

Ligtenberg, M.J.L. *et al.* (2013) EPCAM deletion carriers constitute a unique subgroup of Lynch syndrome patients. Fam Cancer.

Lincoln, S.E. *et al.* (2021) One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genetics in Medicine 2021 23:9, 23(9), pp. 1673–1680.

Mahmoud, M. *et al.* (2019) Structural variant calling: The long and the short of it. Genome Biol, 20(1), pp. 1–14.

Moreno-Cabrera, J.M. *et al.* (2020) Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. European Journal of Human Genetics 2020 28:12, 28(12), pp. 1645–1655.

Moreno-Cabrera, J.M. *et al.* (2022) Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer. J Med Genet, 59(1), pp. 75–78.

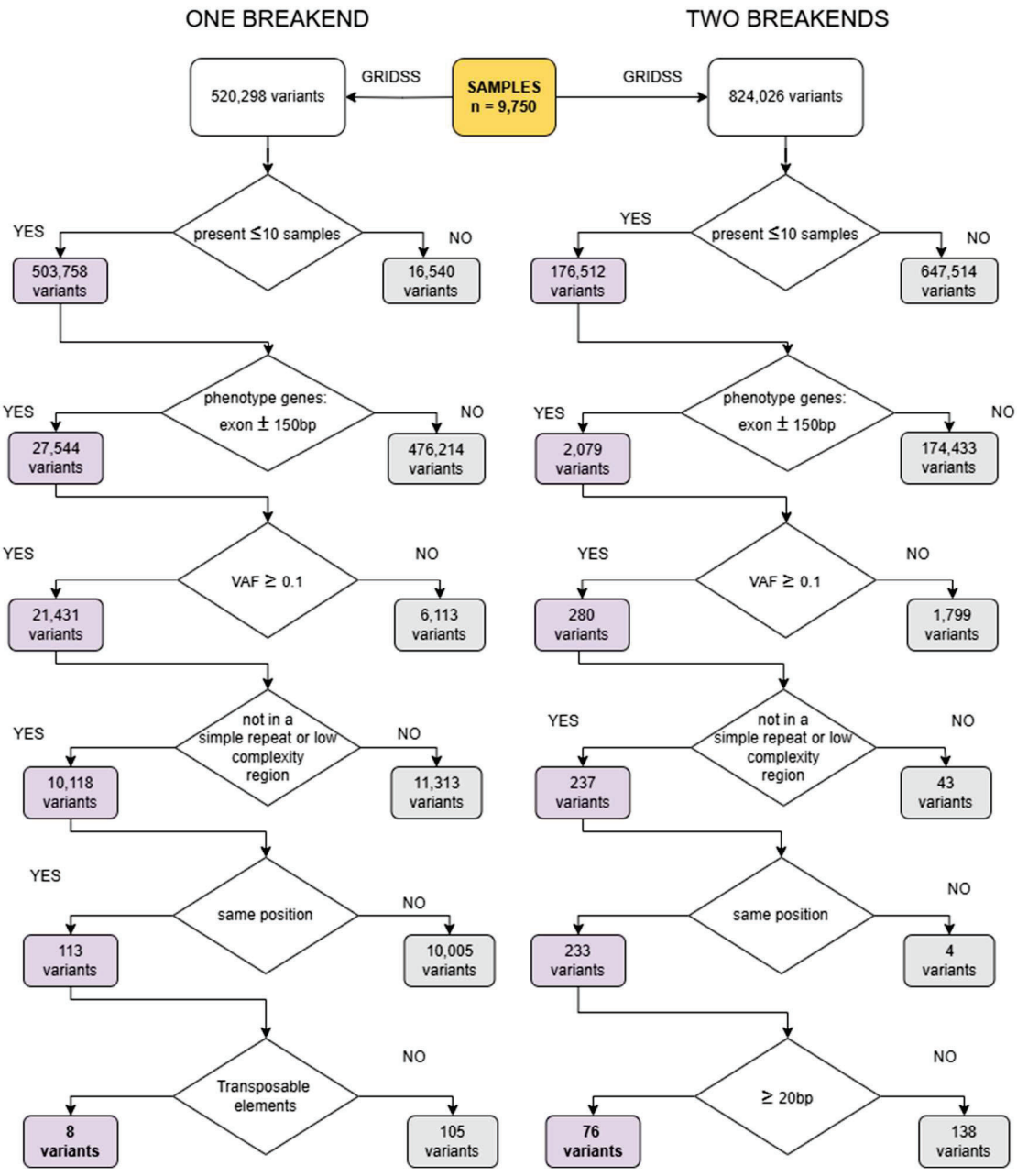
Mur, P. *et al.* (2020) Role of POLE and POLD1 in familial cancer. Genetics in Medicine, 22(12).

Rehm, H.L. (2013) Disease-targeted sequencing: A cornerstone in the clinic. Nat Rev Genet.

Richards, S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5), pp. 405–424.

Walker, L.C. *et al.* (2023) Application of the ACMG/AMP framework to capture evidence relevant to predicted and observed impact on splicing: recommendations from the ClinGen SVI splicing. medRxiv [Preprint].

Supplementary Figure 1. Breakdown of variants filtered at each step.



Resultat no publicat 2

Identifying potential pathogenic variants in 5'UTR regions within a hereditary cancer cohort

Elisabet Munté, Lidia Feliubadaló, Alexandra Martin-Geary, Conxi Lázaro i Nicola Whiffin

Identifying potential pathogenic variants in 5'UTR regions within a hereditary cancer cohort

Elisabet Munté[1,2], Lídia Feliubadaló[1,3], Alex Martin-Geary[4,5], Conxi Lázaro[1,3*] and Nicola Whiffin[4,5,6*].

[1]Hereditary Cancer Program, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge - IDIBELL-ONCOBELL, 08908, L'Hospitalet de Llobregat, Spain.

[2]Doctoral Programme in Biomedicine, University of Barcelona (UB), Barcelona, Spain.

[3] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain.

[4]Big Data Institute, University of Oxford, Oxford, UK

[5]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.

[6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

ABSTRACT

Molecular diagnostics often focus on coding regions, overlooking untranslated regions (UTRs) despite their critical role in post-transcriptional regulation. Variants in 5'UTRs can disrupt translation in various ways, such as affecting the ribosome's ability to recognize the start codon, altering or creating upstream open reading frames (uORFs), or impacting splicing.

We analyzed the 5'UTR of 55 clinically relevant hereditary cancer genes in 4,533 samples. Variants were prioritized according to Martin-Geary et al. (2023) filters, targeting those creating/disrupting uORFs, modifying Kozak sequences, or impacting splicing.

From 86,694 identified variants, 860 were unique, and fourteen remained after applying a filtering process. Four variants were prioritized by UTRannotator, two by SpliceAI, and one based on its location affecting the Kozak sequence. Among all these candidate variants, only one variant, *CDKN2A* c.-34G>T was classified as pathogenic and correlated with the patient phenotype.

This study highlights the feasibility of integrating 5'UTR variant analyses into diagnostics and their potential clinical relevance. Broader UTR coverage and experimental validation are essential to confirm or exclude the functional impact of candidate variants and enhance diagnostic approaches.

INTRODUCTION

To date, most molecular diagnostics units predominantly use exome sequencing (ES) or targeted gene panels, focusing on coding regions and their immediate surroundings. Genome sequencing (GS), which enables detection of variants in all regions of the genome, is becoming more common in diagnostics and is expected to lead the field in the coming years. However, the complexity of variant interpretation often limits studies to coding regions, despite having data from the non-coding parts of the genome.

Therefore, untranslated regions (UTRs), which are present in the mRNA but not translated into protein, are frequently overlooked. However, these regions play a crucial role in post-transcriptional regulation, influencing RNA stability, intracellular localization and translation efficiency

(Bashirullah, Cooperstock and Lipshitz, 2001; Jansen, 2001; Mignone *et al.*, 2002).

Of particular interest are 5'UTR regions, as they play a key role in regulating the rate at which the coding sequence of a gene is translated (Van Der Velden and Thomas, 1999). They contain a multitude of regulatory elements, including upstream open reading frames (uORFs), located before the main coding sequence. uORFs are elements that are present due to an upstream AUG (uAUG) start codon in the 5'UTR (Morris and Geballe, 2000; Pesole *et al.*, 2001). They can either be non-overlapping if they have initiation and termination codons before the main ORF. In contrast, upstream overlapping ORFs (uoORFs) do not have a stop codon before the main ORF. These uoORFs can either be in-frame with the main ORF, resulting in translation of an N-terminally extended protein product (NTE), or out-of-frame, terminating at a different stop codon to the main ORF (Calvo, Pagliarini and Mootha, 2009).

The efficiency of translation initiation from any AUG start codon is largely influenced by the local sequence context, the Kozak sequence, which surrounds the AUG (Kozak, 1986, 1987a, 1987b). Positions -3 and +4 relative to the A in the initiator AUG have been identified as the most critical for ribosomal recognition (Kozak, 1986, 1997). A strong Kozak sequence enhances the likelihood of the ribosome initiating translation at the uORF, thereby influencing the overall expression of the main ORF. uORFs can thus act as modulators of gene expression, often competing with the main ORF for ribosomal binding and impacting the protein level. Subsequent studies have shown that positions -2, -4, and +5 also influence translation efficiency, experimentally evaluating all possible combinations (NNNNNNAUGNN, where N represents any nucleotide) to determine their effect on translational efficiency (Noderer et al., 2014).

Variants in 5'UTRs can disrupt regulatory functions in multiple ways. For example, variants located just upstream of the CDS that alter the Kozak sequence may impact translation efficiency by affecting ribosome's ability to recognize the start codon (Mohan *et al.*, 2014). Additionally, certain variants can create new uORFs by introducing a uAUG (Wright *et al.*, 2021), eliminate existing uORFs by deleting a uAUG, or alter an existing uORF by removing its stop codon (uStop) or causing an upstream frameshift (uFrameshift) that changes the reading frame of a uORF (Filatova *et al.*, 2021). Splicing variants within the 5'UTR can also introduce uFrameshifts or affect the length, inclusion, or exclusion of uORFs, altering translation efficiency in various ways (Filatova *et al.*, 2021).

The Kozak sequence of these uORFs, along with the nature of the uORF, are critical in determining the regulatory impact of these variants, which could potentially contribute to diseases. These types of variants have been demonstrated to be under strong negative selection (Whiffin *et al.*, 2020).

Considering all these, UTRannotator was developed to prioritize variants in 5'UTR regions that create or eliminate uORFs (Zhang *et al.*, 2021). It identifies the type of uORF and assesses the strength of the Kozak sequence. Additionally, SpliceAI is widely used to prioritize splicing variants, aiding in the identification of potential splicing aberrations (Jaganathan *et al.*, 2019). Both UTRannotator and SPliceAI are integrated into Ensembl's Variant Effect Predictor (VEP) (McLaren *et al.*, 2016).

To facilitate the analysis of 5'UTR variants, Martin-Geary et al. (2023) suggested an annotation approach that prioritizes variants in diagnostic settings, allowing for the identification of candidate variants without substantially increasing the number of variants to be classified. These candidate variants must then undergo functional validation to confirm their impact.

In this study, we used UTRannotator and applied the filters proposed by Martin-Geary et al. (2023) to evaluate a cohort of patients with suspected hereditary cancer, with the aim of identifying potential pathogenic variants within their 5'UTR regions.

METHODS

Study Cohort

A total of 4533 samples from patients with suspected hereditary cancer were analyzed. These patients had previously undergone

genetic testing at the Molecular Diagnostic Laboratory of the Catalan Institute of Oncology, and all provided informed consent for both diagnostic and research purposes.

Routine diagnostics pipeline

DNA was extracted from peripheral blood leukocytes, and all samples were sequenced using a NextSeq550 platform (Illumina, San Diego, CA, USA). A customized NGS gene panel, the ICO-IMPPC Hereditary Cancer Panel (I2HCP, v3), was used, including 165 genes. For 24 genes, the capture included all exons of the 5'UTR region. For the remaining genes, only about 150 bp upstream of the AUG start codon were well covered. For diagnostic purposes, only genes related to the individual's phenotype and variants located in coding regions or within 20 base pairs of these regions had been analyzed.

5'UTR regions definition

The bed file utilized in the I2HCP diagnostic pipeline does not encompass the complete 5'UTR regions for most genes. To address this, a custom bed file was created specifically for the 5'UTR regions of the 55 genes included in the Catalan Health Instruction guidelines and covered in I2HCP v3 (Table 1). These genes were selected due to their recognized clinical relevance and, therefore, are the ones used for diagnostic purposes.

The transcripts selected for all genes were the MANE Select transcripts, with the exception of *CDKN2A*, for which the MANE Plus Clinical transcript was also included (Supplementary Table 1). The 5'UTR regions were extracted from the .gff file provided by Ensembl's latest GRCh38 release (https://ftp.ensembl.org/pub/release-113/gtf/homo_sapiens/Homo_sapiens.GRCh38.113.gtf.gz). However, as the alignment in our diagnostic pipeline is performed using the GRCh37 genome, a liftover was applied to convert these positions to their GRCh37 coordinates.

Extent of UTR coverage

To determine the extent of UTR coverage for each gene, we analyzed the distance between the transcription start site and the translation start site for each transcript of interest. Transcripts with complete exonic coverage of the 5'UTR (though not necessarily covering intronic regions) and those with a 5'UTR shorter than 150 bp consisting of a single exon were considered well-covered (Supplementary Table 1).

Variant calling at 5'UTR

Variant calling was performed using VarDict Java (v1.8.3; *https://github.com/AstraZeneca-NGS/VarDictJava*). The variant allele frequency (VAF) threshold was set at 0.1, and the custom bed file was used to restrict the variant calls to 5'UTR regions. All other parameters were left at default settings. Complex variants, as defined by VarDict Java (i.e., multiple alterations in close proximity, such as an insertion near a singe nucleotide variant or a deletion), and variants supported by fewer than six reads were not considered.

Table 1: List of the 55 clinically relevant genes included in the Catalan Health Institute instruction guidelines and in I2HCPv3 panel

AIP	*CDK4*	*MAX*	*POLD1*	*SDHAF2*
APC	*CDKN1B*	*MEN1*	*POLE*	*SDHB*
ATM	*CDKN2A*	*MET*	*POT1*	*SDHC*
BAP1	*CHEK2*	*MLH1*	*PRKAR1A*	*SDHD*
BARD1	*CTNNA1*	*MSH2*	*PRSS1*	*SMAD4*
BMPR1A	*DICER1*	*MSH6*	*PTEN*	*STK11*
BRCA1	*EPCAM*	*MUTYH*	*RAD51C*	*TMEM127*
BRCA2	*FH*	*NF1*	*RAD51D*	*TP53*
BRIP1	*FLCN*	*NTHL1*	*RET*	*TSC1*
CDC73	*KIF18*	*PALB2*	*RNF43*	*TSC2*
CDH1	*KIT*	*PMS2*	*SDHA*	*VHL*

Variant prioritization

All resulting variants were processed using Ensembl's VEP with both UTRannotator and SpliceAI enabled. Only variants from the transcripts of interest were selected (Supplementary Table 1). Variants were prioritized according to any of the following criteria (Martin-Geary et al., 2023): 1. UTRannotator annotation suggested one of the following scenarios: a) Gain of a uAUG creating an oORF with a strong or moderate Kozak sequence; b) Loss of uStop with no other stop codon upstream of the main ORF with a strong or moderate Kozak sequence; c) Loss of a uAUG with a strong Kozak sequence or d) uFrameshift resulting in an oORF with a strong or moderate Kozak sequence. 2. Any of their SpliceAI delta scores were ≥ 0.20. 3. They were located at position -3 relative to the main ORF, where an A or G changed to C or T.

SpliceAI-10k was used to help interpret the splicing effect of variants with any SpliceAI delta score ≥0.20 (Canson *et al.*, 2023).

GnomAD v4.1.0 was utilized to determine the minor allele frequency (MAF) of the candidate variants in the general population (Chen et al., 2023).

The translation efficiency rate predictions used in this study were derived from the work of Noderer et al. (2014).

RESULTS

After analyzing the length of the 5'UTRs, we found that 24 of the 55 clinically relevant genes included in the I2HCP v3 panel have a short, single-exon 5'UTR that are fully covered, while an additional 11 genes have at least all 5'UTR exons and 150 bp of intronic boundaries covered (see Supplementary Table 1). Variant calling was performed on NGS results from 4,485 samples, focusing exclusively on these 5'UTR regions. A total of 86,694 variants were identified, of which 860 were unique. After filtering (see Methods), 14 variants remained.

Variants selected with UTRannotator

Although UTRannotator provided annotations for eighteen variants (Table 2), only four met the specific filtering criteria (Table 3).

Table 2: Summary of the total number of variants annotated in the 5'UTR by UTRannotator, categorized by consequence. For each consequence, the resulting ORF type and the Kozak strength of the gained or lost ORF are specified. Cases highlighted in gray indicate where variants would meet the filtering criteria if present in the dataset.

Consequence		Type of ORF regarding consequence		Kozak strength of each type		
5'UTR consequence	**Total**	**types**	**Total**	**Strong**	**Moderate**	**Weak**
premature start codon gain variant	9	inFrame_oORF	2	**1**	0	1
		OutOfFrame_oORF	4	**1**	**1**	2
		uORF	3	0	2	1
uORF stop codon loss variant	1	another stop codon upstream of the main ORF	0	0	0	0
		no other stop codon upstream of the main ORF	1	**1**	**0**	0
premature start codon loss variant	6	inFrame_oORF	0	0	0	0
		OutOfFrame_oORF	0	0	0	0
		uORF	6	0	3	3
uORF frameshift variant	2	OutOfFrame_oORF	0	0	0	0
		uORF	2	0	1	0
TOTAL	**18**					

Table 3: Variants selected by UTRannotator filters. The type of uORF and the Kozak strength predicted by UTRannotator are shown, as well as the translation efficiency of the main ORF and the new uORF. Additionally, the table indicates the minor allele frequency (MAF) in GnomAD v.4.1.0, the number of carrier patients whose clinical suspicion was related to the variant gene, and the number of carrier patients where it was not.

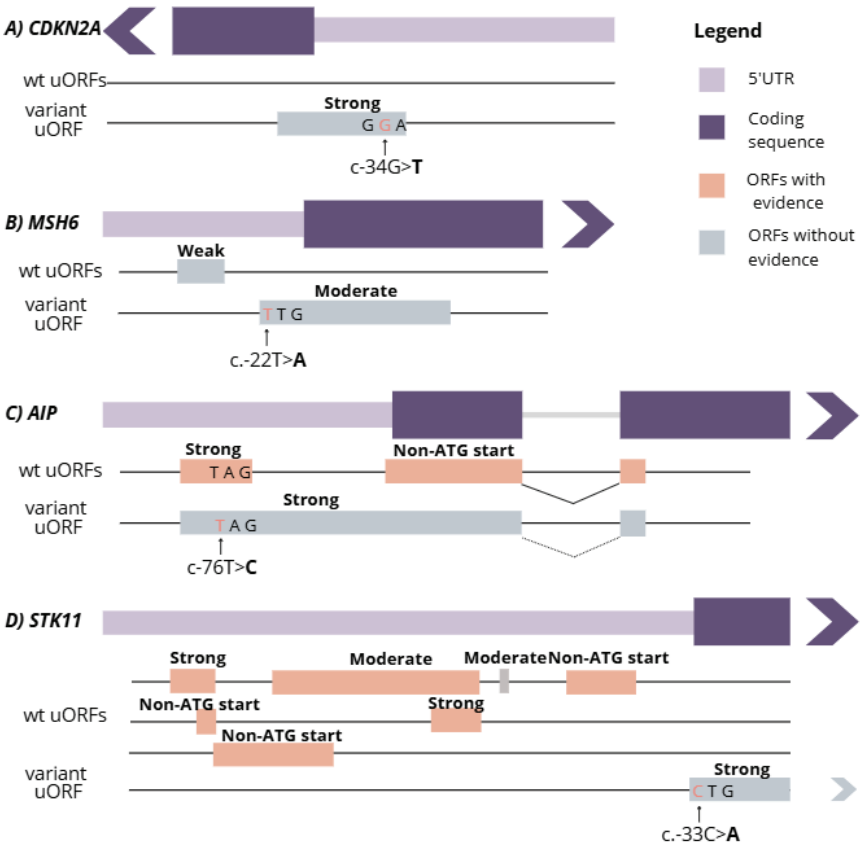
Variant description		UTRannotator		Predicted translation efficiency		GnomAD (v.4.1.0)	Nº of individuals with clinical phenotype*	
Gene	**Variant**	**Type**	**Kozak strength**	**oORF**	**Main ORF**	**MAF (%)**	**Yes**	**No**
CDKN2A	c.-34G>T	premature start codon gain variant -> OutOfFrame oORF	Strong	100	79	0.0060 (91/ 1509400)	1	2
MSH6	c.-22T>A	premature start codon gain variant -> OutOfFrame oORF	Moderate	71	103	0.0001 (2/16109 06)	1	0
AIP	c.-76T>C	uORF stop codon loss variant -> no other stop codon upstream of the main ORF	Strong	109	135	0	0	1
STK11	c.-33C>A	premature start codon gain variant -> InFrame_oORF	Strong	113	80	0	1	0

* The clinical indication is based on guidelines in the Catalan Health Instruction (Supplementary File 1)

Two of them create a uAUG that generates an uoORF out-of-frame with the CDS (Figure 1, A and B). The *CDKN2A* c.-34G>T variant has been previously reported in the literature as pathogenic (Liu *et al.*, 1999). This variant was identified in three unrelated probands. One proband had melanoma at the age of 55 and breast cancer at the age of 70, and a niece with melanoma at the age of 28. The other two probands were diagnosed with ovarian cancer. Additionally, in one of the ovarian cancer cases, there were unconfirmed diagnoses of skin cancer in the family, affecting the proband's mother and aunt. The *MSH6* c.-22T>G variant is located in a moderate Kozak context. This variant was found in a patient with polyposis who had developed endometrial cancer. However, immunohistochemistry performed on the tumor showed preserved expression of the MSH6 protein. Additionally, the patient's mother and grandmother both had colorectal cancer, but genetic panel testing on the mother revealed that the variant did not co-segregate with the disease. This variant was not therefore considered to be a good candidate to explain disease in this family.

The *STK11* c.-33C>A variant creates a uAUG with a Strong Kozak context, predicted to result in an in-frame oORF with the CDS (Figure 1, C). The patient was tested with a polyposis panel that includes *STK11*. Mutations in this gene are typically associated with hamartomatous (juvenile-type) polyps, characteristic of Peutz-Jeghers syndrome. However, after reviewing the only report available for the patient, it specifies that the patient had >100 polyps, some of which were hyperplastic, with no indication that any were hamartomatous. Given this information, we would not initially consider this condition to be *STK11*-related.

Figure 1: Schematic representation of the N terminus of genes with candidate variants prioritized by UTRannotator. For each gene, all wild-type uORFs, both those with and without supporting evidence, are shown. In cases A, B, and C, the variants create an out-of-frame uORF overlapping the coding sequence (CDS) (oORF-creating). If translation initiates at the uAUG, the ribosome will not translate the CDS. In case C, the variant creates a uAUG that is in-frame with the CDS. If translation initiates at this point, an elongated protein will be translated.



The remaining variant, *AIP* c.-76T>C, deletes a uORF stop codon, with no other stop codon before the main ORF's first methionine, leading to an out-of-frame oORF in a strong Kozak context (Figure 1, D). This variant was found in a patient with prostate cancer. However, *AIP* is not indicated for analysis with this phenotype (Supplementary File 1); it is typically indicated for certain pituitary tumors.

Variants selected with SpliceAI

A total of eleven variants had at least one SpliceAI delta score of 0.20 or higher (Table 4). Two variants were dismissed as no splicing aberration was predicted by SpliceAI-10k. Of the nine left, two were filtered out due to a minor allele frequency (MAF) >1%. Finally, five variants were excluded as they were not found exclusively in samples from patients whose phenotype was related to their gene, leaving two variants for further inspection (Figure 2).

The *BRCA2* c.-117C>G variant is predicted to generate a potential new donor site, leading to a partial exon skipping of 77bp, with a delta score of 0.64 (reference score = 0.04, alternative score = 0.68). However, donor loss of the natural site is not predicted (delta score 0.02, reference score = 0.99, alternative score = 0.97) and the score for the natural donor site in the presence of the variant outweighs that of the new potential site. This suggests that while alternative splicing may occur, it would likely be at a lower frequency compared to the use of the natural splice site. Similarly, the *BRIP1* c.-232G>A variant is also predicted to generate a donor gain with a delta score of 0.39 (reference = 0.01, alternative = 0.40), leading to a partial exon skipping of 205bp As with *BRCA2*, the natural donor site has a predicted probability clearly higher than the new one (delta score = 0.04, reference = 0.88, alternative = 0.84).

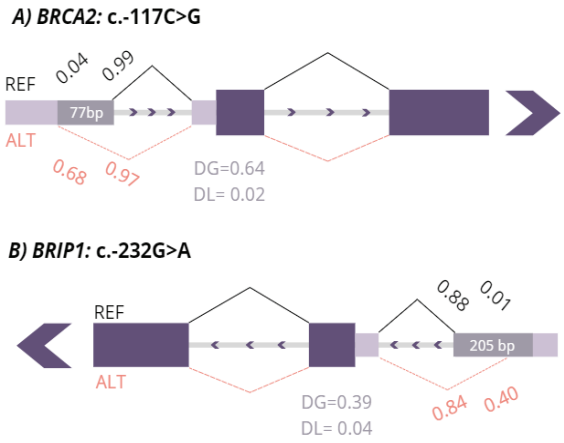


Figure 2: Schematic representation of the N-terminus of genes with candidate variants prioritized by SpliceAI, with splicing aberrations predicted by SpliceAI-10k and found exclusively in samples from patients whose phenotype was related to the gene. In the upper part of each panel, the splicing pattern of the full-length isoform is shown, along with its raw scores for the reference allele. In the lower part, the alternative raw scores for the natural site and the donor gain site caused by the variant are displayed. Both variants cause a partial exon skipping, represented in dark gray, with the lost region size indicated. DG (Donor Gain) and DL (Donor Loss) delta scores reflect the likelihood of gaining or losing donor splice sites, respectively.

Table 4: Variants selected according to SpliceAI delta scores. The four delta scores provided by SpliceAI are shown, along with the SpliceAI-10K splicing aberration prediction. Additionally, the table indicates the minor allele frequency (MAF) in GnomAD v.4.1.0, the number of carrier patients whose clinical suspicion was related to the variant gene, and the number of carrier patients where it was not. Variants selected for further exploration and discussed in the text are highlighted in light gray.

Variant description		**SpliceAI**				**SpliceAI-10K**	**GnomAD (v4.1.0)**	**Nº of individuals with clinical phenotype***	
Gene	**Variant**	**AG**	**AL**	**DG**	**DL**	**Splicing aberration?**	**MAF (%)**	**Yes**	**No**
ATM	c.-31+61A>G	0.02	0	**0.65**	0	PIR	0.5960 (958/160736)	15	13
BMPR1A	c.-287A>G	0.01	0	**0.79**	0.13	PED	0.0388 (59/151976)	0	1
BRCA2	c.-117C>G	0	0	**0.64**	0.02	PED	0.0007 (1/152362)	1	0
BRIP1	c.-31+6T>C	0.03	0	**0.36**	**0.39**	PED	0	0	1
BRIP1	c.-232G>A	0.04	0	**0.39**	0.04	PED	0	1	0
CHEK2	c.-7+286A>T	0.03	0	**0.26**	0.01	NCP	2.011 (3062/152228)	4	0
FLCN	c.-24-394A>G **	0.04	0	**0.48**	0.11	NCP	51.50 (119706/232442)	6	2
MEN1	c.-23-135G>A	0	0	**0.36**	0	No	0.0043 (27/627826)	0	4
MUTYH	c.-7+5C>G	0.02	0	**0.63**	**0.31**	No	98.76 (1384292/1401720)	498	1433
PTEN	c.-714G>A	0.03	0	**0.28**	0	PED	0.0095 (37/390758)	0	2
TP53	c.-28-82G>A	**0.60**	0	0.02	0	PIR & NCP	0.0094 (88/934420)	0	1

* The clinical indication is based on guidelines in the Catalan Health Instruction (Supplementary File 1)
** This region is not well covered by our panel (0.14% of individuals have at least 10 reads covering the position), and we are likely underestimating the number of individuals with the variant in our cohort.
Abbreviations: PIR: partial intron retention; PED: partial exon deletion; NCP: non-coding pseudoexon.

Table 5: Variants selected for -3 position relative to A in the main AUG. The translation efficiency of the main ORF is shown, where "REF" indicates the efficiency in the reference sequence (without the variant) and "ALT" indicates the efficiency with the variant. Additionally, the table indicates the minor allele frequency (MAF) in GnomAD v.4.1.0, the number of carrier patients whose clinical suspicion was related to the variant gene, and the number of carrier patients where it was not.

Variant description		Predicted Translation efficiency main ORF		GnomAD (v.4.1.0)	Nº of individuals with clinical phenotype*	
Gene	**Variant**	**REF**	**ALT**	**MAF (%)**	**Yes**	**No**
BRCA1	c.-3G>T	74	60	0,0001 (1/1610582)	1	0

* The clinical indication is based on guidelines in the Catalan Health Instruction (Supplementary File 1)

Variants at position -3 relative to the main ORF

The only variant found 3 bp upstream the main ORF, where an A/G changes to C/T, is *BRCA1* c.-3G>T (Table 5). This variant is predicted to reduce the translation efficiency of the main ORF. It was observed in a patient diagnosed with triple-negative medullary breast cancer at the age of 52. Her sister had also breast cancer at the age of 59. Their mother was diagnosed with Leukemia at the age of 76.

DISCUSSION

UTR regions are crucial for post-transcriptional regulation of gene expression. However, they are often overlooked in diagnostics settings due to the lack of clear guidelines for their classification.

To address this gap, experts have proposed filters to prioritize promising variants for further study, while maintaining a manageable number of candidates for classification (Martin-Geary et al., 2023). This is particularly important, given that variant classification remains a major bottleneck in molecular diagnostics units.

Using their approach, we prioritized seven variants Four were selected by UTRannotator. Of these, only the *CDKN2A* c.-34G>T variant, which had already been described in the literature (Liu *et al.*, 1999), is definitively pathogenic. This finding has enabled the adaptation of clinical follow-up for the three probands and the initiation of cascade testing in their relatives. However, according to their phenotype, the study of *CDKN2*A was not indicated for two of them. Nevertheless, one of them has an aunt with an unconfirmed diagnosis of skin cancer, which should be further investigated to assess whether *CDKN2A* variant could be the cause, while the other is clearly an incidental finding. The *MSH6* c.-22T>G variant, although it matches the family's clinical context, further analysis revealed that it does not co-segregate with the disease in the proband's mother. Moreover, the IHC showed normal expression of the MSH6 protein, suggesting that this variant is not the cause of the family's condition. The remaining two variants, *STK11* c.-33C>A and *AIP* c.-76T>C, have an unknown effect in the absence of functional follow-up. The first variant predicts an in-frame oORF, resulting in a NTE of eleven amino acids. Although the patient's polyps do not appear to be hamartomatous (typically associated with *STK11*-related conditions), we have limited clinical information, which prevents us from definitively excluding the presence of these polyps. The *AIP* variant is unrelated to the patient's phenotype, as AIP is typically linked to pituitary tumors, while the patient has prostate cancer.

We prioritized two variants based on potential splicing effects due to their position; however, neither appeared to have a significant impact on splicing after further investigation. Although they are predicted to create new donor sites, neither is expected to surpass the strength of the natural donor sites, suggesting that any alternative splicing events would likely occur at a lower frequency compared to the full-length isoform. While *in silico* tools like SpliceAI are highly predictive, they do not always accurately reflect biological outcomes, which are influenced by many other factors. Functional

studies, such as mini gene assays or RNA-seq, would be necessary to confirm the putative splicing effects.

We identified one variant at a -3 position in *BRCA1* gene, which is predicted to slightly reduce translation efficiency; although experimental validation, such as luciferase assays, would be needed to confirm this reduction. Nonetheless, RNA studies on other *BRCA1* variants suggest that even with only 20-30% of BRCA1 tumor suppressor function is sufficient to avoid a high risk of cancer (de la Hoya *et al.*, 2016). Therefore, it is unlikely that this variant is the cause of the family's condition.

Here, we applied a strict filtering approach, and it is important to consider that the restrictiveness of the filters used may inadvertently exclude other potentially relevant variants. For example, while UTRannotator is a powerful tool, it does not consider non-canonical start codons, which have been shown to initiate translation with considerable efficiency under certain Kozak sequence contexts (De Arce, Noderer, and Wang, 2018). Additionally, it does not consider variants that may modify the strength of the Kozak sequence of existing uORFs which could potentially impact gene expression.

This study has some limitations. First, although we achieved full exonic coverage of the 5'UTR region for 35 genes, for 19 genes we only captured approximately 150 bp upstream of the start codon, and for one gene, 500bp, in both cases, this coverage is insufficient to encompass the entire exonic 5'UTR (Supplementary Table 1). Therefore, we may have missed relevant variants in uncovered regions. Future versions of our panel will hopefully cover the whole 5'UTR of these genes.

Secondly, we studied single nucleotide variants and short indels, without considering copy number variants in the 5'UTR regions. They should be included to provide a more comprehensive understanding of the regulatory impact of UTR variants. For example, deletions affecting the first exon and promoter of the *MEF2C* 5'UTR are predicted to disrupt enhancer function and have been identified in patients with developmental disorders (Wright *et al.*, 2021). Additionally, our focus was limited to variants in the 5'UTR regions, but future studies should also consider variants in 3'UTR, internal ribosome entry sites and transcription factor binding sites, which can also play key regulatory roles.

Finally, this study represents a prioritization approach proposal, and as mentioned earlier, experimental validation is necessary to confirm the functional impact of the identified variants.

In summary, our study highlights the potential of analyzing 5'UTR regions to uncover clinically relevant variants in hereditary cancer. The use of strict prioritization filters demonstrates that incorporating these regions into diagnostic workflows can be feasible, paving the way for more comprehensive genetic analyses in the future.

REFERENCES

Bashirullah, A. *et al.* (2001) Spatial and temporal control of RNA stability. Proc Natl Acad Sci U S A, 98(13).

Calvo, S.E. *et al.* (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A, 106(18), p. 7507.

Canson, D.M. *et al.* (2023) SpliceAI-10k calculator for the prediction of pseudoexonization, intron retention, and exon deletion. Bioinformatics. Edited by C. Kendziorski, pp. 0–0.

Chen, S. *et al.* (2023) A genomic mutational constraint map using variation in 76,156 human genomes. Nature 2023 625:7993, 625(7993), pp. 92–100.

Filatova, A.Y. *et al.* (2021) Upstream ORF frameshift variants in the PAX6 5'UTR cause congenital aniridia. Hum Mutat, 42(8), pp. 1053–1065.

Jaganathan, K. *et al.* (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), pp. 535-548.e24.

Jansen, R.P. (2001) mRNA localization: Message on the move. Nat Rev Mol Cell Biol.

Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell, 44(2), pp. 283–292.

Kozak, M. (1987a) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res, 15(20), pp. 8125–8148.

Kozak, M. (1987b) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol, 196(4), pp. 947–950.

Kozak, M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J, 16(9), pp. 2482–2492.

de la Hoya, M. *et al.* (2016) Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Hum Mol Genet, 25(11).

Liu, L. *et al.* (1999) Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. Nature Genetics 1999 21:1, 21(1), pp. 128–132.

Martin-Geary, Alexandra C. *et al.* (2023) Systematic identification of disease-causing promoter and untranslated region variants in 8,040 undiagnosed individuals with rare disease. medRxiv [Preprint].

Martin-Geary, Alexandra C *et al.* (2023) variants in 8 , 040 undiagnosed individuals with rare disease Abstract Background, pp. 1–39.

McLaren, W. *et al.* (2016) The Ensembl Variant Effect Predictor. Genome Biol, 17(1), pp. 1–14.

Mignone, F. *et al.* (2002) Untranslated regions of mRNAs. Genome Biol.

Mohan, R.A. *et al.* (2014) A mutation in the Kozak sequence of GATA4 hampers translation in a family with atrial septal defects. Am J Med Genet A, 164(11).

Morris, D.R. and Geballe, A.P. (2000) Upstream Open Reading Frames as Regulators of mRNA Translation, 20(23), pp. 8635–8642.

Noderer, W.L. *et al.* (2014) Quantitative analysis of mammalian translation initiation sites by FACS -seq . Mol Syst Biol, 10(8), p. 748.

Pesole, G. *et al.* (2001) Structural and functional features of eukaryotic mRNA untranslated regions. Gene, 276(1–2), pp. 73–81.

Van Der Velden, A.W. and Thomas, A.A.M. (1999) The role of the 5' untranslated region of an mRNA in translation regulation during development. International Journal of Biochemistry and Cell Biology.

Whiffin, N. *et al.* (2020) Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun, 11(1).

Wright, C.F. *et al.* (2021) Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. Am J Hum Genet, 108(6), pp. 1083–1094.

Zhang, X. *et al.* (2021) Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics, 37(8), pp. 1171–1173.

Supplementary File 1:

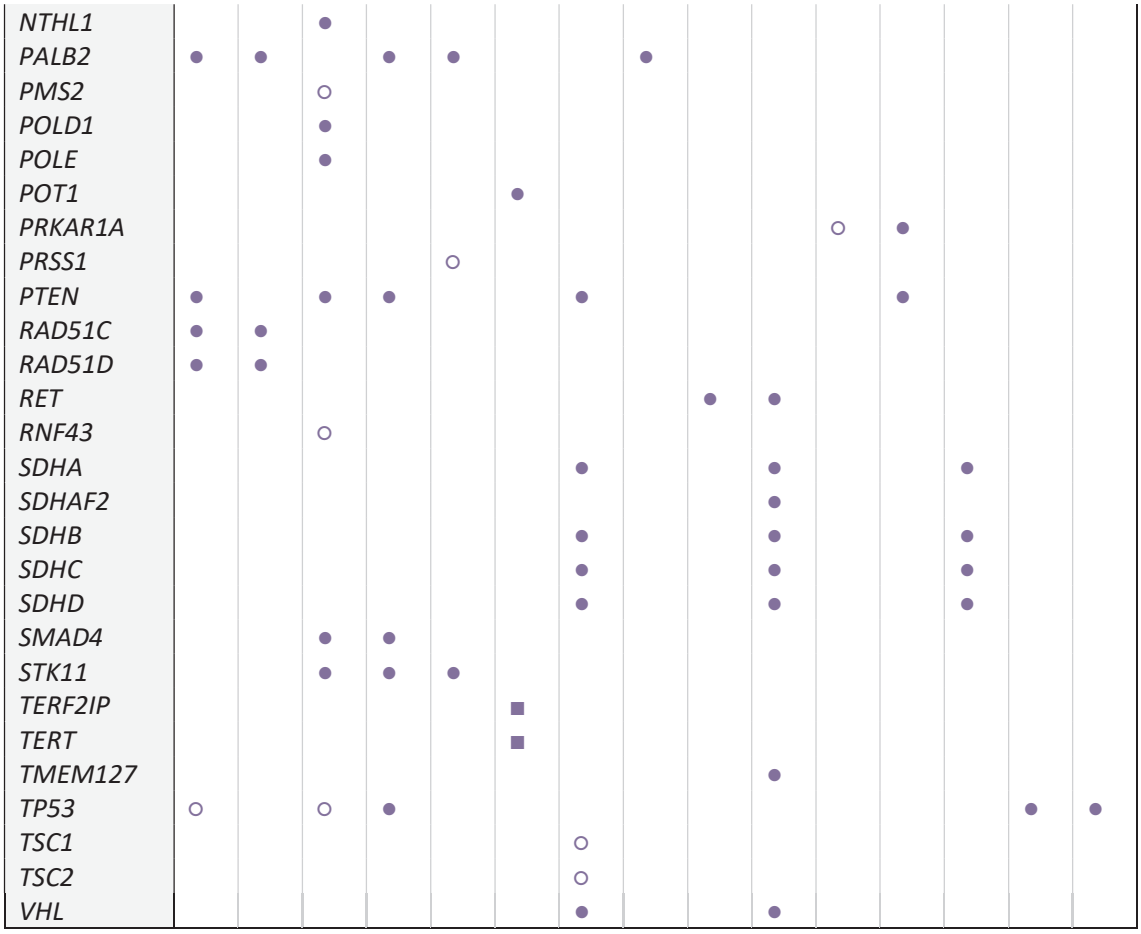
Genes and Clinical Criteria for Hereditary Cancer Predisposition Syndromes

Note: This document details only the syndromes relevant to the patients in our study. For a comprehensive list of hereditary cancer predisposition syndromes, please refer to the full Catalan Health Instruction document:

https://scientiasalut.gencat.cat/bitstream/handle/11351/8438.3/determinacions_perfil_genetic_sindromes_hereditaries_cancer_adult_pediatria_2023.pdf?sequence=4&isAllowed=y

Genes to be analyzed by phenotype: *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, and *MSH6* are always analyzed regardless of phenotype due to their high clinical relevance (highlighted in light purple). Figures shaded in purple represent genes that must always be analyzed for that phenotype, while those shaded in white indicate genes that are only analyzed if the proband meets additional phenotypic conditions. Circles represent genes included in the I2HCP v3 panel, whereas squares indicate genes that are not yet included in the panel but will be added in future versions.

Genes	Phenotypes														
	HBOC	OV	CRC	GAS	PAN	MEL	REN	PR	MTC	PPGL	HP	TS	GIST	AC	TP53
ACD						■									
AIP											○				
APC			●									●		●	
ATM	●			●	●			●							
BAP1						●	●								
BARD1	●														
BMPR1A			●	●											
BRCA1	●	●			●			●							
BRCA2	●	●			●	●		●							
BRIP1	●	●													
CaSR											□				
CDC73											●				
CDH1	○			●											
CDK4						●									
CDKN1B											●				
CDKN2A					●	●									
CHEK2	●							●							
CTNNA1				●											
DICER1												●			
EPCAM			●												
FH							●			●				●	
FLCN							●								
HOXB13								●							
KIF18										●					
KIT													●		
MAX										●					
MEN1										●					
MEN1											●				
MET							●								
MITF						●									
MLH1		●	●	●	●			●							
MSH2		●	●	●	●			●							
MSH6		●	●	●	●			●							
MUTYH			●												
NF1										○					



Abbreviations of syndromes: HBOC: Hereditary Breast and Ovarian Cancer; OV: Ovarian cancer; CRC: Colorectal and endometrial cancer and polyposis; GAS: Gastric cancer; PAN: Pancreatic cancer; MEL: Melanoma; REN: Renal cancer; PR: Prostate cancer; MTC: Medullary thyroid carcinoma; PPGL: Pheochromocytoma/paraganglioma; HP: Primary hyperparathyroidism or manifestations of MEN1/MEN4; TS: Syndromic thyroid cancer; GIST: Gastrointestinal stromal tumor; AC: Adrenocortical tumor; TP53: Hereditary cancer syndromes related to *TP53*.

Clinical Criteria:

<u>i. Hereditary Breast and Ovarian Cancer (HBOC)</u>

- Breast cancer at age ≤ 40.
- Breast cancer at age ≤ 50 in cases of non-informative family history.
- Triple-negative breast cancer at age ≤ 60.
- Male breast cancer.
- Three or more first- or second-degree relatives with breast cancer (at least one ≤ 60 years).
- Three or more cases of breast, pancreatic, ovarian, and/or prostate cancer (Gleason score ≥ 7).
- Two cases of breast cancer at age ≤ 50.
- Bilateral breast cancer (first diagnosis at age ≤ 50).
- Bilateral breast cancer in one individual plus another case in the family (one diagnosis at age ≤ 60).
- HER2-negative breast cancer for patients eligible for PARP inhibitor treatment per CatSalut indications.
- Breast cancer with family history of ovarian cancer.
- Invasive non-mucinous epithelial ovarian cancer (consider age, family history, and potential benefit for relatives in low-grade cases).

ii. Colorectal and Endometrial Cancer Predisposition Syndromes

Screening Requirement: Lynch syndrome screening using immunohistochemistry (IHC) for DNA repair gene proteins and/or microsatellite instability (MSI) analysis should be performed on all colorectal and endometrial cancers.

- MSI or altered IHC (if MLH1/PMS2 loss is detected, rule out MLH1 promoter hypermethylation or BRAF mutation in the tumor).
- Colorectal cancer at age ≤ 50 or Amsterdam criteria met.

iii. Hereditary Polyposis Syndromes

Adenomatous Polyposis:

- Accumulation of ≥ 20 adenomas.
- 10-19 adenomas in cases meeting one of the following: age < 40, synchronous/metachronous CRC before age 60, or family history of adenomatous polyposis or CRC < 60 years.

Non-Adenomatous Polyposis:

- Syndrome-specific criteria (e.g., Cowden, Peutz-Jeghers).

Serrated Polyposis (WHO 2019):

- Age < 50 and/or
- At least one first-degree relative with serrated polyposis syndrome.

iv. Hereditary Prostate Cancer

- Metastatic prostate cancer with Gleason score ≥ 7 (either metastatic at diagnosis or as recurrence).
- Prostate cancer with Gleason score ≥ 7 and any of the following:
    - Age < 55 years,
    - Family history of breast and/or ovarian cancer, or
    - Two or more cases of prostate cancer in the same lineage.
- Prostate cancer diagnosed at < 55 years with family history of two or more cases of prostate cancer or HBOC.
- Prostate cancer with cribriform histological pattern (ductal or intraductal).
- Prostate cancer not meeting the above criteria but with indication for PARP inhibitors.

v. Familial Melanoma

- Two melanoma cases in first- or second-degree relatives, with at least one diagnosed before age 60.
- Multiple melanomas: two or more melanomas in an individual, with the first diagnosis before age 60.
- At least one melanoma case under age 60 plus a family member with pancreatic cancer.

Supplementary Table 1: Summary of the 5'UTR coverage for selected transcripts in the ICO-IMPPC Hereditary Cancer Panel (I2HCP, v3). For each transcript, chromosome location (chr, start, end) based on both GRCh37 and GRCh38 assemblies, transcript identifier, length of the 5'UTR region, and coverage status (UTR covered) are provided. "Y" indicates full 5'UTR coverage, "Y, but only exons" indicates all UTR exons and 150 bp of intronic boundaries covered while "150bp" indicates partial coverage up to 150 bp upstream of the translation start site.

chr	start_38	end_38	start_37	end_37	transcript	version	RefSeq Catalan Instruction	type	length_utr	UTR covered?
11	67483026	67483158	67250497	67250629	AIP|ENST00000279146	8	NM_003977.4	MANE_Select	133	Y
5	112737885	112754890	112073582	112090587	APC|ENST00000257430	9	NM_000038.6	MANE_Select	17006	Y, but only exons
11	108223067	108227624	108093794	108098351	ATM|ENST00000675843	1	NM_000051.4	MANE_Select	4558	Y, but only exons
3	52409879	52410008	52443895	52444024	BAP1|ENST00000460680	6	NM_004656.4	MANE_Select	130	Y
2	214809570	214809683	215674294	215674407	BARD1|ENST00000260947	9	NM_000465.4	MANE_Select	114	Y
10	86756619	86876018	88516376	88635775	BMPR1A|ENST00000372037	8	NM_004329.3	MANE_Select	119400	Y, but only exons
17	43124097	43125364	41276114	41277381	BRCA1|ENST00000357654	9	NM_007294.4	MANE_Select	1268	Y, but only exons
13	32315086	32316460	32889223	32890597	BRCA2|ENST00000544455	8	NM_000059.4	MANE_Select	1375	Y, but only exons
17	61861540	61863528	59938901	59940889	BRIP1|ENST00000259008	7	NM_032043.3	MANE_Select	1989	Y, but only exons
1	193122031	193122200	193091161	193091330	CDC73|ENST00000367435	5	NM_024529.5	MANE_Select	170	150bp
16	68737292	68737415	68771195	68771318	CDH1|ENST00000261769	10	NM_004360.5	MANE_Select	124	Y
12	57751718	57752310	58145501	58146093	CDK4|ENST00000257904	11	NM_000075.4	MANE_Select	593	150bp
12	12717368	12717839	12870302	12870773	CDKN1B|ENST00000228872	9	NM_004064.5	MANE_Select	472	150bp
9	21974828	21974857	21974827*	21974856*	CDKN2A|ENST00000304494	10	NM_000077.5	MANE_Select	30	Y
9	21994332	21994392	21994331	21994391	CDKN2A|ENST00000579755	2	NM_058195.4	Plus Clinical	61	Y
22	28734722	28741820	29130710	29137808	CHEK2|ENST00000404276	6	NM_007194.4	MANE_Select	7099	Y, but only exons
5	138753425	138781924	138089114	138117613	CTNNA1|ENST00000302763	12	NM_001903.5	MANE_Select	28500	150bp
14	95133459	95157529	95599796	95623866	DICER1|ENST00000343455	8	NM_177438.3	MANE_ Select	24071	150bp
2	47369311	47369505	47596450	47596644	EPCAM|ENST00000263735	9	NM_002354.3	MANE_Select	195	150bp
1	241519723	241519755	241683023	241683055	FH|ENST00000366560	4	NM_000143.4	MANE_Select	33	Y
17	17228138	17237168	17131452	17140482	FLCN|ENST00000285071	9	NM_144997.7	MANE_Select	9031	150bp
1	10210570	10232328	10270628	10292386	KIF1B|ENST00000676179	1	NM_001365951.3	MANE_Select	21759	150bp
14	65102340	65102517	65569058	65569235	MAX|ENST00000358664	9	NM_002382.5	MANE_Select	178	150bp
11	64810110	64810551	64577582	64578023	MEN1|ENST00000450708	7	NM_001370259.2	MANE_Select	442	150bp
7	116672196	116699084	116312250	116339138	MET|ENST00000397752	8	NM_000245.4	MANE_Select	26889	150bp
3	36993518	36993547	37035009	37035038	MLH1|ENST00000231790	8	NM_000249.4	MANE_Select	30	Y

2	47403156	47403191	47630295	47630330	MSH2|ENST00000233146	7	NM_000251.3	MANE_Select	36	Y
2	47783145	47783233	48010284	48010372	MSH6|ENST00000234420	11	NM_000179.3	MANE_Select	89	Y
1	45334506	45339970	45800178	45805642	MUTYH|ENST00000456914	7	NM_001048174.2	MANE_Select	5465	Y, but only exons
17	31094977	31095309	29421995	29422327	NF1|ENST00000358273	9	NM_001042492.3	MANE_Select	333	Y
16	2047834	2047866	2097835	2097867	NTHL1|ENST00000651570	2	NM_002528.7	MANE_Select	33	Y
16	23641158	23641310	23652479	23652631	PALB2|ENST00000261584	9	NM_024675.4	MANE_Select	153	Y
7	6009020	6009049	6048651	6048680	PMS2|ENST00000265849	12	NM_000535.7	MANE_Select	30	Y
19	50384347	50398851	50887604	50902108	POLD1|ENST00000440232	7	NM_002691.4	MANE_Select	14505	Y, but only exons
12	132687316	132687342	133263902	133263928	POLE|ENST00000320574	10	NM_006231.4	MANE_Select	27	Y
7	124897174	124929825	124537228	124569879	POT1|ENST00000357628	8	NM_015450.3	MANE_Select	32652	150bp
17	68512430	68515399	66508571	66511540	PRKAR1A|ENST00000589228	6	NM_002734.5	MANE_Select	2970	150bp
7	142749472	142749484	142457323	142457335	PRSS1|ENST00000311737	12	NM_002769.5	MANE_Select	13	Y
10	87863625	87864469	89623382	89624226	PTEN|ENST00000371953	8	NM_000314.8	MANE_Select	845	500bp
17	58692602	58692643	56769963	56770004	RAD51C|ENST00000337432	9	NM_058216.3	MANE_Select	42	Y
17	35119614	35119860	33446633	33446879	RAD51D|ENST00000345365	11	NM_002878.4	MANE_Select	247	Y
10	43077069	43077258	43572517	43572706	RET|ENST00000355710	8	NM_020975.6	MANE_Select	190	150bp
17	58415578	58417534	56492939	56494895	RNF43|ENST00000407977	7	NM_017763.6	MANE_Select	1957	150bp
5	218320	218355	218435	218470	SDHA|ENST00000264932	11	NM_004168.4	MANE_Select	36	Y
11	61430124	61430146	61197596	61197618	SDHAF2|ENST00000301761	7	NM_017841.4	MANE_Select	23	Y
1	17054020	17054032	17380515	17380527	SDHB|ENST00000375499	8	NM_003000.3	MANE_Select	13	Y
1	161314381	161314405	161284171	161284195	SDHC|ENST00000367975	7	NM_003001.5	MANE_Select	25	Y
11	112086873	112086907	111957597	111957631	SDHD|ENST00000375549	8	NM_003002.4	MANE_Select	35	Y
18	51030213	51047046	48556583	48573416	SMAD4|ENST00000342988	8	NM_005359.6	MANE_Select	16834	Y, but only exons
19	1205778	1206913	1205777	1206912	STK11|ENST00000326873	12	NM_000455.5	MANE_Select	1136	150bp
2	96265382	96265997	96931120	96931735	TMEM127|ENST00000258439	8	NM_017849.4	MANE_Select	616	150bp
17	7676595	7687490	7579913	7590808	TP53|ENST00000269305	9	NM_000546.6	MANE_Select	10896	Y, but only exons
9	132928873	132944616	135804260	135820003	TSC1|ENST00000298552	9	NM_000368.5	MANE_Select	15744	150bp
16	2047985	2048615	2097986	2098616	TSC2|ENST00000219476	9	NM_000548.5	MANE_Select	631	150bp
3	10141778	10141847	10183462	10183531	VHL|ENST00000256474	3	NM_000551.4	MANE_Select	70	Y

*CAGE data for CDKN2A support a longer isoform as considered in previous transcript versions, thus start_37 coordinate will be considered 21975097 (ENST00000304494.5). UTR is well covered.

RESUM
DE RESULTATS

Article publicat 1: Algoritme bioinformàtic de codi obert per optimitzar el diagnòstic de *PMS2* utilitzant dades d'NGS de lectura curta

Motivació

L'objectiu d'aquest article era optimitzar el diagnòstic genètic del gen *PMS2*. Fins ara, totes les mostres amb pèrdua exclusiva de PMS2 en el tumor, s'estudiaven experimentalment mitjançant PCR (3 LR-PCR i 15 PCR curtes per cobrir tot el gen, i posterior seqüenciació dels productes de totes les PCRs curtes). Aquesta tècnica és complicada, laboriosa i costosa, i en la majoria de casos no es detecten variants patogèniques, malgrat la pèrdua de PMS2 en el tumor. Per aquest motiu, es va plantejar l'objectiu de reduir el nombre de mostres que han de ser validades experimentalment.

Desenvolupament de l'algoritme

Per abordar aquesta problemàtica, es va desenvolupar un algoritme bioinformàtic com a cribratge inicial per identificar mostres amb possibles variants patogèniques candidates. Es consideren candidates perquè poden procedir tant de *PMS2* com de *PMS2CL*, de manera que cal posteriorment validar-les experimentalment per distingir la seva procedència. No obstant això, aquest enfocament permet limitar les PCR únicament a les regions de les variants identificades, fet que implica una reducció significativa del nombre de PCRs necessàries.

L'algoritme elimina la seqüència de *PMS2CL* del genoma de referència, forçant l'alineament de totes les lectures a *PMS2*. Per les mostres que no presenten pèrdua de PMS2 al tumor, es recomana aplicar l'estratègia proposada per Gould et al. (2018), segons la qual, per l'exó 11 – l'exó amb més PSVs- , només es consideren les lectures que intersequen posicions invariants específiques. Totes les variants identificades són analitzades mitjançant un algoritme de decisió que té en compte diversos factors com són la VAF, si és PSV, la localització i la classificació de la variant. Aquest flux de treball determina quines variants han de ser validades experimentalment.

Validació de l'algoritme

L'algoritme es va validar en una cohort de 40 pacients, on es va demostrar un augment de la sensibilitat (de 0.853 a 0.956) respecte a l'anàlisi bioinformàtica utilitzada per la rutina diagnòstica habitual.

Prevalença de variants patogèniques a *PMS2* en una cohort de pacients amb càncer hereditari

L'algoritme desenvolupat es va utilitzar per reanalitzar les dade NGS de 5.619 pacients amb càncer hereditari. Aquesta cohort incloïa 13 pacients amb sospita de LS i pèrdua exclusiva de l'expressió de PMS2 en el tumor (sense haver estat prèviament analitzats amb LR-PCR), 36 pacients diagnosticats amb càncer colorectal d'aparició primerenca (<50 anys) amb expressió conservada de les proteïnes MMR o sense informació d'imunohistoqúimica (IHQ), 798 pacients amb altres criteris de sospita de LS (criteris d'Amsterdam o pèrdua d'expressió MMR no exclusiva de PMS2), i 4772 pacients analitzats per sospita d'altres síndromes de càncer hereditari. Es van identificar 16 variants patogèniques candidates. D'aquestes, 15 es van confirmar experimentalment com a variants patogèniques de *PMS2*: Cinc pacients presentaven tumors amb pèrdua exclusiva de PMS2 detectada amb IHQ (38,462%, 5/13), cinc complien altres criteris de sospita de LS (quatre amb pèrdua de PMS2 i MSH6, i un amb pèrdua de PMS2 i MLH1) (0,627%, 5/798), i cinc es van analitzar per altres sospites de càncer hereditari (un tumor d'ovari amb expressió heterogènia de PMS2 i quatre tumors amb expressió MMR conservada) (0,105%, 5/4.772).

Article publicat 2: Detecció de CNVs germinals a partir de dades de panells de gens: comparativa de les eines actuals

Motivació

L'objectiu d'aquest article era avaluar les eines existents per a la identificació de CNVs a partir de dades de panells de gens, amb la finalitat que els resultats serveixin com a referència tant en contextos de diagnòstic com de recerca. Aquest treball amplia l'avaluació realitzada pel nostre grup a l'article "*Evaluation of CNV detection tools for NGS panel data in genetic diagnostics*", considerant les noves eines sorgides des del 2018 fins 2024, i incloent també les eines seleccionades en l'estudi anterior. A més, s'ha analitzat l'impacte de la modificació de 107 paràmetres en el rendiment de les eines i s'han avaluat combinacions d'eines per parelles.

Desenvolupament de l'algoritme

Es va fer una revisió bibliogràfica i es van escollir 12 eines: clearCNV v0.306, GATK-gCNV v4.5.0, Atlas-CNV v.0, Cobalt v0.8.0, ClinCNV v1.18.3, CNVkit v0.9.10, VisCap v0.8, DeCoN v2.0.1, panelcn.MOPS v1.20.0, ExomeDepth v1.1.16, CoNVaDING v1.2.1 i CODEX2 v1.3.0. La selecció es va basar en el següents criteris:

- Codi obert.
- Capaces d'identificar CNVs germinals.
- Dissenyades per analitzar panells de gens.
- No dissenyades per amplicons.

Les eines es van avaluar utilitzant 4 conjunts de dades de panells validats, emprant dues mètriques (a nivell de gen i de regió d'interès) i en tres contexts diferents: avaluació utilitzant els paràmetres per defecte, un estudi de com influencia la modificació dels diferents paràmetres en el resultat final i una combinació de les eines per parelles.

Per dur a terme l'avaluació es va desenvolupar CNVbenchmarker2, un codi d'accés lliure que permet als usuaris realitzar les seves pròpies avaluacions amb les seves dades.

Avaluació amb els paràmetres per defecte

Considerant l'F1 score, ClinCNV i GATK-gCNV van mostrar un millor rendiment a nivell de regió d'interès en comparació amb la resta d'eines. A nivell de gen, GATK-gCNV va destacar com a la millor eina.

Pel que fa a la sensibilitat a nivell de gen, les eines CoNVaDING, GATK-gCNV, CODEX2 i DeCoN van ser les més destacades. D'una banda, CoNVaDING només va tenir tres falsos negatius en els quatre conjunts de dades. D'altra banda, GATK-gCNV, tot i tenir més falsos negatius, va mostrar una millor relació entre sensibilitat i especificitat.

Avaluació dels paràmetres

Es va demostrar que alguns els paràmetres poden afectar en gran mesura el rendiment, i es van identificar quatre patrons principals:

- Sense efecte en el rendiment.
- Compromís classificadors binaris: un augment de la sensibilitat comporta una disminució de l'especificitat, i viceversa.
- Forma de campana: el rendiment de l'eina pot ser optimitzat al voltant d'un valor determinat.
- Efectes en el rendiment sense un patró clar.

Es va suggerir provar 13 paràmetres per intentar optimitzar el valor de F1 de les eines, recomanant als usuaris que ho avaluessin en els seus propis conjunts de dades.

Combinació per parelles

L'avaluació de la unió de les parelles va demostrar que cap combinatòria va detectar tots els vertaders positius per regió d'interès, però en canvi sí que cinc parelles van detectar-ho a nivell de gen; Atlas-CNV/CoNVaDING, CODEX2/CoNVaDING, CNVkit/CoNVaDING, panelcn.MOPS/CoNVaDING and DECoN/CODEX2.

L'avaluació de la intersecció va mostrar que la combinació de CODEX2 amb GATK-gCNV o DeCoN va identificar totes les CNVs en un dels panells a nivell de gen, sense generar falsos positius.

Article publicat 3: vaRHC: un paquet de R per semi-automatitzar la classificació de variants en els gens de càncer hereditari considerant les guies ACMG/AMP i les guies específiques de gen de Clingen

Motivació

L'objectiu d'aquest article era proporcionar una eina per optimitzar la classificació de variants mitjançant l'automatització del màxim nombre d'evidències possibles. Aquest enfocament permet als curadors de variants concentrar els seus esforços en aquells aspectes que realment requereixen la seva expertesa, agilitant el procés i reduint el coll d'ampolla associat a la classificació de variants en les unitats de diagnòstic molecular.

Desenvolupament de l'algoritme

En primer lloc, es van analitzar els criteris establerts en les guies actualitzades de l'ACMG/AMP per determinar quins podien ser completament automatitzats, quins eren parcialment automatitzables i quins requerien una intervenció manual.

A partir d'aquesta anàlisi, es va desenvolupar vaRHC, un paquet d'R que extreu dades de les principals fonts d'informació utilitzades per classificar variants i integrar aquesta informació per assignar o descartar els criteris que poden ser parcialment o totalment automatitzables. A més, pels gens *ATM*, *CDH1*, *CHEK2*, *MLH1*, *MSH2*, *MHS6*, *PMS2*, *PTEN* i T*P53,* es van aplicar les guies específiques de l'ACMG.

Per combinar aquests criteris, es va utilitzar la metaestructura bayesiana proposada per Tavtigian, així com les recomanacions del grup CanVig-UK.

Validació de l'algoritme

El paquet es va validar utilitzant 659 variants classificades per panells d'experts de ClinGen, La validació es va realitzar criteri per criteri per evitar qualsevol biaix degut a la impossibilitat d'automatitzar certs criteris, assegurant així una avaluació precisa del rendiment de l'algoritme en cadascun dels criteris automatitzables. A més, es va incloure com a material suplementari la llista completa de les variants utilitzades, així com una especificació dels casos en què l'assignació no es corresponia amb la de ClinGen.

Els resultats van mostrar una concordança elevada entre les classificacions automàtiques de vaRHC i les classificacions fetes pels experts de ClinGen, sent del 97% per aquelles evidències que poden ser totalment automatitzades.

Comparació amb CancerSIGVAR

Pels gens *CDH1* i *PTEN* es van comparar també els resultats amb els de Cancer SIGVAR, una eina que també considera les guies ClinGen específiques per a aquests dos gens. Aquesta comparació va demostrar que vaRHC oferia una millora significativa en la classificació de variants.

Article no publicat 1: Creació d'un codi per la priorització de variants estructurals clínicament rellevants basat en GRIDSS

Motivació:

L'objectiu d'aquest article era identificar SVs en les regions codificants i adjacents dels gens d'alt i moderat risc que podrien estar passant desapercebudes amb els *pipelines* diagnòstics actuals, que utilitzen exclusivament el mètode de profunditat de cobertura. A més, es buscava establir una estratègia, mitjançant l'ús de filtres, que faci viable la detecció de variants estructurals en el context de la pràctica diagnòstica.

Desenvolupament de l'algoritme

En aquest estudi es van incloure 9.750 mostres de pacients amb sospita de CH. Es va utilitzar el programari GRIDSS, que combina lectures dividida (SR), mapeig d'extrems aparellats (PEM) i assemblatge per identificar variants estructurals, detectant inicialment 1.344.324 SVs. Per prioritzar-les, es va desenvolupar en R un *pipeline*, amb filtres per excloure variants recurrents i per descartar probables polimorfismes o falsos positius. A més, es va centrar l'anàlisi en les regions clínicament rellevants de gens accionables i es van eliminar altres artefactes d'alineament.

Selecció de variants

Els filtres van retenir 84 variants, 30 de les quals ja havien estat prèviament detectades per rutina diagnòstica i, per tant, no es van considerar. De les 54 restants, 30 es van descartar després d'inspecció visual amb l'IGV (integrative genome viewer) per falta d'evidències de SVs. La inspecció consistia a analitzar els patrons de cobertura per detectar regions amb una profunditat de lectures anormal, examinar l'orientació de les lectures i també identificar bases *soft-clipped* que delimiten potencials punts de trencament. De les 24 variants restants, només 9 estaven en regions codificants i adjacents (+/- 20bp) de gens clínicament accionables i es van validar experimentalment.

Així es van identificar dues duplicacions *frameshift* patogèniques a *MSH6* en dos pacients amb pèrdua d'expressió de proteïna MSH6 al tumor, una deleció probablement patogènica a *BARD1* en una pacient amb càncer de mama, que afectava el donador canònic de l'exó 9 i causava skipping dels exons 8 i 9, activant el NMD. També es va trobar una duplicació en pauta de *PALB2* en una pacient amb càncer de mama (classificada com a VSD). A més, es van trobar cinc insercions de transposons en regió codificant: quatre *Alus* (a una a *PALB2*, una a *ATM* i dues a *BRCA2*) i un *LINE* a *APC*, totes relacionades amb els fenotips dels pacients. De moment aquestes variants s'han considerat probablement patogèniques, a falta de completar encara la seva classificació final.

Article no publicat 2: Identificació de variants potencialment patogèniques a les regions 5'UTR en una cohort de CH

Motivació:

L'objectiu d'aquest estudi era identificar possibles variants patogèniques dins de les 5'UTRs.

Detecció de variants

En aquest estudi es van reanalitzar 4485 mostres de pacients amb sospita de CH, centrant-se en les regions 5'UTR dels 55 gens clínicament rellevants (segons la Instrucció del CatSalut: https://scientiasalut.gencat.cat/bitstream/handle/11351/8438.3/determinacions_perfil_genetic_sindromes_hereditaries_cancer_adult_pediatria_2023.pdf?sequence=4&isAllowed=y) i inclosos en el panell I2HCP v3. En aquestes regions, que no s'analitzen per rutina, es van identificar 86.694 variants, de les quals 860 eren úniques. Es van analitzar mitjançant Ensembl's VEP, amb UTRannotator i SpliceAI habilitats, i considerant únicament les variants en els transcrits especificats en la Instrucció del CatSalut .

Priorització de variants

Es van aplicar els filtres descrits a Martin-Geary et al., (2023), que prioritzen les variants segons diversos criteris: la creació o destrucció d'oORFs amb potenical d'interferir en l'expressió de l'ORF principal o d'allargar-lo (segons prediccions d'UTRannotator), l'impacte en l'empalmament (predit per SpliceAi) i les modificacions en la seqüència de Kozak de l'ORF principal. Aquests filtres van reduir les variants inicials a setze.

Quatre van ser seleccionades per les anotacions d'UTRannotator. Dues d'aquestes generen un codó d'inici aigües amunt (uAUG) creant un uoORF fora de pauta: la variant *CDKN2A* c.-34G>T, en un context de Kozak fort, descrita a la literatura com a patogènica, i la variant *MSH6 c.*-22T>G, en context moderat, que no mostra pèrdua de *MSH6* a la IHQ del tumor del pacient ni co-segrega amb la malaltia en la mare afecta. Una altra variant, *STK11* c.-33C>A, crea un uAUG que genera un uoORF en pauta amb la regió codificant, potencialment allargant l'extrem N-terminal de la proteïna. Malgrat que el pacient presenta poliposi, no hi ha indicis de pòlips hamartomatosos, típics en deficiències associades a *STK11.* Finalment, la variant *AIP* c.-76T>C elimina un codó de terminació d'un uORF, sense cap altre codó de terminació abans de la metionina inicial de l'ORF principal, generant un oORF fora de pauta (Kozak fort). Així i tot, l'anàlisi d'AIP no està indicat en aquest pacient, diagnosticat amb càncer de pròstata.

D'altra banda es van seleccionar onze variants com a candidates a alterar l'empalmament, per tenir un valor de delta score de SpliceAI ≥ 0.2. D'entre aquestes, *BRCA2* c.-117C>G i *BRIP1* c.-232G>A es van trobar exclusivament en pacients amb fenotip relacionat amb el gen. No obstant això, en ambdós casos, els llocs naturals mantenien puntuacions més altes, suggerint que qualsevol esdeveniment d'empalmament alternatiu seria de menor freqüència.

També es va identificar una variant al lloc -3 respecte de l'ORF principal en el gen *BRCA1*, que podria reduir l'eficiència de la traducció de l'ORF principal.

DISCUSSIÓ

Millora del rendiment diagnòstic a partir de l'optimització de dades de panell

Contextualització dels estudis genètics assistencials en càncer hereditari

Tot i que el càncer té sempre un origen genètic, només entre un 5% i un 10% dels casos es deuen a variants genètiques constitucionals que poden passar a la descendència (J.E. and K., 2005; Jahn *et al.*, 2022). En aquests casos, anomenats càncer hereditari, el que s'hereta no és el càncer en si mateix, sinó una predisposició a desenvolupar-lo. Aquestes famílies solen tenir una alta agregació de tumors específics, diagnosticats en edats primerenques, amb individus amb múltiples diagnòstics i, en alguns casos tumors amb histologia poc freqüent. La correcta identificació d'aquestes famílies amb càncer hereditari és clau. Per aquest motiu, a Catalunya s'han establert guies de consens que harmonitzen els criteris de selecció i els gens a analitzar segons el fenotip familiar (https://scientiasalut.gencat.cat/handle/11351/8438.3).

En aquest context, els panells multigen que s'utilitzen al SDMCH de l'ICO inclouen 165 gens, però pel diagnòstic només s'analitzen aquells que es relacionen amb el fenotip de la família i tenen una accionabilitat clínica establerta en les guies clíniques de consens internacional. La identificació de la causa genètica de predisposició al càncer no només és rellevant pel pacient, sinó també pels seus familiars directes, a qui se'ls ofereix la possibilitat de realitzar un estudi de la variant identificada a la família. En cas que la tinguin, se'ls proporcionen mesures de prevenció personalitzades; en cas contrari, si es valora que el seu risc és el mateix que el de la població general, se segueixen els protocols de cribratge poblacional.

Ara bé, avui dia només aconseguim identificar la causa genètica en aproximadament un 11% de les famílies, tot i que aquest percentatge varia en funció del fenotip (Taula 15) i dels criteris d'inclusió de cada període. Aquest fet deixa un nombre considerable de famílies sense una explicació genètica, cosa que obliga a proporcionar recomanacions als pacients i als seus familiars de primer grau basades exclusivament en la història personal i familiar.

Taula 15: Taxa de detecció de variants (probablement) patogèniques a l'ICO durant els 3 últims anys, desglossada pels principals panells sol·licitats

Panell	**2021**	**2022**	**2023**
CMOH	12,8% (81/632)	9,7% (89/912)	8,8% (82/934)
Poliposis	13,1% (10/76)	17,4% (15/86)	17,9% (20/112)
Síndrome de Lynch	15,9% (18/113)	14,2% (32/225)	17,5% (39/223)
Qualsevol panell	**13,3% (109/821)**	**11,1% (136/1223)**	**11,1% (141/1269)**

Limitacions actuals en la detecció de variants

Diversos factors poden limitar la capacitat per identificar la causa de la predisposició genètica al càncer. Una part dels casos no resolts es pot atribuir al fet que el nostre coneixement sobre els gens implicats en el càncer hereditari encara és limitat. Tot i que s'han identificat molts gens d'alta i moderada penetrància, encara es continuen descobrint nous gens candidats associats a diferents condicions clíniques. A més, la contribució de múltiples variants de risc baix acumulades en un individu podria augmentar la susceptibilitat sense que una única variant patogènica sigui evident, com s'ha demostrat en els estudis de *Polygenic Risk Scores*, per exemple en el de Mavaddat et al. (2019).

Un altre factor important són les limitacions associades als algoritmes diagnòstics utilitzats. Actualment, els panells diagnòstics solen emprar tecnologia NGS de lectures curtes, una tècnica que tot i haver revolucionat el diagnòstic genètic en permetre l'estudi simultani de múltiples gens, presenta certes limitacions en la detecció de certs tipus de variants (Lincoln *et al.*, 2021), com es recull a l'apartat sis de la introducció.

Una altra limitació és que la majoria d'estudis solen analitzar només les regions codificants i les bases intròniques adjacents. En concret, al SDMCH de l'ICO, s'analitza la regió codificant i les 20 bases adjacents, seguint les indicacions establertes pel Programa d'Oncologia de Precisió del CatSalut. Aquesta limitació es deu principalment al disseny del panell, centrat en les regions comentades per tal d'optimitzar recursos i facilitar la interpretació de resultats donada la manca de guies clares per interpretar i classificar variants en aquestes regions. No obstant això, s'ha descrit que aquestes regions no analitzades, poden tenir un impacte crucial en la regulació post-transcripcional i que variants en les mateixes poden desregular el gen en qüestió (Liu et al., 1999; Calvo, Pagliarini and Mootha, 2009; Schulz et al., 2018). Per aquest motiu, s'estan realitzant esforços per millorar la comprensió de l'efecte de les variants genètiques en aquestes regions i el seu impacte clínic (Ellingford *et al.*, 2022).

Per últim, en alguns casos l'agregació de càncers en una família o l'aparició de tumors en edats primerenques poden ser resultat de l'atzar, degut a la naturalesa parcialment estocàstica de les mutacions i la seva reparació.

Aquesta tesi s'ha centrat en l'optimització de solucions bioinformàtiques per maximitzar el rendiment diagnòstic a partir de les dades de panells de gens amb lectures curtes, la tecnologia actualment disponible al SDMCH de l'ICO.

Impacte dels pseudogens en l'anàlisi de variants amb NGS de lectures curtes

Un dels principals reptes de la seqüenciació de lectures curtes és la identificació precisa de variants en gens que comparteixen alta homologia amb els seus pseudogens (Lincoln *et al.*, 2021). Sovint, les lectures no s'estenen més enllà de les regions d'homologia, provocant que tant les seqüències del gen com les del pseudogèn siguin capturades i seqüenciades. Per la mateixa raó, sovint l'alineament de les lectures és inespecífic i no es pot determinar l'origen de la variant. Per poder discernir entre aquestes regions, és necessari recórrer a altres tècniques, com la utilització de lectures llargues o la realització de LR-PCR, posant els encebadors fora de les regions homòlogues (Clendenning *et al.*, 2006; Etzler *et al.*, 2008; Senter *et al.*, 2008; Vaughn *et al.*, 2010; van der Klift *et al.*, 2016; Wagner *et al.*, 2022; Schwenk *et al.*, 2023). Tanmateix, aquestes estratègies són complexes i costoses, fet que les fa poc viables com a cribratge rutinari.

Alguns dels gens de càncer hereditari inclosos a la Instrucció del CatSalut (Taula 5) tenen pseudogens (Taula 6) que poden comprometre la precisió de l'anàlisi quan s'estudien amb NGS de lectures curtes a causa de l'alta homologia que hi presenten.

Estratègies prèvies per analitzar *PMS2*

Un exemple clar és el gen *PMS2*, que presenta alta homologia amb el pseudogèn *PMS2CL*. Davant d'aquest repte, molts laboratoris opten per descartar l'anàlisi d'aquest gen, el que implica perdre informació potencialment rellevant. En altres laboratoris, com es feia anteriorment al SDMCH, s'opta per realitzar LR-PCR a totes les mostres amb pèrdua exclusiva de PMS2 al tumor, assegurant una major precisió però amb un cost i una complexitat elevats. Finalment, altres laboratoris analitzen les dades obtingudes per NGS de lectures curtes sense adaptacions específiques, fet que compromet tant la sensibilitat com l'especificitat dels resultats.

Millora de l'anàlisi mutacional del gen *PMS2*

En aquesta tesi s'ha desenvolupat PMS2_vaR, un *pipeline* en R, disponible a https://github.com/emunte/PMS2_vaR i fàcilment integrable a la rutina diagnòstica, que optimitza l'anàlisi mutacional del gen *PMS2*. A partir de les dades ja disponibles dels panells genètics de diagnòstic i utilitzant eines bioinformàtiques comunes, el *pipeline* identifica variants candidates a pertànyer al gen *PMS2* i les anota amb la classificació que consta a la base de dades del laboratori, alimentada amb els estudis de més de 25 anys d'experiència en l'anàlisi genètica de la síndrome de Lynch. Es recomana l'anàlisi posterior amb LR-PCR/PCR només per a les mostres que presenten variants potencialment patogèniques. Tot i que PMS2_vaR ha estat dissenyat principalment per a l'anàlisi de dades procedents de panells de gens, també es pot aplicar a l'anàlisi d'exomes i genomes. Això és possible gràcies al fet que el *pipeline* utilitza com a dades d'entrada un fitxer BAM, que és el resultat habitual generat per qualsevol tecnologia de seqüenciació NGS.

Els resultats de la validació del nostre estudi demostren que l'ús de PMS2_vaR incrementa la sensibilitat en comparació amb el *pipeline* de diagnòstic anterior, millorant-la del 85,3% al 95,6% sense comprometre l'especificitat. Malgrat no assolir un 100% de sensibilitat, es van detectar totes les variants patogèniques. Les 10 variants que es van descartar eren PSVs classificades com a polimorfismes, que no es van considerar perquè la seva VAF era inferior al 60%. Aquesta decisió es basa en la necessitat d'evitar haver de realitzar LR-PCR a totes les mostres, ja que dues PSVs de l'exó 11 (c.1730dup i c.1864_1865del) serien patogèniques si es localitzessin al gen. Fixant aquest llindar del 60%, es busca que la variant es trobi en com a mínim en tres al·lels, suggerint que estaria al gen (3 de 4 al·lels dels gens *PMS2* i *PMS2CL*).

Algunes PSVs, polimòrfiques al gen i al pseudogèn, com les 10 variants descartades pel *pipeline*, poden donar lloc a excepcions d'aquesta regla i presentar un percentatge inferior al 60% de VAF, malgrat que la variant es trobi al gen (vegeu Figures Suplementàries 2 i 4 del primer article publicat). No obstant això, aquestes variants estan classificades com a benignes i per tant, no tenen impacte clínic. Gould et al. (2018) va proposar utilitzar posicions invariants de l'exó 11, segons l'observació en 707 mostres, per filtrar les variants candidates. S'ha incorporat aquesta estratègia en el codi desenvolupat, que retorna totes les variants i marca específicament aquelles que serien filtrades segons aquesta aproximació. Es recomana que, en casos de pacients amb pèrdua exclusiva de PMS2 en el tumor, s'adopti una actitud conservadora i es confirmi qualsevol variant patogènica a l'exó 11, tant si ha estat filtrada pel codi com si no, per evitar perdre possibles variants del gen en casos de conversió gènica. A més, se suggereix també estudiar les variants recurrents del pseudogèn c.2186_2187del i c.2243_2246del només quan el fenotip clínic ho indiqui.

Prevalença de variants patogèniques en *PMS2* en una cohort de càncer hereditari

Per tal d'estimar la prevalença de variants patogèniques al gen *PMS2*, es va analitzar una cohort de 5.619 mostres de càncer hereditari utilitzant el *pipeline* desenvolupat en aquesta tesi doctoral, PMS2_vaR.

Es van identificar 16 variants patogèniques candidates (0,28%; 16/5619), de les quals 15 van resultar ser del gen i només 1 del pseudogèn. Dels 15 casos identificats, només 10 famílies tenien sospita de síndrome de Lynch (1,28%; 10/847), i en nou dels deu pacients, els tumors mostraven pèrdua de PMS2 o de MSH6/PMS2, cosa que indicava que la deficiència de PMS2 era el principal motor de la carcinogènesi. Per contra, en la cohort amb sospita d'altres síndromes hereditàries, la prevalença va ser de 0,11% (5/4772), en el rang de la prevalença estimada en la població general (0,14%; 1/714) (Wimmer *et al.*, 2014). A més, dels 5 casos en què es va identificar la variant, quatre mostraven immunohistoquímica (IHQ) conservada en el tumor, suggerint que es tracta de tumors esporàdics. Tot i que la prova d'IHQ pot generar falsos negatius, especialment en variants *missense*, aquest resultat també s'alinea amb estudis recents que indiquen que alguns individus portadors de variants

patogèniques en *PMS2* poden desenvolupar tumors sense deficiència del sistema MMR. És important destacar, però, que aquesta prevalença pot estar subestimada, ja que PMS2_vaR no ha estat optimitzat encara per detectar CNVs.

Pros i contres del cribratge oportunista del gen *PMS2*

Actualment, els altres gens MMR (MLH1, MSH2 i MSH6) s'estudien de manera oportunista, tal com recomana la Instrucció del CatSalut (https://scientiasalut.gencat.cat/handle/11351/8438.3). Segons els resultats obtinguts (5 casos identificats entre 4.772 sense sospita prèvia; 0,11%), el nombre de casos a validar per LR-PCR seria reduït, fet que faria viable l'aplicació de PMS2_vaR per al cribratge oportunista en la rutina diagnòstica.

Aquest cribratge permetria, d'una banda, identificar portadors monoal·lèlics de variants en *PMS2* (individus amb LS). Tot i que la utilitat clínica de la identificació de portadors monoal·lèlics és controvertida, ja que les variants en *PMS2* presenten una penetrància més baixa en comparació amb altres gens MMR, aquesta estratègia podria refinar el fenotip associat a alteracions germinals en aquest gen i així contribuir a millorar el coneixement sobre el LS vinculat a *PMS2*. A més, *PMS2* és considerat clínicament accionable segons l'ACMG, que recomana reportar les troballes incidentals en aquest gen (Kalia *et al.*, 2017; Miller *et al.*, 2023).

D'altra banda, el *pipeline* desenvolupat també permetria identificar de manera més eficient pacients amb CMMRD, pels quals el diagnòstic ràpid i precís és crucial per a l'assessorament genètic, prevenció i decisions terapèutiques. En aquest sentit, podria ser d'utilitat el cribratge oportunista de *PMS2* en famílies consanguínies o procedents de poblacions amb colls d'ampolla genètics.

Implementació de PMS2_vaR en el diagnòstic a l'ICO

L'aplicació de PMS2_vaR de manera rutinària es realitza a les mostres amb pèrdua exclusiva de PMS2 al tumor. Anteriorment, aquestes mostres eren analitzades per PCR, on es requerien 3 LR-PCR i 15 PCR curtes per cobrir tot el gen. Gràcies a la utilització del *pipeline*, no només es redueix l'anàlisi a les mostres que tenen possibles variants patogèniques, sinó que en aquestes la realització de PCR (llargues o curtes) es limita a únicament les regions on es troben aquestes variants.

Posteriorment a l'anàlisi de la sèrie inclosa en el treball publicat, s'han analitzat 79 noves mostres amb pèrdua exclusiva de PMS2 al tumor o en les quals no es va poder realitzar la IHQ i tenien sospita clínica de síndrome de Lynch. S'han identificat vuit variants patogèniques candidates, tres de les quals es trobaven en exons amb alta homologia amb *PMS2CL*. Per validar aquestes variants, només han sigut necessàries cinc PCR llargues i vuit PCR curtes, aconseguint una reducció del 97% i el 99%, respectivament (Taula 16). S'ha confirmat que totes les variants pertanyen al gen *PMS2*. Paral·lelament, s'han analitzat totes les mostres amb MLPA, identificant mutacions patogèniques en tres pacients addicionals.

Cal tenir en compte que la interpretació de la IHQ d'MLH1 pot ser complicada i, en alguns casos, una pèrdua aparentment exclusiva de PMS2 pot tractar-se en realitat d'una pèrdua de PMS2 i MLH1, que no s'associa a una alteració genètica a *PMS2*. Això pot distorsionar els càlculs de prevalença, degut a una assignació errònia dels pacients a les diverses cohorts.

Impacte clínic i difusió de PMS2_vaR

Actualment, el SDMCH rep mostres procedents de diferents hospitals de Catalunya per fer l'estudi genètic en casos de sospita de síndrome de Lynch o CMMRD. PMS2_vaR s'està utilitzant en aquells casos en què els pacients presenten pèrdua exclusiva de l'expressió de PMS2 al tumor.

A més, amb l'objectiu d'afavorir la seva integració en el diagnòstic clínic al nostre entorn, es farà difusió d'aquest *pipeline* al consorci SpadaHC, una xarxa de laboratoris d'abast estatal (Moreno-

Cabrera *et al.*, 2024). Es facilitarà suport als hospitals interessats perquè puguin incorporar fàcilment *PMS2_vaR*, optimitzant el diagnòstic i homogeneïtzant la detecció de variants en *PMS2* a escala nacional.

Taula 16: Llistat de variants trobades en les 79 mostres germinals analitzades amb pèrdua exclusiva de *PMS2* al tumor

Id	Variant	Proteïna	Localit-zació	Fenotip personal	Fenotip familiar	Mètode de detecció	Validació	
							LR-PCR	PCR curta
1	c.1A>G	p.?	E01	CCR (63)	CCR (87)	*PMS2_vaR*	Sí	Sí
2	c.137G>T	p.Ser46Ile	E02	CCR (40)	/	*PMS2_vaR*	Sí	Sí
3	c.780del	p.(Asp261Metfs*46)	E07	CCR (79)	STO_nc	*PMS2_vaR*	No	Sí
4	c.780del	p.(Asp261Metfs*46)	E07	CCR (87a)	-ENDO (53) -BR_nc	*PMS2_vaR*	No	Sí
5	c.823C>T	p.(Gln275*)	E08	ENDO (50)	/	*PMS2_vaR*	No	Sí
6	c.1687C>T	p.(Arg563∗)	E11	CCR (70)	-CCR_nc (69) -OV_nc	*PMS2_vaR*	Sí	Sí
7	c.1882C>T	p.(Arg628∗)	E11	CCR (61)	-LG_nc (52) -CCR_nc (60) -CCR_nc (88)	*PMS2_vaR*	Sí	Sí
8	c.1882C>T	p.(Arg628∗)	E11	CCR (52)	/	*PMS2_vaR*	Sí	Sí
9	delE06-E10	p.?	E06-E10	CCR(59) TM(60)	-PAN_nc(74)	MLPA	/	/
10	del E09	p.?	E09	CCR(59)	/	MLPA	/	/
11	del E12-E14	p.?	E12-E14	CCR(60)	/	MLPA	/	/

El número entre parèntesis en els fenotips indica l'edat a la que van desenvolupar el càncer. Abreviatures: BR: càncer de mama; CCR: càncer colorectal; E: exó; ENDO: càncer d'endometri, I: intró, LG: càncer de pulmó; OV: càncer d'ovari; STO: càncer d'estómac, TM: teratoma mediastínic.

Adaptació de PMS2_vaR a l'estudi d'altres gens

PMS2_vaR podria ser adaptat a l'estudi d'altres gens inclosos en la Instrucció del CatSalut que també presenten pseudogens, amb els ajustaments necessaris en el codi. Un exemple destacat és *SDHA*, que té quatre pseudogens no processats amb els quals comparteix més d'un 93% d'identitat: *SDHAP1* (93,65% d'identitat), *SDHAP2* (93,31%), *SDHAP3* (95,22%) i *SDHAP4* (93,72%). Cal destacar que la identitat és més elevada en els exons (Taula 17).

Tot i que s'han dissenyat aproximacions experimentals per a la validació de variants en *SDHA* mitjançant LR-PCR (Bahrambeigi *et al.*, 2016), no existeix cap *pipeline* bioinformàtic específic per prioritzar variants detectades per NGS en aquest gen. El desenvolupament d'un algoritme per aquesta finalitat seria més complex que pel gen *PMS2*, pel fet que, si es força l'alineament de les lectures dels pseudogens cap al gen, en alguns exons podríem arribar a tenir fins a vuit al·lels diferents, complicant així la identificació de variants patogèniques. Això posa de manifest la

necessitat de continuar desenvolupant i refinant eines bioinformàtiques específiques per aquests gens.

És important esmentar que les variants de pèrdua de funció a *SDHA* són les que s'identifiquen amb més freqüència respecte a la resta de gens de la família SDH a gnomAD v4.1.0 (Taula 18). Paradoxalment, *SDHA* és el gen amb menor penetrància (Maniam et al., 2018), de forma similar a *PMS2*. Una possible explicació podria ser que algunes de les variants atribuïdes a *SDHA* provinguin realment dels pseudogens. De fet, al nostre panell tenim lectures que alineen a les seqüències de referència dels quatre pseudogens. Tot i que *SDHD* també presenta set pseudogens amb alta homologia, i per tant podria semblar que té la mateixa problemàtica, aquests són pseudogens processats, és a dir, no contenen introns, fet que podria facilitar-ne la identificació mitjançant lectures curtes.

Taula 17: Percentatge d'identitat compartida entre els exons i les 30 pb annexes dels introns del gen *SDHA* i els seus pseudogens *SDHAP1, SDHAP2, SDHAP3 i SDHAP4*.

SDHA	*SDHAP1*	*SDHAP2*	*SDHAP3*	*SDHAP4*
E01	89,8	89,8	/	89,2
E02	**97,3**	**97,3**	/	**97,3**
E03	**97,3**	**97,3**	/	**97,8**
E04	**95,1**	**95,6**	/	**96,1**
E05	94,7	94,7	/	**96,5**
E06	**95,2**	**95,2**	/	**94,7**
E07	92,4	94,1	/	
E08	94,8	94,8	/	
E09	**95,7**	**95,7**	/	
E10	**95,3**	94,8	**98,3**	
E11	90,9	91,4	**98,3**	
E12	**97,1**	**97,1**	92,1	
E13	94,8	94,8	**97,4**	
E14	**96,6**	**96,0**	**99,4**	
E15	87,7	87,2	92,9	

Les seqüències s'han extret d'Ensembl GRCh38 (SDHA: ENST00000264932.11, SDHAP2: ENST00000455183.1, SDHAP3: ENST00000515467.2, SDHAP4: ENST00000454517.1) i s'ha realitzat un alineament global amb EMBOSS Needle. Es marquen en negreta els exons que comparteixen més d'un 95% d'identitat.

Taula 18: Nombre de variants de pèrdua de funció observades a GnomAD V4.1.0 per a cadascun dels gens *SDHA*, *SDHB*, *SDHC* i *SDHD*

Gen	Variants úniques	Recompte total d'al·lels	Recompte d'al·lels de la variant més recurrent	Classificació a ClinVar
SDHA	84	1161	651	Probablement patogènica / patogènica
SDHB	40	145	26	No classificada
SDHC	18	84	40	Patogènica
SDHD	29	134	32	No classificada

La columna de "Variants Úniques" indica el nombre de variants diferents identificades per gen. La columna de "Recompte Total d'Al·lels" mostra el nombre total d'al·lels observats considerant totes les variants del gen. La columna "Recompte d'al·lels de la variant més recurrent" indica la freqüència de la variant més comuna, i la "Classificació a ClinVar" mostra la classificació segons el seu impacte clínic, quan aquesta està disponible.

Estratègies per millorar la identificació de variants estructurals

La naturalesa de les variants patogèniques és extremadament diversa; abastant des de canvis puntuals en un lloc concret del genoma fins a alteracions estructurals que poden afectar cromosomes sencers. La tria d'una eina per identificar aquestes variants no és un procés trivial, ja que cal considerar diversos factors: el tipus de dades generades (WGS, WES o panell, que influeixen no només per l'abast de la regió estudiada sinó també per la cobertura de la mateixa), la mida de les lectures, el tipus de variants que es pretén identificar (puntuals, indels o SVs) i el mètode que utilitza l'eina per a la seva detecció. En general, una sola eina no pot capturar tota aquesta diversitat, per això és necessari combinar-ne diverses. A la unitat de diagnòstic molecular de l'ICO s'utilitzen VarScan per a detectar mutacions puntuals i petites indels, i DECoN per a CNVs. Aquesta última va ser seleccionada després d'una comparativa de cinc eines amb dades de panell, realitzada pel nostre grup, on va obtenir els millors resultats un cop optimitzats els seus paràmetres per al diagnòstic (Moreno-Cabrera *et al.*, 2020).

Reptes de les anàlisis comparatives d'eines bioinformàtiques

Realitzar anàlisis comparatives és fonamental, especialment perquè constantment estan sorgint noves eines que podrien millorar el rendiment diagnòstic. No obstant això, moltes comparatives presenten limitacions importants, ja que estan fetes pels mateixos autors de les eines avaluades, utilitzen conjunts de dades reduïts, dades simulades o avaluen un nombre limitat d'eines, fets que poden introduir biaixos. Per això, és necessari dur a terme validacions independents, objectives i basades en dades reals de qualitat diagnòstica. D'aquí, se'n destaca la gran importància de compartir dades anonimitzades, que són molt valuoses per avançar en la investigació.

Ara bé, abans d'usar una eina en un context diagnòstic, cada laboratori hauria de fer sempre les seves pròpies validacions amb les seves dades, ja que el rendiment de l'eina pot estar condicionat per la naturalesa d'aquestes dades.

Comparativa de les eines més recents per a la detecció de CNVs a partir de dades de panells

Les comparatives publicades sobre eines per detectar CNVs germinals per a dades de panells de gens presenten algunes de les limitacions descrites anteriorment. Per exemple, a Roca et al. (2019), s'utilitza dades simulades , a Lepkes et al. (2021) s'avalua un nombre limitat d'eines i només es realitza MLPA en els casos identificats com a positius, fet que impossibilita detectar variants no identificades i calcular la sensibilitat i l'especificitat, i a Johansson et al. (2016) s'inclou una comparativa en el mateix article en què presenten la seva eina.

La comparativa prèviament realitzada pel nostre grup (Moreno-Cabrera et al., 2020), aborda la majoria d'aquestes limitacions. En el treball s'avaluaven cinc eines utilitzant dos conjunts de dades públics de panells de gens i es complementen amb dos conjunts de dades procedents del SDMCH de l'ICO. No obstant això, l'aparició de noves versions dels programes, així com d'un gran nombre d'eines noves per a la detecció de CNVs, ha fet que aquesta comparativa quedi desactualitzada.

Per aquest motiu, es va decidir fer-ne una de nova on s'ha incorporat les versions actualitzades (si n'hi havia) dels cinc programes analitzats prèviament, juntament amb noves eines que complissin els requisits metodològics detallats en el segon article publicat d'aquesta tesi doctoral.

Malgrat els esforços per identificar nous conjunts de dades per ampliar la nova comparativa, no s'han trobat conjunts addicionals disponibles (última revisió abril 2024). Per tant, s'han utilitzat els mateixos conjunts que en la comparativa anterior. A part d'avaluar el rendiment de les eines a nivell individual amb els valors per defecte, com es va fer en el darrer estudi, en aquesta comparativa s'ha avaluat les combinacions per parelles i també s'ha analitzat l'impacte de modificar paràmetres

individuals en el rendiment. Les mètriques s'han calculat tant per regió d'interès com a escala de gen.

Anàlisi dels valors per defecte

En l'anàlisi dels valors per defecte a escala individual, els resultats suggereixen l'ús de diferents eines segons l'objectiu perseguit. Si es vol maximitzar el rendiment global (*F1 score,* una mètrica que combina precisió i sensibilitat per mesurar l'equilibri entre encerts i detecció de casos reals), ClinCNV i GATK-gCNV són les millors opcions. En canvi, si l'objectiu és augmentar la sensibilitat, les eines més efectives són CoNVaDING, GATK-gCNV, CODEX2 i DECoN. La sensibilitat és un factor clau en contextos diagnòstics, on es prioritza la minimització de falsos negatius. En aquest escenari, resulta més útil calcular les mètriques a escala de gen, ja que només cal detectar alguna alteració en el gen per després validar-la amb MLPA, sense necessitat de conèixer tots els exons implicats. Tanmateix, és important no comprometre gaire l'especificitat, perquè totes les alteracions potencials han de ser validades a posteriori experimentalment, i amb l'augment del nombre de gens analitzats, augmenta també el nombre de falsos positius a validar. Així doncs, tot i que CoNVaDING genera pocs falsos negatius, produeix un nombre elevat de falsos positius, fet que la fa poc adequada per l'ús diagnòstic. L'eina que millor equilibra sensibilitat i especificitat és GATK-gCNV, convertint-la en una candidata sòlida per ser usada en rutina diagnòstica. El nostre estudi és el primer que avalua aquesta eina amb dades de panell.

Combinació d'eines per parelles per optimitzar la detecció de CNVs

Si una eina és capaç de detectar tots els CNVs, es podria utilitzar com a cribratge i només caldria confirmar amb MLPA els gens en els quals s'hagi detectat una possible variant candidata, reduint així significativament els costos. No obstant això, cap eina va assolir un 100% de sensibilitat de manera individual en tots els conjunts de dades. Per aquest motiu es va avaluar si la combinació d'eines per parelles podia assolir aquest objectiu, donat que és comú en bioinformàtica combinar o intersecar els resultats de diferents eines per generar meta-identificadors més útils en un context clínic (Samarakoon et al., 2014; Liu et al., 2016; Gabrielaite et al., 2021).

A escala de gen, cinc combinacions d'eines van aconseguir detectar totes les variants, tot i que l'especificitat va variar en funció del conjunt de dades. Aquesta variabilitat en l'especificitat, torna a posar de manifest la importància que els laboratoris realitzin les seves pròpies comprovacions amb les seves dades, motiu pel qual es va posar a disposició oberta el codi informàtic desenvolupat en el present treball (https://github.com/emunte/PMS2_vaR).

Impacte de l'ajust de paràmetres en el rendiment de les eines

Un problema comú és que la documentació de les eines bioinformàtiques sovint és limitada, cosa que fa difícil comprendre l'impacte dels paràmetres sobre el rendiment. Per abordar aquesta qüestió, vam repetir les execucions de la comparativa modificant cada paràmetre de manera individual dins d'un rang de valors. En total, es van generar 436 figures que poden servir com a guia de referència per a laboratoris de recerca i diagnòstic que utilitzen o volen utilitzar aquestes eines.

Els paràmetres seguien quatre patrons de comportament. Alguns no van tenir cap efecte en el rendiment i no haurien de ser considerats per millorar-lo. D'altres presentaven el compromís típic dels classificadors binaris: un augment en la sensibilitat implica una disminució en l'especificitat, i viceversa, pel que es recomana ajustar aquests paràmetres per equilibrar sensibilitat i especificitat. Altres paràmetres mostraven un comportament en forma de campana, suggerint que el rendiment de l'eina es pot optimitzar al voltant d'un valor concret. Finalment, alguns paràmetres van afectar el rendiment sense un patró clar, amb variacions en la sensibilitat. Aquests s'han de modificar amb precaució, ja que la seva variabilitat fa difícil preveure'n l'impacte.

D'acord amb aquests patrons, els nostres resultats suggereixen modificar 13 paràmetres de diferents eines, en els quals vam observar una millora en l'F1 *score*. Aquestes millores van anar des d'increments modestos del 0,07% a increments de 4,38% per alguns paràmetres.

Dues hipòtesis podrien explicar per què els valors per defecte no aconsegueixen la màxima *F1 score*. En primer lloc, pot ser que les eines s'hagin optimitzat utilitzant un conjunt de dades específic de referència, de manera que el rendiment no es generalitza bé a altres conjunts de dades. En segon lloc, és possible que aquestes eines es dissenyessin optimitzant mètriques diferents de l'*F1 score*. En qualsevol cas, aquestes hipòtesis reafirmen la necessitat que cada laboratori validi els ajustos de paràmetres amb les seves pròpies dades. La nostra aportació, per tant, constitueix una guia sobre els paràmetres que podrien ser més interessants avaluar en els processos de validació.

DECoN o GATK -gCNV en el context diagnòstic

En l'anterior comparativa del nostre grup es van optimitzar els paràmetres de DECoN per tal d'utilitzar-lo com a eina de cribratge en un context diagnòstic. L'optimització pels conjunts de dades internes provinents de seqüenciadors Illumina dels models MiSeq i HiSeq va permetre detectar totes les CNVs analitzades excepte una, sense comprometre l'especificitat. L'única variant no detectada era d'un pacient amb mosaïcisme genètic, fet que en dificultava la detecció a causa de la seva baixa VAF. El rendiment global observat va implicar la implementació del programa DECoN com a mètode de cribratge previ a validació per MLPA a la nostra rutina diagnòstica.

En la comparativa realitzada en la present tesi doctoral, entre els molts resultats obtinguts, hem observat que pels nostres conjunts de dades, GATK-gCNV també destaca pel seu rendiment a escala de gen, obtenint una *F1 score* i una sensibilitat altes amb els seus paràmetres per defecte.

Tanmateix, a diferència de l'estudi previ, aquest treball ha analitzat la influència de cada paràmetre de manera individual, en lloc d'una optimització conjunta, per comprendre millor el pes específic de cadascun. Per aquest motiu, no es poden extreure conclusions sobre quina eina seria millor per a un context diagnòstic.

Com a pas futur, seria interessant explorar l'optimització conjunta dels paràmetres de GATK-gCNV, tot i que aquesta tasca seria computacionalment exigent, ja que és l'eina amb més paràmetres (vegeu Taula 2 del segon article publicat). Així mateix, altres eines prometedores, com clinCNV, també podrien ser candidates per avaluar si poden detectar el 100% de les variants i reduir el nombre de falsos positius generats per DECoN, fet que disminuiria la necessitat de proves MLPA addicionals i per tant reduiria costos mantenint o incrementant la sensibilitat.

Ara bé, donat que la versió optimitzada de DECoN ja ofereix una sensibilitat molt alta, qualsevol possible millora podria ser difícil de quantificar amb els conjunts de dades actuals. Per aquest motiu, caldria ampliar el nombre de mostres per aconseguir una millor discriminació de possibles diferències. En cas de no disposar de noves mostres, una alternativa podria ser l'ús de tècniques com el *downsampling* de les lectures per "complicar" el conjunt de referència, simulant condicions més exigents i permetent una avaluació més precisa de les possibles millores.

Detecció d'indels de mida mitjana i variants estructurals equilibrades

Hi ha insercions i delecions de mida mitjana, així com SVs equilibrades, que escapen a la detecció dels *pipelines* convencionals. Això es deu al fet que la seqüència alterada és prou llarga perquè els algoritmes interpretin les lectures com a quimèriques, resultant en manca d'alineament o *soft-clipping*, que no es detecten pels *variant callers* de SNVs i indels habitualment utilitzats. A més, si no tenen almenys la mida d'un exó complet o no impliquen un canvi en el nombre de còpies, tampoc es detecten pels algoritmes basats en profunditat de cobertura.

La identificació d'aquestes variants requereix altres estratègies, com les de lectures dividides, el mapatge d'extrems aparellats i l'assemblatge *de novo*. Com que cadascuna d'aquestes tècniques té els seus avantatges i inconvenients, cada cop més eines combinen diversos mètodes per maximitzar la detecció de variants (Wang et al., 2022).

Anàlisi de l'eina GRIDSS per la detecció d'aquestes variants

En aquest sentit, l'eina GRIDSS esdevingué un candidat amb potencial, ja que integra l'ús de lectures dividides, el mapatge d'extrems aparellats i l'assemblatge *de novo* (Cameron *et al.*, 2017). No obstant això cal tenir en compte una sèrie de limitacions. D'una banda, GRIDSS va ser dissenyat originalment per a dades de genoma i el seu ús no ha estat validat en dades de panells de gens. D'altra banda, retorna les variants en notació de *breakend*, una representació útil per descriure tota l'heterogeneïtat de les SVs, però molt complexa d'interpretar, ja que tots els esdeveniments es reporten com a punts de trencament simples, requerint passos addicionals per categoritzar-los correctament. A més, GRIDSS genera un gran nombre de variants, fet que el fa inviable per a la pràctica diagnòstica sense filtres adequats per prioritzar les variants. Mentre que els autors de l'article van desenvolupar un codi amb filtres per prioritzar variants somàtiques, no han desenvolupat cap per filtrar variants germinals.

Per avaluar GRIDSS i abordar les limitacions exposades, en aquesta tesi s'ha desenvolupat un codi que: 1) processa tots els fitxers VCF generats per GRIDSS en notació *breakend*, assegurant que cada variant es representi en una sola fila; 2) aplica filtres per reduir el nombre de variants a validar. Els filtres exclouen variants recurrents (n > 10) en el nostre conjunt de dades, se centren en regions clínicament rellevants, eviten artefactes d'alineament i regions hipervariables, i descarten variants més petites de 20 parells de bases, que habitualment són capturades pels *pipelines* de rutina diagnòstica habitual.

Resultats obtinguts amb la implementació dels filtres

Així doncs, es va aplicar GRIDSS en una cohort de 9.750 mostres seqüenciades en 256 carreres, i es van identificar 1.141.613 variants. Després d'aplicar els filtres i descartar les variants detectades prèviament per les eines habituals, es van seleccionar 54 variants per a inspecció visual. Això equival a una mitjana d'una variant per revisar visualment per cada 4,74 carreres, un volum assumible en rutina diagnòstica.

Després de la inspecció visual, 24 variants es van considerar com a bones candidates. D'aquestes es van seleccionar nou per validació experimental ja que eren les úniques clínicament rellevants. Les nou variants es van confirmar a nivell experimental, vuit es van classificar com a (probablement) patogèniques, la qual cosa va representar un augment del rendiment diagnòstic del 0,6%, mentre que la novena es va classificar com a VSD. Concretament aquests resultats van permetre diagnosticar dos pacients amb síndrome de Lynch, un pacient amb poliposi adenomatosa familiar, quatre pacients amb variants en gens relacionats amb càncer de mama i un amb una variant relacionada amb càncer de pròstata. Les noves troballes tenen un alt valor clínic, ja que involucren gens d'alt o moderat risc amb protocols ben establerts per a la prevenció i tractament. A més, obren la porta a estudis de portadors en cascada als familiars de primer grau, cosa que permet personalitzar el risc de cada individu i establir les mesures de prevenció necessàries.

Cal destacar que cinc de les vuit variants (probablement) patogèniques detectades eren insercions d'elements transposables en la regió codificant (quatre Alus i un LINE), un tipus de variants que no es detectaven amb els *pipelines* rutinaris i que són especialment rellevants, ja que es calcula que representen fins a un 0,3% de les variants patogèniques identificades en gens de succeptibilitat al càncer (Qian et al., 2017).

Limitacions de GRIDSS

Tot i el seu gran potencial, GRIDSS presenta diverses limitacions que cal considerar, especialment si es parteix de dades procedents de panells de gens. En primer lloc, requereix que en el punt exacte de trencament hi hagi suficient cobertura per poder identificar les variants. Mentre que això no és un problema si tenim dades de genoma, sí que pot suposar un inconvenient si s'usen dades d'exoma o de panells, ja que pot provocar la pèrdua de variants. Per aquest motiu, recomanem combinar GRIDSS amb eines que identifiquin CNVs basant-se en patrons de cobertura, com és el cas de DECoN.

En segon lloc, el propi GRIDSS pot descartar variants reals per criteris de qualitat. Un exemple il·lustratiu és la inserció d'*Alu* de *BRCA2* c.2197_2198insAluYb8, detectada (i no filtrada) en la pacient 23, diagnosticada de càncer de mama. També es van incloure dos familiars en l'estudi: la seva germana, diagnosticada amb càncer de mama als 36 anys, i una cosina llunyana, diagnosticada amb càncer de mama als 37 anys. Inicialment, la inserció *Alu* no es va detectar en cap de les dues familiars. No obstant això, en inspeccionar amb més detall els fitxers VCF, es va observar que la variant estava present en les dades primàries de la germana, però no havia superat el llindar de qualitat de GRIDSS (qualitat de 918 d'un mínim de 1.500) i, per tant, havia estat descartada. A més, hi havia poques lectures que donaven suport a aquesta variant (VAF = 6,4%). Tanmateix, la inspecció visual mitjançant IGV va suggerir que la variant era present. Per tant, és plausible que altres SVs amb paràmetres de qualitat baixos hagin quedat sense detectar per aquest mateix motiu.

La rigidesa dels filtres emprats pot excloure variants potencialment rellevants. Per exemple, restringir l'anàlisi a variants detectades en més de 10 mostres, amb l'objectiu de reduir els falsos positius, podria descartar variants patogèniques recurrents o fundadores, no identificades per altres eines. No obstant això, sabent que aquest criteri podia excloure SVs rellevants, es va relaxar fins a 15 mostres, la qual cosa va permetre identificar una inserció AluYa5 a l'intró 54 del gen *ATM* (c.8010+30_8010+31insAluYa5; p.(Val2671Serfs*17)). Aquesta variant es va trobar en cinc pacients amb càncer de mama i en sis individus sense sospita clínica d'alteracions relacionades amb *ATM*. Segons es descriu a Klein et al. (2023), l'anàlisi de transcrits de l'RNA de tres portadors heterozigots va mostrar la pèrdua de l'exó 54 en el 38% dels transcrits totals d'*ATM*. En el seu estudi, la variant es va detectar en 6 de 303 pacients amb sospita de càncer hereditari (1,98%). Per avaluar-ne la freqüència en població control, els autors van analitzar les dades del projecte 1000 Genomes amb Mobster (Thung *et al.*, 2014) i SCRAMble (Torene *et al.*, 2020) i la van identificar en 3 de 2.984 individus (0,1%), considerant que això posava de manifest la seva baixa presència en poblacions sanes. Tanmateix, la comparació de freqüències en cohorts de malaltia pròpies d'un laboratori o regió amb dades poblacionals de gnomAD pot patir de biaix d'estratificació.

En analitzar la proporció de mostres de la nostra cohort on el gen *ATM* formava part del conjunt de gens a analitzar (és a dir, el gen *ATM* podria explicar la sospita inicial), es va observar que la freqüència de la inserció era fins i tot inferior (0,08%) en comparació amb aquelles on no hi estava inclòs (0,16%). Aquestes dades, per tant, suggereixen que aquesta inserció no sembla estar associada amb un risc incrementat de tumors associats a *ATM*. Aquests resultats recolzen el llindar escollit de 10 mostres.

D'altra banda, establir un llindar de VAF del 10% podria ser massa restrictiu, especialment per a SVs en mosaïcisme o en regions amb alineament dificultós, on les lectures poden ser escasses o parcialment alineades. També s'eviten regions repetitives per minimitzar artefactes, fet que pot comportar la pèrdua de variants reals patogèniques en aquestes zones.

Implementació de GRIDSS en la rutina diagnòstica: viabilitat i futur

No obstant això, el codi desenvolupat per aplicar els filtres és altament modulable i permet ajustar els llindars en funció del tipus de població analitzada i dels recursos disponibles per a la visualització

i validació de variants. Aquesta flexibilitat facilita l'adaptació de l'eina a les necessitats específiques de cada laboratori, aconseguint un balanç òptim entre sensibilitat i especificitat.

Per tant, l'aplicació de GRIDSS en la rutina diagnòstica té un gran potencial, però per garantir una implementació eficient caldria adaptar el codi, desenvolupant una base de dades interna per emmagatzemar variants detectades retrospectivament i prospectivament. Això permetria aplicar filtres de recurrència (descartar variants MAF>=0,1%, 10/9750 a la sèrie interna), malgrat que s'estiguin analitzant poques variants a la vegada. L'aplicació d'aquest filtre és particularment eficient quan es comparen un gran nombre de mostres, però si només es processen mostres d'una única carrera, es perdria gran part del seu avantatge.

L'ús de bases de dades poblacionals com gnomAD no seria tan útil en aquest context, ja que no eliminaria els falsos positius generats, perquè aquests es condicionen a les tècniques utilitzades i la naturalesa de les dades i a més, l'anotació de les SVs és complexa, fet que faria difícil comparar les dades per identificar els artefactes i polimorfismes.

La resta de filtres no suposen cap repte significatiu i podrien ser aplicats de manera semblant a com s'ha descrit en l'article. Posteriorment, és imprescindible que un expert revisi les variants visualment i que aquelles que es considerin com potencialment rellevants a la clínica es validin experimentalment.

Alternatives tecnològiques per a la identificació de variants no detectades

Davant les limitacions inherents a l'ús de l'NGS de lectures curtes (segona generació), és important considerar tecnologies alternatives que podrien superar algunes d'aquestes barreres. Una opció prometedora és la seqüenciació de tercera generació que utilitza lectures llargues i ofereix avantatges significatius en l'estudi de certes variants. Aquesta tecnologia permet identificar diferents isoformes, SVs complexes, modificacions nucleotídiques que poden passar desapercebudes amb l'ús de lectures curtes. També pot ajudar a resoldre regions amb pseudogens o altres zones d'alta homologia.

Avui dia, aquesta tecnologia encara presenta algunes dificultats, com el seu cos elevat, una menor qualitat de les lectures respecte als mètodes de segona generació i la necessitat d'utilitzar eines bioinformàtiques i infraestructures especialitzades per al seu processament i emmagatzematge, ja que genera un gran volum de dades (Scarano *et al.*, 2024). No obstant això, el seu gran potencial fa que valgui la pena començar a explorar-la, especialment tenint en compte la ràpida evolució tecnològica i la previsió d'una reducció dels costos.

Variants més enllà de les regions d'interès habituals: anàlisi de les 5'UTR

Les guies de classificació actuals se centren principalment en l'anàlisi de les regions codificants i les regions canòniques d'empalmament. Tanmateix, també s'ha demostrat l'existència de variants patogèniques en àrees com les regions intròniques profundes, regions promotores o regions UTRs.

Les variants localitzades en les regions 5'UTR, tot i no codificar per proteïnes, són essencials per la regulació de l'expressió gènica a escala post-transcripcional. Tot i que part d'aquestes regions estan cobertes al nostre panell, en el SDMCH no s'analitzen de manera habitual donada la complexitat de la interpretació i validació funcional d'aquestes variants.

Un pas endavant en el diagnòstic genètic consistirà probablement en l'estudi rutinari d'aquestes regions, ja que probablement s'estan perdent variants que podrien explicar la predisposició al càncer en algunes famílies. En aquest sentit, el grup liderat per la professora Whiffin ha proposat l'ús de filtres per prioritzar variants prometedores en aquestes zones. Aquests permeten mantenir un

nombre de variants candidates viable per a una posterior validació funcional i classificació (Martin-Geary et al., 2023). Els filtres proposats es basen en criteris com la creació o destrucció d'un oORF que puguin influir sobre la traducció de l'ORF principal segons les prediccions de l'UTRannotator, prediccions de SpliceAI que indiquin alteracions potencials en l'empalmament o variants que modifiquin la seqüència de Kozak de l'ORF principal, condicionant l'eficiència de la traducció.

Variants identificades a la regió 5'UTR emprant els filtres descrits

Gràcies a l'estança de la doctorand en el laboratori de la Dra. Whiffin es van poder analitzar en retrospectiu les dades d'NGS de les 5'UTR del SDMCH emprant els programes i filtres descrits a l'apartat anterior. Aquests filtres han permès prioritzar set variants per a una anàlisi més detallada. D'aquestes, quatre van ser seleccionades per UTRannotator i una d'elles, la variant *CDKN2A* c.-34G>T, ja descrita en la literatura, és patogènica (Liu *et al.*, 1999). Aquest resultat podrà integrar-se amb la resta d'informació clínica dels tres probands portadors, així com amb la història clínica familiar de cadascun, per ajustar el seu risc i les mesures de seguiment adequades.

En relació a la variant *MSH6* c.-22T>G, malgrat concordar amb el fenotip familiar, la IHQ del tumor i els estudis de co-segregació van suggerir que aquesta variant no seria la causa de la predisposició al càncer observada a la família. Les altres dues variants, *STK11* c.-33C>A i *AIP* c.-76T>C, tenen un efecte desconegut en absència d'estudis funcionals. La primera variant prediu un oORF en el marc de lectura que allarga la proteïna 11 aminoàcids a l'extrem N terminal. Tot i que els pòlips del pacient no semblen hamartomatosos (associats típicament a condicions relacionades amb *STK11*), la manca d'informació clínica completa impedeix descartar definitivament aquesta possibilitat. La variant d'*AIP* no concorda amb el fenotip del pacient, ja que *AIP* s'associa normalment amb tumors hipofisiaris, i el pacient presenta càncer de pròstata.

Les dues variants que es van prioritzar per efectes d'empalmament, no sembla que hagin de tenir un impacte significatiu en aquest després d'analitzar-les més a fons. Tot i que es prediu la creació de nous llocs donadors, cap d'ells supera la força dels llocs donadors naturals, suggerint que qualsevol empalmament alternatiu es produiria amb menor proporció que la isoforma completa. Tanmateix, encara que eines *in silico* com SpliceAI tenen un valor predictiu molt alt, no sempre reflecteixen perfectament els resultats biològics, i serien necessaris estudis d'RNA per confirmar l'efecte real en l'empalmament.

L'última variant que es va prioritzar afecta la posició -3 de *BRCA1*. Les prediccions fetes per Noderer et al. (2014) suggereixen que el canvi disminuiria una mica l'eficiència de la traducció. No obstant això, estudis d'RNA fets a altres variants de *BRCA1* indiquen que només una funció residual d'entre el 20 i el 30% de la proteïna és suficient per evitar un risc alt de càncer (de la Hoya *et al.*, 2016), per la qual cosa és poc probable que aquesta variant sigui la causa de la condició familiar, però caldrien estudis funcionals per poder-ho afirmar.

Limitacions del nostre estudi

Tot i que els resultats obtinguts en aquest primer estudi, semblen interessants, cal tenir en compte que aquest ha estat un estudi molt preliminar i amb poca profunditat pel que podríem estar perdent variants potencialment rellevants per diversos motius.

Limitacions d'UTRannotator

UTRannotator és una eina potent, però per exemple, no prediu uORFs que es creïn a partir de codons d'inici no canònics, que s'ha demostrat que poden iniciar la traducció amb un eficiència considerable en determinats contextos de la seqüència de Kozak (De Arce, Noderer and Wang, 2018; Andreev *et al.*, 2022; Chothani *et al.*, 2022). A més, tampoc considera les variants que poden modificar la força de la seqüència de Kozak dels uORFs existents, la qual cosa podria impactar en l'expressió gènica.

Cobertura i variants cridades

Tot i que tenim cobertura completa de la regió 5'UTR exònica per a 35 gens de la Instrucció del CatSalut, per a 19 gens no tan freqüentment analitzats només s'han capturat aproximadament 150 bp aigües amunt del codó d'inici, i per a un gen, 500 bp. En aquests 20 casos, aquesta extensió és insuficient per abastar tota la regió exònica de la 5'UTR. Per tant, no s'haurien detectat variants rellevants en les regions no cobertes. S'està estudiant la viabilitat de cobrir completament els exons de la 5'UTR d'aquests gens en futures versions del panell.

A més, el present estudi s'ha centrat en SNVs i indels petites, sense considerar les CNVs. En futurs estudis caldria incloure aquestes últimes per poder oferir una fotografia més completa de l'abast d'aquestes variants. Per exemple, s'ha predit que delecions que afecten el primer exó i el promotor de la 5'UTR de *MEF2C* poden interrompre la funció dels potenciadors, i de fet, s'han identificat en pacients amb trastorns del desenvolupament (Wright *et al.*, 2021).

Filtres aplicats

Els filtres aplicats en aquest estudi són restrictius per mantenir un nombre viable de variants candidates (Martin-Geary et al., 2023). Podria ser que variants amb conseqüències diferents a les prioritzades puguin tenir un efecte regulador. En serien exemples les variants que creen uORFs no solapants que redueixin l'expressió de l'ORF principal, les variants no prioritzades per SpliceAI que tinguin una conseqüència en l'empalmament o variants que puguin influir la seqüència de Kozak sense afectar la posició -3.

Validació experimental

Aquest estudi s'ha centrat en la priorització de les variants més prometedores a la regió 5'UTR mitjançant predictors *in silico*, sense abordar validacions experimentals. Les prediccions *in silico* actuals per a les 5'UTR es basen en algoritmes i models computacionals que, tot i ser molt útils, presenten limitacions a l'hora de predir amb precisió l'impacte funcional de les variants. Per aquest motiu, a dia d'avui, cal realitzar estudis funcionals per poder interpretar clínicament aquestes variants.

Tot i que en altres apartats de la tesi s'han dut a terme validacions experimentals per confirmar les variants prioritzades i/o el seu impacte funcional, en aquest cas no ha estat possible per les limitacions de temps i l'abast del projecte.

En el futur, la realització d'assajos massius de mutagènesi funcional (MAVEs, de l'anglès *Multiplexed assays of variant effect*), podria permetre crear predictors més acurats, amb un valor predictiu prou alt com per evitar la necessitat de validació en alguns casos.

Altres regions d'interès

En aquest estudi ens hem centrat exclusivament en les variants de les regions 5'UTR, però estudis futurs haurien de considerar també altres regions no codificants, com les 3'UTR i les IRES, ja que també poden tenir un paper regulador important. Ara bé, les 3'UTR són més llargues que les 5'UTR (longitud mitjana d'aproximadament 1775 pb en comparació amb els 202 bp de les 5'UTR), fet que n'incrementa la complexitat i el cost de la seva anàlisi (Ellingford *et al.*, 2022; Wieder *et al.*, 2024).

Taula 18: Adaptacions de les guies ACMG per a variants en regions no codificants.						
	Criteris a favor de benignitat		Criteris a favor de patogenicitat			
Pes	**Fort**	**Suport**	**Suport**	**Moderat**	**Fort**	**Molt fort**
Dades de població	La MAF és massa alta per al trastorn BA1/BS1 O observació en controls incompatible amb la penetrància de la malaltia BS2.		Absent en bases de dades poblacionals PM2_Sup		Prevalença en individus afectats estadísticament superior a la dels controls PS4	
Dades computacionals i predictives		Diverses línies d'evidència computacional suggereixen una manca d'impacte en el gen/producte del gen BP4.	Diverses línies d'evidència computacional suggereixen un efecte perjudicial sobre el gen/producte del gen PP3. Variant d'empalmament en el mateix nucleòtid que una variant patogènica establerta PS1_Sup.	Mateix impacte predit que una variant patogènica establerta PM5. Variant que canvia una proteïna PM4.		Variant nul·la en un gen on la pèrdua de funció és un mecanisme conegut de la malaltia PVS1.
Dades funcionals	Estudis funcionals quantitatius ben establerts en teixits/cèl·lules derivades de pacients mostren manca d'efecte perjudicial BS3.		*Hot spot* mutacional o domini funcional ben estudiat sense variants benignes PM1_Sup.		Estudis funcionals quantitatius en teixits/cèl·lules derivats de pacients mostren un efecte perjudicial PS3	
Dades de segregació	No segregació amb la malaltia BS4.		Segregació amb la malaltia en múltiples membres afectats de la família PP1.	Més dades de segregació		
Dades de novo				*De novo* (sense paternitat/maternitat confirmada) PM6.	De novo (amb paternitat i maternitat confirmada) PS2.	
Dades al·lèliques		Observada en *trans* amb una variant patogènica dominant BP2. Observada en *cis* amb una variant patogènica establerta BP2.		Per a trastorns recessius, detectada en *trans* amb una variant patogènica PM3.		
Altres dades		Cas trobat amb una causa alternativa BP5	El fenotip del pacient o HPO és altament específic per a la malaltia PP4.			

Les regles que necessiten orientació addicional per a variants en regions no codificants estan en negre, mentre que aquelles que requereixen consideracions extres o adaptació estan en color. Traduït d'Ellingford et al. (2022), originalment adaptat de Richards et al. (2015).

Guies de classificació específiques per regions no codificants

A causa de la manca de directrius específiques per a les regions no codificants a les guies ACMG, un grup d'experts ha desenvolupat recomanacions per adaptar els criteris ACMG a aquestes regions (Ellingford *et al.*, 2022). Donada la dificultat per predir amb precisió l'efecte i l'impacte fenotípic de moltes d'aquestes variants, les guies s'han formulat amb un enfocament conservador, reduint la força de l'evidència aplicada a cada criteri. Així, s'han suggerit interpretacions adaptades per a cadascun dels criteris establerts (Taula 18). Per exemple, el criteri PM5, originalment destinat a variants *missense*, s'ha ajustat per incloure variants no codificants amb un impacte similar sobre el mateix gen, seguint l'evidència d'altres variants patogèniques conegudes.

Un dels reptes fonamentals en l'estudi d'aquestes regions és la manca de conjunts de dades NGS a gran escala que incloguin cohorts representatives de diferents poblacions i amb fenotips associats. Aquesta limitació redueix el poder estadístic i dificulta establir correlacions fiables entre variants i fenotips de malaltia. Això subratlla un tema recurrent en aquesta tesi: la importància de compartir dades.

Automatització de la classificació de variants

En els últims anys, la transició de l'estudi de gens individuals a l'ús d'NGS ha revolucionat el camp de la genètica clínica, generant un volum creixent de variants que s'han de classificar. Aquesta classificació s'ha convertit en un factor limitant que sovint és un coll d'ampolla en el procés diagnòstic. A més, si bé anteriorment es feien servir guies generals per classificar variants, avui existeixen nombroses guies gen-específiques. La publicació d'aquestes és essencial per assegurar una classificació precisa i adaptada als mecanismes propis de cada gen. Tanmateix, incrementen les consideracions que els classificadors de variants han de tenir en compte, augmentant la complexitat de les decisions i les possibilitats d'error humà.

Aquest escenari posa de manifest la necessitat d'eines que puguin automatitzar part del procés, ajudant a agilitzar-lo i minimitzant els possibles errors humans. Així i tot, aquests sistemes mai han de reemplaçar la figura de l'especialista, que continua sent fonamental per integrar adequadament les dades clíniques, genètiques, els estudis funcionals i la revisió bibliogràfica.

En resposta a aquestes necessitats, els darrers anys, s'han desenvolupat diverses eines d'automatització (Taula 13), moltes de les quals es basen en les guies generals proposades per Richards et al. (2015). Tanmateix, només una petita part d'aquestes eines inclou guies específiques per a alguns gens. En l'àmbit del càncer hereditari, una de les poques eines que integra algunes guies específiques és Cancer SIGVAR, però no incorpora actualitzacions recents, limitant el seu valor diagnòstic en un camp amb recomanacions que evolucionen constantment (Li *et al.*, 2021).

Desenvolupament de vaRHC

En aquesta tesi es presenta el desnvolupament del paquet d'R vaRHC, que semi-automatitza el procés de classificació de variants utilitzant guies específiques per a gens de càncer hereditari. El programa recopila informació de diverses bases de dades sobre les variants detectades i la combina per assignar o rebutjar criteris segons les guies més actualitzades en el moment de la publicació. Per fer la classificació més flexible i acurada, vaRHC empra l'enfocament bayesià de Tavtigian et al. (2018, 2020) i incorpora la major part de les recomanacions de combinació d'evidències de CanVIG-UK, evitant la doble assignació de criteris solapants i minimitzant el risc d'errors per puntuació duplicada (Garrett *et al.*, 2021). Els resultats obtinguts per consola es poden descarregar en un format editable de full de càlcul (.xlsx). Així es facilita una revisió i classificació àgils i precises de les variants, i un

registre acurat de les evidències no automatitzables junt a les automatitzades, fins i tot per a aquells que no són experts en bioinformàtica.

Selecció del conjunt de dades per la validació

La selecció del conjunt de dades adequat per validar un programa de classificació de variants no és un procés trivial, ja que cal garantir que l'eina avaluï les variants amb precisió i compleixi les guies establertes per cada gen. Molts programes opten per validar la classificació global de les variants utilitzant grans bases de dades com ClinVar (Landrum *et al.*, 2018). Tot i que ClinVar és una font valuosa, presenta limitacions: la majoria de les seves entrades provenen de laboratoris individuals sense una verificació uniforme, altres presenten discrepàncies en la classificació i sovint no inclou informació detallada sobre els criteris seguits per classificar les variants, només sobre la classificació final. Això impossibilita garantir que s'han aplicat les guies de manera consistent i de manera gen-específica si s'escau.

Per a la validació de vaRHC, vam voler assegurar-nos que cada criteri completament automatitzable s'estigués aplicant correctament. Per aquest motiu, vam decidir utilitzar el conjunt de dades del registre de ClinGen, centrant-nos en gens d'alt risc associats a càncer hereditari amb guies específiques en data de març de 2022 (*ATM, CDH1, PTEN* i *TP53*). Tot i que aquests conjunts de dades són més reduïts en comparació amb ClinVar, estan curosament classificats per experts i proporcionen informació exhaustiva sobre l'assignació de cada criteri per a cada variant, cosa que ens permet avaluar criteri per criteri. A més, vam ampliar el conjunt de dades amb variants del gen *CHEK2* proporcionades a l'article Vargas-Parra et al. (2020) i amb variants de gens MMR de la base de dades interna del SDMCH de l'ICO, aplicant per tots aquests gens també una avaluació criteri per criteri.

Finalment, per comparar el nostre programa amb Cancer-SIGVAR, vam utilitzar únicament els conjunts de dades de ClinGen per *CDH1* i *PTEN*, ja que són els únics gens amb guies específiques recollides a Cancer-SIGVAR.

Validació

Dels resultats de la validació podem extreure diverses conclusions. En primer lloc, hem comprovat la robustesa del programa en l'assignació dels criteris completament automatitzables, aconseguint una concordança del 97,7% amb els criteris assignats manualment. Només un 63,4% de les variants coincidien amb la classificació final manual, discrepància que s'explica perquè molts criteris no són automatitzats o només ho són parcialment. En un ús real del programa es compta amb la participació de l'usuari per afegir-los dins l'arxiu excel generat per vaRHC, que llavors recalcula automàticament la classificació final. Algunes eines, per compensar la manca d'informació derivada de criteris no automatitzables, són menys restrictives en l'assignació de certs criteris, com ara els llindars de freqüència de població per als criteris BA1, BS1 i PM2 en programes com Varsome. Addicionalment, programes com Cancer SIGVAR, InterVar, Varsome, Franklin, Pathoman i CharGer, utilitzen criteris que actualment no són aplicables (com PP5 i BP6) per aproximar els seus resultats als de ClinVar, usant la pròpia classifcació a ClinVar com un criteri que pot arribar a tenir força forta o fins i tot molt forta (Biesecker *et al.*, 2018). Com que en moltes validacions d'eines només s'avalua la classificació final de la variant, aquest enfocament les fa tenir millor rendiment, però això pot introduir biaixos i contravé el principi d'independència, clau en les recomanacions de ClinGen. Des del nostre punt de vista, tot i que aquestes eines poden ser útils per a la priorització de variants en un context de recerca, no són adequades per a ús diagnòstic, on l'aplicació de criteris ha de ser precisa i justificada, ja que condiciona posteriors decisions mèdiques.

En analitzar cada criteri hem observat que, malgrat els esforços d'ACMG/AMP i ClinGen per estandarditzar els criteris de les guies, alguns són encara ambigus i podrien causar discrepàncies

entre laboratoris. Quan s'han identificat aquestes ambigüitats, les hem etiquetat com a 'criteris refinats'. Alguns exemples inclouen l'ús de bases de dades de poblacions excloent el subconjunt de dades provinents de The Cancer Genome Atlas Program (TCGA) per evitar una sobrerepresentació de variants que predisposin al càncer i subpoblacions aïllades amb efecte fundador per calcular criteris relacionats amb les dades poblacionals, o la selecció de predictors i/o llindars en casos on les guies no els estableixen. En particular, s'ha utilitzat SpliceAI (Jaganathan *et al.*, 2019) com a predictor únic per esdeveniments d'empalmament, amb l'excepció dels gens MMR, on s'ha aplicat també el model de Prior (UTAH) d'acord amb les guies específiques. Tot i que algunes guies suggereixen puntuar les variants mitjançant un consens de dos o tres predictors, recentment s'ha demostrat que això aporta pocs avantatges (Wai *et al.*, 2020). Una comparativa d'eines de predicció d'empalament recomana l'ús de SpliceAI per la seva alta àrea sota la corba (Riepe *et al.*, 2021). També avalen l'ús d'aquest predictor en solitari les recents guies de ClinGen per la classificació de variants envers els efectes predits o trobats sobre l'empalmament (Walker *et al.*, 2023). Cal destacar que vaRHC utilitza tots els llindars proposats com a valors per defecte, però és fàcilment modificable mitjançant un fitxer de text.

Aquesta validació confirma que l'automatització de part del procés redueix l'error humà, ja que fins i tot ha assenyalat casos en els quals no s'havien aplicat correctament les guies en els conjunts de dades de referència. En el nostre article es proporciona una comparativa detallada de totes les variants utilitzades per la validació, amb explicacions específiques per als casos en que l'assignació dels criteris no coincideix amb la de ClinGen i, quan s'escau, suggeriments per a la curació manual dels criteris. Finalment, en la comparativa feta amb Cancer SIGVAR, el test de Kappa ha mostrat diferències significatives en l'assignació de criteris, afavorint vaRHC. Aquest fet és raonable si considerem que Cancer SIGVAR no ha estat actualitzat recentment i, per tant, utilitza una versió antiga de les guies i criteris no recomanats per ACMG.

Per tot això, considerem que vaRHC és rigurós en l'assignació de criteris, la qual cosa el fa especialment adequat per a l'ús diagnòstic.

Impacte de vaRHC

Des de la publicació de l'article, vaRHC s'ha incorporat com a eina de rutina a la unitat de diagnòstic molecular de l'ICO. S'utilitza per analitzar totes les noves variants i aquelles classificades fa més de dos anys. Als experts classificadors se'ls proporciona un full de càlcul (.xlsx) per variant amb els resultats preliminars, que revisen i completen amb dades no automatitzables abans de transferir la informació a Pandora, la base de dades del servei.

Inicialment, l'eina es va implementar exclusivament per classificar variants detectades en estudis de panell. Tanmateix, la seva bona acollida ha permès ampliar-ne l'ús a les variants d'estudis dirigits i a altres projectes, com el projecte iD.BRCA o el projecte de recerca de la suscetpibilitat genètica a patir una COVID greu.

L'eina utilitza codi obert disponible a GitHub (https://github.com/emunte/vaRHC), i tot i que no disposem d'informació per poder mesurar l'impacte de vaRHC fora del nostre centre, sabem que també ha estat implementada a l'Hospital del Mar i hem rebut consultes de professionals d'arreu del món interessats en l'eina.

Actualitzacions de vaRHC després de la publicació de l'article

Des de la publicació de l'article, vaRHC s'ha anat actualitzat per incorporar noves guies o versions millorades de les existents (Taula 19), actualment s'està utilitzant la versió 3.0.0.

A més, s'ha incorporat al full de càlcul (.xlsx) els suggeriments dels classificadors de variants de l'ICO, les dades de nous estudis funcionals i el predictor SpliceAI-10k (Canson *et al.*, 2023). També s'ha modificat part del codi per adaptar-se a actualitzacions de pàgines web sobre les quals es fa *web scrapping,* com és el cas de ClinVar.

Taula 19: Versions actuals de les guies ClinGen per a diversos gens implicats en càncer hereditari (a data de 4/11/2024), en comparació amb les versions de guies incorporades a vaRHC en el moment de la publicació de l'article i les incloses en la v3.0.0 de vaRHC

Gens	Versió actual guies ClinGen	vaRHC publicació article	vaRHC v.3.0.0
APC	v.2.1.0 (24/11/2023)	-	v.1.0.0 (1/10/2023)
ATM	v.1.3.0 (27/3/2024)	v1.1.0 (25/2/2022)	v1.1.0 (25/2/2022)
BRCA1	v.1.1.0 (21/12/2023)	-	v.1.0[1] (9/8/2023)
BRCA2	v.1.1.0 (21/12/2023)	-	v.1.0[1] (9/8/2023)
CDH1	v.3.1.0 (7/12/2022)	v.3.1.0 (7/12/2022)	v.3.1.0 (7/12/2022)
CHEK2	-	Vargas-Parra et al. (2020)Vargas-Parra et al. (2020)	Vargas-Parra et al. (2020)Vargas-Parra et al. (2020)
DICER1	v.1.3.0 (30/1/2024)	-	-
MLH1	v.1.0.0 (8/9/2024)	V1 (esborrany)[2] (11/2021)	V1 (esborrany)[2] (11/2021)
MSH2	v.1.0.0 (8/9/2024)	V1 (esborrany)[2] (11/2021)	V1 (esborrany)[2] (11/2021)
MSH6	v.1.0.0 (8/9/2024)	V1 (esborrany)[2]	V1 (esborrany)[2] (11/2021)
PALB2	v.1.1.0 (28/11/2023)	-	v.1.0.0 (28/11/2023)
PMS2	v.1.0.0 (8/9/2024)	V1 (esborrany)[2] (11/2021)	V1 (esborrany)[2] (11/2021)
PTEN	v.3.1.0 (14/3/2024)	v.2.0.0 (10/9/2019)	v.2.0.0 (10/9/2019)
RUNX1	v.2.0.0 (15/9/2021)	-	-
TP53	v.2.2.0 (30/9/2024)	v.1.2 (6/8/2019)	v.1.4.0 (5/7/2023)
VHL	v.1.0.0 (29/2/2024)	-	-

[1] Només s'han incorporat els criteris PVS1, PM2, PM5, BA1, BS1.

[2] Esborrany disponible a https://www.insight-group.org/wp-content/uploads/2021/11/DRAFT_Nov_2021_TEMPLATE_SVI.ACMG_Specifications_InSiGHT_MMR_V1.pdf.

Limitacions

En l'article original ja es van detallar algunes limitacions de vaRHC. Una d'elles és que l'eina no funciona per a tots els tipus ni llargades de variants. El seu temps d'execució és d'aproximadament 15–30 segons per variant, però pot arribar fins als 2 minuts per a insercions i delecions en què SpliceAI no es pot precalcular. A més, la connexió de vaRHC a bases de dades com ClinVar depèn del bon rendiment i de la connectivitat dels seus llocs web. Aquesta dependència del *web scraping* implica un manteniment continuat, ja que qualsevol canvi en l'estructura HTML de les pàgines consultades pot provocar errors. Finalment, la base de dades de vaRHC conté informació sobre alguns assajos funcionals publicats, però l'actualització és manual, i no pot fer una extracció automàtica de la literatura.

Tot i que s'ha fet un esforç continuat per anar incorporant noves guies, la involucració en altres projectes dins aquesta tesi doctoral ha fet impossible mantenir vaRHC completament al dia. En aquest sentit, serà imprescindible comptar amb algú dedicat en bona mesura a l'actualització de les guies i els criteris, destacant la necessitat d'eines flexibles que s'adaptin ràpidament a l'evolució de les guies i a les exigències del context diagnòstic. Una dedicació plena al manteniment i evolució de vaRHC asseguraria que l'eina estigui sempre en línia amb les guies més actualitzades i segueixi incrementant el rendiment del servei en la demandant tasca de la classifcació de variants.

Perspectives de futur de vaRHC

Amb l'experiència de dos anys de funcionament, hem constatat que vaRHC és una eina molt útil per al diagòstic, però que alhora presenta un significatiu marge de millora. En particular, la dependència de vaRHC d'elements externs, com bases de dades i llocs web a través de web *scraping*, ha resultat ser un obstacle per garantir la seva integració dins de Pandora, la base de dades de variants del SDMCH. Aquests recursos externs poden fallar perquè el servidor cau o dona temps de resposta molt llargs. També, l'especificació de l'API o l'estructura de la web poden canviar, la qual cosa obliga a modificar el codi de vaRHC que accedeix a aquests recursos. Aquestes incidències poden comprometre la fiabilitat del servei i dificulten el seu ús continuat en un entorn clínic on la robustesa és essencial.

A més, l'ús d'informació consultada a temps real planteja problemes de reproductibilitat, ja que des de vaRHC no es pot controlar que un recurs extern doni el mateix resultat en dos moments diferents. Aquesta manca de consistència és especialment preocupant en el context de les ISOs de diagnòstic, on la reproductibilitat és un requisit.

Per abordar aquestes limitacions, hem iniciat el desenvolupament d'una segona versió de vaRHC amb l'objectiu d'eliminar totes les dependències externes. Aquest projecte implica encapsular totes les funcionalitats dins d'una imatge Docker, que inclourà totes les eines necessàries eliminant així la necessitat de connectar-se a serveis externs.

El nostre pla per aquesta nova versió inclou la creació d'una API per al Docker de vaRHC, que s'integrarà directament a Pandora. Aquesta imatge dockeritzada incorporarà programes com VEP per a l'anotació de variants i SpliceAI per a prediccions d'empalmament. Pel que fa a fonts com gnomAD i ClinVar, encara s'està valorant si s'integraran en forma de bases de dades internes o si es gestionaran com a fitxers comprimits que només incloguin la informació necessària. En el cas de ClinVar, es preveu una actualització periòdica automàtica per mantenir les dades actualitzades sense sobrecarregar el sistema. Totes les versions usades quedaran registrades, per poder garantir la reproductibilitat. Això permetrà als usuaris seleccionar fàcilment un conjunt de variants per ser classificades o reclassificades i generar els informes en format Excel de manera directa i eficient, sense necessitat de sol·licitar aquesta execució a un bioinformàtic, tal com es fa actualment. A més,

aquesta API estarà disponible per a usuaris externs, facilitant la integració de vaRHC en *pipelines* diagnòstics de tercers, és a dir, es facilitarà la seva integració a qualsevol sistema emprat per altres centres. Per tal d'ampliar l'accessibilitat, també es desenvoluparà una aplicació web, que permetrà a usuaris no experts en bioinformàtica utilitzar l'eina amb facilitat.

Amb aquestes millores, esperem consolidar vaRHC com una referència en la classificació automatitzada de variants, fomentant el seu ús tant dins com fora del nostre centre.

Noves eines per classificar variants

Malgrat que han passat gairebé dos anys des de la publicació de l'article, no tenim notícia que hagi sorgit cap programa que utilitzi les guies gen-específiques per a càncer hereditari, fet que posa de manifest la dificultat de mantenir les eines constantment actualitzades.

Paral·lelament, la intel·ligència artificial (IA) està cada cop més present, de manera que no sorprèn que comencin a aparèixer programes de classificació de variants que en facin ús, una tendència que probablement es continuarà expandint. L'ús d'IA per optimitzar alguns criteris o ajudar en l'obtenció de dades pot ser molt positiu, ja que pot agilitzar la recollida d'evidències i millorar la precisió en l'aplicació d'aquests criteris, com ja es veu en casos com SpliceAI per a la predicció d'esdeveniments d'empalmament. Per exemple, podria ser beneficiós per l'aplicació d'evidències predictives. Ara bé, cal assegurar que els conjunts d'entrenament i la informació que nodreixen aquests models no solapin d'altres criteris, de manera que es mantingui el supòsit d'independència que requereix la combinació bayesiana de les guies ACMG. Per altra banda, no tots els laboratoris disposen d'expertesa suficient per avaluar la validesa dels càlculs interns, de manera que cal confiar en un procés de validació rigorós. A més, la majoria d'aquestes eines són de pagament, la qual cosa en limita l'accessibilitat i fa que el cost d'incorporar-les de manera continuada sigui un altre factor a considerar.

Per tant, des del nostre punt de vista, si es garanteix la qualitat de les dades d'entrenament, una validació rigorosa, la independència dels criteris i la possibilitat de traçar els passos que duen a la classificació final, les eines basades en IA poden resultar molt valuoses més enllà de l'entorn de recerca, afavorint una classificació més ràpida i precisa també en l'àmbit clínic.

CONCLUSIONS

- S'ha desenvolupat l'algoritme d'anàlisi anomenat PMS2_vaR per optimitzar l'anàlisi mutacional del gen *PMS2* a partir de dades de panells d'NGS. La seva implementació ha permès reduir temps i recursos, ja que només es validen experimentalment les mostres amb variants patogèniques o probablement patogèniques candidates i la PCR es limita a la regió especifica. Actualment, PMS2_vaR s'utilitza al SDMCH per analitzar mostres amb pèrdua exclusiva de PMS2, facilitant la identificació de portadors. L'èxit de PMS2_vaR suggereix la possibilitat d'adaptar-lo a altres gens amb pseudogens tot ajustant el codi pertinentment.

- L'anàlisi global detallada d'una bateria d'eines de detecció de CNVs germinals indica que algunes, com ClinCNV i GATK-gCNV, tenen un rendiment diagnòstic superior. En concret, GATK-gCNV destaca per un excel·lent equilibri entre sensibilitat i especificitat. Els resultats de la comparativa realitzada, juntament amb l'avaluació de paràmetres, la combinació d'eines per parelles i el marc de treball CNVbenchmarkeR2, proporcionen una guia valuosa per optimitzar aquestes eines en contextos clínics concrets.

- L'aplicació de GRIDSS ha millorat el rendiment diagnòstic un 0,6%, en identificar vuit variants estructurals patogèniques o probablement patogèniques que estudis previs no havien detectat. L'algoritme presenta una alta especificitat i càrrega manual moderada amb els filtres emprats.

- L'exploració de variants a regions 5'UTR ha identificat variants genètiques interessats en el context del diagnòstic del càncer hereditari. Una d'elles, *CDKN2A* c.-34G>T, s'havia descrit prèviament com a patogènica i, s'ha identificat en tres probands. Tanmateix calen esforços per homogenitzar la classificació de variants en aquestes regions així com l'estandardització d'estudis funcionals per a la seva correcta classificació.

- S'ha desenvolupat vaRHC, un paquet d'R que semiautomatitza la classificació de variants seguint les guies ACMG així com guies específiques de gen de ClinGen. VaRHC redueix temps de classificació, limita errors i permet personalitzar paràmetres, però no substitueix l'especialista, clau per integrar dades clíniques i genètiques. Els resultats de vaRHC es poden exportar en format xlsx, facilitant la incorporació de criteris no automatitzats.

REFERÈNCIES

A

Abdellah, Z. et al. (2004) Finishing the euchromatic sequence of the human genome. Nature 2004 431:7011, 431(7011), pp. 931–945.

Abou Tayoun, A.N. et al. (2018) Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat, 39(11), pp. 1517–1524.

Amaral, P. et al. (2023) The status of the human gene catalogue. Mark Yandell, 622, p. 19.

Anand, P. et al. (2008) Cancer is a preventable disease that requires major lifestyle changes. Pharm Res, 25(9), pp. 2097–2116.

Andreev, D.E. et al. (2022) Non-AUG translation initiation in mammals. Genome Biol, 23(1), pp. 1–17.

Ao-Kondo, H. et al. (2011) Emergence of the Diversified Short ORFeome by Mass Spectrometry-Based Proteomics. Computational Biology and Applied Bioinformatics [Preprint].

B

Bahrambeigi, V. et al. (2016) An Approach for Accurate Molecular Diagnosis of Highly Homologous SDHA Gene. Am J Clin Pathol, 146(suppl_1), pp. 94–100.

Bailey, M.H. et al. (2018) Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell, 173(2).

Baltimore, D. (1970) Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature, 226(5252).

Barbitoff, Y. and Predeus, A. (2024) Negligible effects of read trimming on the accuracy of germline short variant calling in the human genome [version 1; peer review: awaiting peer review]. F1000Res, 13, p. 506.

Barbitoff, Y.A. et al. (2024) Bioinformatics of germline variant discovery for rare disease diagnostics: current approaches and remaining challenges. Brief Bioinform, 25(2).

Biesecker, L.G. et al. (2018) The ACMG/AMP reputable source criteria for the interpretation of sequence variants. Genet Med, 20(12), p. 1687.

Brnich, S.E. et al. (2019) Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med, 12(1), pp. 1–12.

C

Calvo, S.E. et al. (2009a) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A, 106(18), p. 7507.

Calvo, S.E. et al. (2009b) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A, 106(18), p. 7507.

Cameron, D.L. et al. (2017a) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res, 27(12), pp. 2050–2060.

Cameron, D.L. et al. (2017b) GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res, 27(12), pp. 2050–2060.

Canson, D.M. et al. (2023) SpliceAI-10k calculator for the prediction of pseudoexonization, intron retention, and exon deletion. Bioinformatics. Edited by C. Kendziorski, pp. 0–0.

Castellanos, E. et al. (2017) A comprehensive custom panel design for routine hereditary cancer testing: preserving control, improving diagnostics and revealing a complex variation landscape. Scientific Reports 2017 7:1, 7(1), pp. 1–12.

Chen, X. et al. (2020) Re-recognition of pseudogenes: From molecular to clinical applications. Theranostics, 10(4), pp. 1479–1499.

Chénais Biosse, B. (2022) Transposable Elements and Human Diseases: Mechanisms and Implication in the Response to Environmental Pollutants. Int. J. Mol. Sci, 2022, p. 2551.

Chothani, S.P. et al. (2022) A high-resolution map of human RNA translation. Mol Cell, 82(15), pp. 2885-2899.e8.

Clendenning, M. et al. (2006) Long-range PCR facilitates the identification of PMS2-specific mutations. Hum Mutat, 27(5), pp. 490–495.

Cooke, D.P. et al. (2021) A unified haplotype-based method for accurate and comprehensive variant calling. Nature Biotechnology 2021 39:7, 39(7), pp. 885–892.

D

De Arce, A.J.D. et al. (2018a) Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. Nucleic Acids Res, 46(2), p. 985.

de la Hoya, M. et al. (2016) Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Hum Mol Genet, 25(11).

E

Eichler, E.E. (2019) Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. New England Journal of Medicine, 381(1), pp. 64–74.

Ellingford, J.M. et al. (2022) Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med, 14(1).

Escaramís, G. et al. (2015) A decade of structural variants: description, history and methods to detect structural variation.

Etzler, J. et al. (2008) RNA-based mutation analysis identifies an unusual MSH6 splicing defect and circumvents PMS2 pseudogene interference. Hum Mutat, 29(2), pp. 299–305.

F

Feliubadaló, L. et al. (2019) Opportunistic testing of BRCA1, BRCA2 and mismatch repair genes improves the yield of phenotype driven hereditary cancer gene panels. Int J Cancer, 145(10), pp. 2682–2691.

G

Gabrielaite, M. et al. (2021) A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. Cancers (Basel), 13(24), p. 6283.

Ganster, C. et al. (2010) Functional PMS2 hybrid alleles containing a pseudogene-specific missense variant trace back to a single ancient intrachromosomal recombination event. Hum Mutat, 31(5), pp. 552–560.

Garrett, A. et al. (2021) Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: CanVIG-UK consensus recommendations. J Med Genet, 58(5), pp. 297–304.

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing.

Gelb, B.D. et al. (2018) ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. Genetics in Medicine, 20(11), pp. 1334–1345.

Genovese, G. et al. (2014) Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. New England Journal of Medicine, 371(26), pp. 2477–2487.

Gould, G.M. et al. (2018) Detecting clinically actionable variants in the 3' exons of PMS2 via a reflex workflow based on equivalent hybrid capture of the gene and its pseudogene 06 Biological Sciences 0604 Genetics. BMC Med Genet, 19(1), pp. 1–13.

H

Hanahan, D. (2022) Hallmarks of Cancer: New Dimensions. Cancer Discov, 12(1), pp. 31–46.

Hanahan, D. and Weinberg, R.A. (2000) The Hallmarks of Cancer. Cell, 100(1), pp. 57–70.

Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. Cell, 144(5), pp. 646–674.

Hayward, B.E. et al. (2007) Extensive gene conversion at the PMS2 DNA mismatch repair locus. Hum Mutat, 28(5), pp. 424–430.

Hogrefe, H.H. and Borns, M.C. (2011) Long-Range PCR with a DNA Polymerase Fusion.in Methods in Molecular Biology.

I

Ignatov, K.B. et al. (2014) A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. Biotechniques, 57(2).

International Human Genome Sequencing Consortium* (2004) Finishing the euchromatic sequence of the human genome, 431(7011), pp. 931–945.

J

Jaganathan, K. et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), pp. 535-548.e24.

Jahn, A. et al. (2022) Comprehensive cancer predisposition testing within the prospective MASTER trial identifies hereditary cancer patients and supports treatment decisions for rare cancers. Annals of Oncology, 33(11).

J.E., G. and K., O. (2005) Hereditary cancer predisposition syndromes. Journal of Clinical Oncology.

Johansson, L.F. et al. (2016) CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. Hum Mutat, 37(5), pp. 457–464.

Joseph, V. et al. (2017) Pathogenicity of mutation analyzer (PathoMAN): A fast automation of germline genomic variant curation in clinical sequencing. https://doi.org/10.1200/JCO.2017.35.15_suppl.1529, 35(15_suppl), pp. 1529–1529.

K

Kalia, S.S. et al. (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genetics in Medicine, 19, pp. 249–255.

Kamps, R. et al. (2017) Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. Int J Mol Sci. MDPI AG, p. 308.

Kim, S. et al. (2018) Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods 2018 15:8, 15(8), pp. 591–594.

Klein, J. et al. (2023) A Novel Alu Element Insertion in ATM Induces Exon Skipping in Suspected HBOC Patients. Hum Mutat, 2023.

van der Klift, H.M. et al. (2016) Comprehensive Mutation Analysis of PMS2 in a Large Cohort of Probands Suspected of Lynch Syndrome or Constitutional Mismatch Repair Deficiency Syndrome. Hum Mutat, 37(11), pp. 1162–1179.

Knudson, A.G. (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. Proc Natl Acad Sci U S A, 68(4), p. 820.

Koboldt, D.C. et al. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25(17), p. 2283.

Koboldt, D.C. (2020) Best practices for variant calling in clinical sequencing. Genome Med.

Kopanos, C. et al. (2019) VarSome: the human genomic variant search engine. Bioinformatics, 35(11), pp. 1978–1980.

Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell, 44(2), pp. 283–292.

Kozak, M. (1987a) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res, 15(20), pp. 8125–8148.

Kozak, M. (1987b) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol, 196(4), pp. 947–950.

Kozak, M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J, 16(9), pp. 2482–2492.

L

Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. Nature 2001 409:6822, 409(6822), pp. 860–921.

Landrum, M.J. et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res, 46(D1), pp. D1062–D1067.

Lee, K. et al. (2018) Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. Hum Mutat, 39(11), pp. 1553–1568.

Lepkes, L. et al. (2021a) Performance of In Silico Prediction Tools for the Detection of Germline Copy Number Variations in Cancer Predisposition Genes in 4208 Female Index Patients with Familial Breast and Ovarian Cancer. Cancers (Basel), 13(1), pp. 1–12.

Leppek, K. et al. (2018) Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol, 19(3), p. 158.

Levy-Sakin, M. et al. (2019) Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun, 10(1), pp. 1–14.

Li, H. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), p. 2078.

Li, H. et al. (2021) Cancer SIGVAR: A semiautomated interpretation tool for germline variants of hereditary cancer-related genes. Hum Mutat, 42(4), pp. 359–372.

Li, H. and Wren, J. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics, 30(20), pp. 2843–2851.

Li, Q. and Wang, K. (2017) InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet, 100, pp. 267–280.

Liao, W.W. et al. (2023) A draft human pangenome reference. Nature 2023 617:7960, 617(7960), pp. 312–324.

Lincoln, S.E. et al. (2021) One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genetics in Medicine 2021 23:9, 23(9), pp. 1673–1680.

Liu, L. et al. (1999) Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. Nature Genetics 1999 21:1, 21(1), pp. 128–132.

Liu, S. et al. (2016) Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res, 44(5), pp. e47–e47.

Luo, X. et al. (2019) ClinGen Myeloid Malignancy Variant Curation Expert Panel recommendations for germline RUNX1 variants. Blood Adv, 3(20), pp. 2962–2979.

M

Mahmoud, M. et al. (2019) Structural variant calling: The long and the short of it. Genome Biol, 20(1), pp. 1–14.

Maniam, P. et al. (2018) Pathogenicity and Penetrance of Germline SDHA Variants in Pheochromocytoma and Paraganglioma (PPGL). J Endocr Soc, 2(7).

Manolio, T.A. et al. (2009) Finding the missing heritability of complex diseases. Nature 2009 461:7265, 461(7265), pp. 747–753.

Martin-Geary, Alexandra C. et al. (2023) Systematic identification of disease-causing promoter and untranslated region variants in 8,040 undiagnosed individuals with rare disease. medRxiv [Preprint].

Mavaddat, N. et al. (2019) Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet, 104(1).

McKenna, A. et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20(9), pp. 1297–1303.

McLaren, W. et al. (2016) The Ensembl Variant Effect Predictor. Genome Biol, 17(1), pp. 1–14.

Melidis, D.P. et al. (2022) GenOtoScope: Towards automating ACMG classification of variants associated with congenital hearing loss. PLoS Comput Biol, 18(9), p. e1009785.

Mester, J.L. et al. (2018) Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. Hum Mutat, 39(11), pp. 1581–1592.

Mignone, F. et al. (2002) Untranslated regions of mRNAs. Genome Biol.

Miller, D.T. et al. (2023) ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med, 25(8).

Moorthie, S. et al. (2013) Informatics and clinical genome sequencing: opening the black box. Genetics in Medicine, 15(3), pp. 165–171.

Moreno-Cabrera, J.M. et al. (2020) Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. European Journal of Human Genetics, 28(12), pp. 1645–1655.

Moreno-Cabrera, J.M. et al. (2022) Screening of CNVs using NGS data improves mutation detection yield and decreases costs in genetic testing for hereditary cancer. J Med Genet, 59(1), pp. 75–78.

Moreno-Cabrera, J.M. et al. (2024) SpadaHC: a database to improve the classification of variants in hereditary cancer genes in the Spanish population. Database, 2024.

Morris, D.R. and Geballe, A.P. (2000) Upstream Open Reading Frames as Regulators of mRNA Translation, 20(23), pp. 8635–8642.

Mount, S.M. (1982) A catalogue of splice junction sequences. Nucleic Acids Res, 10(2), p. 459.

N

Nakagawa, H. et al. (2004) Mismatch repair gene PMS2: disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation. Cancer Res, 64(14), pp. 4721–4727.

Nakken, S. et al. (2021) Cancer Predisposition Sequencing Reporter (CPSR): A flexible variant report engine for high-throughput germline screening in cancer. Int J Cancer, 149(11), pp. 1955–1960.

Nicolaides, N.C. et al. (1995) Genomic Organization of the HumanPMS2Gene Family. Genomics, 30(2), pp. 195–206.

Nicora, G. et al. (2018) CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. Hum Mutat, 39(12), pp. 1835–1846.

Noderer, W.L. et al. (2014) Quantitative analysis of mammalian translation initiation sites by FACS - seq . Mol Syst Biol, 10(8), p. 748.

Nurk, S. et al. (2022) The complete sequence of a human genome. Science (1979), 376(6588), pp. 44–53.

Nykamp, K. et al. (2017) Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. Genetics in Medicine, 19, pp. 1105–1117.

O

Olson, N.D. et al. (2023) Variant calling and benchmarking in an era of complete human genome sequences. Nature Reviews Genetics 2023 24:7, 24(7), pp. 464–483.

Oza, A.M. et al. (2018) Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. Hum Mutat, 39(11), pp. 1593–1613.

P

Pakay, J. et al. (2023) Threshold Concepts in Biochemistry. Thereshold Concepts in Biochemistry [Preprint].

Patel, R.Y. et al. (2017) ClinGen Pathogenicity Calculator: A configurable system for assessing pathogenicity of genetic variants. Genome Med, 9(1), pp. 1–9.

Pedersen, B.S. and Quinlan, A.R. (2017) Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am J Hum Genet, 100(3), p. 406.

Pejaver, V. et al. (2022) Evidence-based calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for clinical use of PP3/BP4 criteria. bioRxiv, p. 2022.03.17.484479.

Peng, J. et al. (2021) VIP-HL: Semi-automated ACMG/AMP variant interpretation platform for genetic hearing loss. Hum Mutat, 42(12), pp. 1567–1575.

Perales, R. and Bentley, D. (2009) 'Cotranscriptionality': the transcription elongation complex as a nexus for nuclear transactions. Mol Cell, 36(2), pp. 178–191.

Pesole, G. et al. (2001) Structural and functional features of eukaryotic mRNA untranslated regions. Gene, 276(1–2), pp. 73–81.

Pinto, D. et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 2010 466:7304, 466(7304), pp. 368–372.

Plon, S.E. et al. (2008) Sequence Variant Classification and Reporting: Recommendations for Improving the Interpretation of Cancer Susceptibility Genetic Test Results. Hum Mutat, 29(11).

Poliseno, L. et al. (2024) Coding, or non-coding, that is the question. Cell Res, 0, pp. 1–21.

Poplin, R. et al. (2018) A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 2018 36:10, 36(10), pp. 983–987.

Q

Qian, Y. et al. (2017) Identification of pathogenic retrotransposon insertions in cancer predisposition genes. Cancer Genet, 216–217.

R

Richards, S. et al. (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5), pp. 405–424.

Riepe, T. V. et al. (2021) Benchmarking deep learning splice prediction tools using functional splice assays. Hum Mutat, 42(7), pp. 799–810.

Riggs, E.R. et al. (2020) Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genetics in Medicine, 22(2).

Roca, I. et al. (2019) Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutation Research/Reviews in Mutation Research, 779, pp. 114–125.

Rodríguez-Santiago, B. and Armengol, L. (2012) Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal. Diagnostico Prenatal, 23(2).

Roy, A.L. (2005) Core Promoters. Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine, pp. 338–341.

S

Samarakoon, P.S. et al. (2014) Identification of copy number variants from exome sequence data. BMC Genomics, 15(1), pp. 1–11.

Santarosa, M. and Ashworth, A. (2004) Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1654(2), pp. 105–122.

Scarano, C. et al. (2024) The Third-Generation Sequencing Challenge: Novel Insights for the Omic Sciences. Biomolecules 2024, Vol. 14, Page 568, 14(5), p. 568.

Schmidt, R.J. et al. (2024) Recommendations for risk allele evidence curation, classification, and reporting from the ClinGen Low Penetrance/Risk Allele Working Group. Genetics in Medicine, 26(3).

Schulz, J. et al. (2018) Loss-of-function uORF mutations in human malignancies. Sci Rep, 8(1).

Schwenk, V. et al. (2023) Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. J Med Genet, 60(8), pp. 747–759.

Scott, A.D. et al. (2019) CharGer: clinical Characterization of Germline variants, 35(5), pp. 865–867.

Sedlazeck, F.J. et al. (2018) Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods 2018 15:6, 15(6), pp. 461–468.

Senter, L. et al. (2008) The clinical phenotype of Lynch syndrome due to germline PMS2 mutations. Gastroenterology, 135(2), p. 419.

Shafee, T. and Lowe, R. (2017) Eukaryotic and prokaryotic gene structure, 4(1).

Shaikh, T.H. (2017) Copy Number Variation Disorders. Curr Genet Med Rep, 5(4), p. 183.

Shlien, A. and Malkin, D. (2009) Copy number variations and cancer. Genome Med, 1(6), pp. 1–9.

Singh, R.R. (2022) Target Enrichment Approaches for Next-Generation Sequencing Applications in Oncology. Diagnostics, 12(7).

Smith, S.D. et al. (2017) Lightning-fast genome variant detection with GROM. Gigascience.

Spurdle, A.B. et al. (2012) OFFICIAL JOURNAL ENIGMA-Evidence-Based Network for the Interpretation of Germline Mutant Alleles: An International Initiative to Evaluate Risk and Clinical Significance Associated with Sequence Variation in BRCA1 and BRCA2 Genes. Hum Mutat, 33, pp. 2–7.

Stawiński, P. and Płoski, R. (2024) Genebe.net: Implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. Clin Genet, 106(2), pp. 119–126.

T

Tabori, U. et al. (2017) Clinical Management and Tumor Surveillance Recommendations of Inherited Mismatch Repair Deficiency in Childhood. Clinical Cancer Research, 23(11), pp. e32–e37.

Tarasov, A. et al. (2015) Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31(12), pp. 2032–2034.

Tavtigian, S. V. et al. (2018) Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med, 20(9), pp. 1054–1060.

Tavtigian, S. V. et al. (2020) Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. Hum Mutat, 41(10), pp. 1734–1737.

Thompson, B.A. et al. (2014) Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet. Nature Publishing Group, pp. 107–115.

Thung, D.T. jwan *et al.* (2014) Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol, 15(10).

Torene, R.I. *et al.* (2020) Mobile element insertion detection in 89,874 clinical exomes. Genetics in Medicine, 22(5).

V

Van der Auwera, G.A. et al. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics, 43(1), pp. 11.10.1-11.10.33.

Vargas-Parra, G. et al. (2020) Comprehensive analysis and ACMG-based classification of CHEK2 variants in hereditary cancer patients. Hum Mutat, 41(12), pp. 2128–2142.

Vaughn, C.P. et al. (2010) Clinical analysis of PMS2: mutation detection and avoidance of pseudogenes. Hum Mutat, 31(5), pp. 588–593.

Vogelstein, B. et al. (2013) Cancer Genome Landscapes. Science, 339(6127), p. 1546.

W

Wagner, J. et al. (2022) Benchmarking challenging small variants with linked and long reads. Cell Genomics, 2(5), p. 100128.

Wai, H.A. et al. (2020) Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genetics in Medicine, 22(6), pp. 1005–1014.

Walker, L.C. et al. (2023) Application of the ACMG/AMP framework to capture evidence relevant to predicted and observed impact on splicing: recommendations from the ClinGen SVI splicing. medRxiv [Preprint].

Wang, K. et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 38(16), pp. e164–e164.

Wang, N. et al. (2022) Tool evaluation for the detection of variably sized indels from next generation whole genome and targeted sequencing data. PLoS Comput Biol, 18(2), p. e1009269.

Weill, L. et al. (2012) Translational control by changes in poly(A) tail length: recycling mRNAs. Nature Structural & Molecular Biology 2012 19:6, 19(6), pp. 577–585.

Weinberg, R.A. (2013) The Biology of Cancer.

Wheeler, E. et al. (2013) Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nature Genetics 2013 45:5, 45(5), pp. 513–517.

Whiffin, N. et al. (2018) CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. Genetics in Medicine 2018 20:10, 20(10), pp. 1246–1254.

Wicker, T. et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet, 8(12), pp. 973–982.

Wieder, N. et al. (2024) Differences in 5'untranslated regions highlight the importance of translational regulation of dosage sensitive genes. Genome Biol, 25(111).

Wimmer, K. et al. (2014) Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'Care for CMMRD' (C4CMMRD) on behalf of the EU-Consortium Care for CMMRD (C4CMMRD). J Med Genet, 51, pp. 283–293.

Win, A.K. et al. (2017) Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer. Cancer Epidemiol Biomarkers Prev, 26(3), p. 404.

Wright, C.F. et al. (2021) Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. Am J Hum Genet, 108(6), pp. 1083–1094.

X

Xavier, A. et al. (2019) TAPES: A tool for assessment and prioritisation in exome studies. PLoS Comput Biol, 15(10).

Y

Yan Lim, C. et al. (1999) The MTE, a new core promoter element for transcription by RNA polymerase II. Brivanlou and Darnell [Preprint].

Yang, Y. and Wang, Z. (2019) IRES-mediated cap-independent translation, a path leading to hidden proteome. J Mol Cell Biol, 911(10), pp. 911–919.

Z

Zarrei, M. et al. (2015) A copy number variation map of the human genome. Nature Reviews Genetics 2015 16:3, 16(3), pp. 172–183.

Zhang, X. et al. (2021) Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics, 37(8), pp. 1171–1173.

Zhao, Z. et al. (2022) STI PCR: An efficient method for amplification and de novo synthesis of long DNA sequences. Mol Plant, 15(4).

ANNEX

Taules suplementàries

Taula Suplementària 1: Exemples d'eines per detectar SVs.							
Eina	**Any creació**	**Disponibilitat**	**Dades crues**	**RD**	**SR**	**PEM**	**A**
AGE	2011	http://archive2.gersteinlab.org/proj/sv/age/	WGS		X		
Atlas-CNV	2019	https://github.com/theodorc/Atlas-CNV	Panell	X			
BIC-seq2	2016	https://github.com/ding-lab/BICSEQ2	WGS	X			
BreakDancer	2014	https://github.com/genome/breakdancer	WGS			X	
BreaKmer	2015	https://github.com/ccgd-profile/BreaKmer	Panell		X		X
CANOES	2014	https://github.com/ShenLab/CANOES	WES	X			
Canvas	2015	https://github.com/Illumina/canvas	WGS/WES	X			
CLAMMS	2015	https://github.com/rgcgithub/clamms	WES	X			
clearCNV	2022	https://github.com/bihealth/clear-cnv	panel	X			
Clever-sv	2012	https://bitbucket.org/tobiasmarschall/clever-toolkit/src/master/	WGS		X	X	
ClinCNV	2022	https://github.com/imgag/ClinCNV/	WGS/WES/panell	X			
ClipCrop	2011	https://github.com/shinout/clipcrop	NA		X		
CMDS	2010	https://github.com/ding-lab/cmds	WGS	X			
Cn.MOPS	2012	https://bioconductor.org/packages/release/bioc/html/cn.mops.html	WGS	X			
CNVer	2010	http://compbio.cs.toronto.edu/CNVer/	WGS	X		X	
cnvHiTSeq	2012	https://bioinformaticshome.com/tools/cnv/descriptions/cnvHiTSeq.html#gsc.tab=0	WGS	X	X	X	
CNVkit	2016	https://github.com/etal/cnvkit	WES/panell	X			
CNVnator	2015	https://github.com/abyzovlab/CNVnator	WGS	X			
cnvOffSeq	2014	https://sourceforge.net/projects/cnvoffseq/	WGS	X			
CNVPanelizer		https://www.bioconductor.org/packages/release/bioc/html/CNVPanelizer.html	WES	X			
CNV-Z	2023	https://github.com/eol017/CNV-Z	WES/panell	X			
CNVind	2022	https://github.com/wkusmirek/CNVind	WES	X			
Cobalt	2022	https://github.com/ARUP-NGS/cobalt	Panell	X			
CODEX	2015	https://www.bioconductor.org/packages/devel/bioc/html/CODEX.html	WES	X			
CODEX2	2018	https://github.com/yuchaojiang/CODEX2	WES/panell	X			
CONDEX	2011	https://code.google.com/archive/p/condr/	WES	X			
CoNIFER	2012	https://github.com/nkrumm/CoNIFER	WES	X			
CONTRA	2012	https://contra-cnv.sourceforge.net/	WES	X			
Control-FREEC	2012	https://boevalab.inf.ethz.ch/FREEC/	WGS/WES	X			
CoNVaDING	2016	https://github.com/molgenis/CoNVaDING	WES/panell	X			
CONVector	2015	https://github.com/parseq/convector	Panell	X			
CopyWriteR	2015	https://github.com/PeeperLab/CopywriteR	WES/Panell	X			
Cortex assembler	2008	https://cortexassembler.sourceforge.net/	WGS				X
Custom pipeline	2021	https://github.com/ash9nov/Target-panel-based-CNV-detection	Panell	X			
DECoN	2016	https://github.com/RahmanTeam/DECoN	Panell	X			
DELLY	2012	https://github.com/dellytools/delly	WGS		X	X	
ERDS	2012	https://github.com/igm-team/ERDS	WGS	X			

EXCAVATOR	2013	https://sourceforge.net/projects/excavatortool	WES	X			
ExomeCNV	2012	https://github.com/cran/ExomeCNV	WES	X			
ExomeCopy	2012	https://www.bioconductor.org/packages/2.10/bioc/html/exomeCopy.html	WES	X			
ExomeDepth	2012	https://github.com/vplagnol/ExomeDepth	WES	X			
FermiKit	2015	https://github.com/lh3/fermikit	WGS				X
FishingCNV	2013	https://sourceforge.net/projects/fishingcnv/	WES	X			
GASV	2009	https://bio.tools/gasv	WGS			X	
GASVPro	2012	https://code.google.com/archive/p/gasv/	WGS	X		X	
GATK-gCNV	2020	https://gatk.broadinstitute.org/hc/en-us	WES	X			
Genome STRiP 2.0	2015	https://software.broadinstitute.org/software/genomestrip/	WGS	X		X	
Gindel	2014	https://sourceforge.net/projects/gindel/	WGS	X	X	X	
GRIDSS	2017	https://github.com/PapenfussLab/gridss	WGS		X	X	X
HadoopCNV	2017	https://github.com/WGLab/HadoopCNV	WGS	X			
Hydra-Multi	2014	https://github.com/arq5x/Hydra	WGS			X	X
iCopyDAV	2016	https://github.com/vogetihrsh/icopydav	WGS	X			
inGAP-sv	2011	https://ingap.sourceforge.net/	WGS	X		X	
JointSLM	2011	https://bioinformaticshome.com/tools/cnv/descriptions/JointSLM.html#gsc.tab=0	WGS	X			
LUMPY	2014	https://github.com/arq5x/lumpy-sv	WGS	X	X	X	
Magnolya	2012	https://sourceforge.net/projects/magnolya/	WGS				X
Manta	2016	https://github.com/Illumina/manta	WGS		X	X	X
mrCaNaVar	2018	https://github.com/BilkentCompGen/mrcanavar	WGS	X			
panelcn.MOPS	2017	https://bioconductor.org/packages/release/bioc/html/panelcn.mops.html	Panell	x			
PatternCNV	2014	https://bioinformaticstools.mayo.edu/research/patterncnv/	WES	X			
PEcnv	2022	https://github.com/Sherwin-xjtu/PEcnv	WGS/WES/panell	X			
Pindel	2009	https://gmt.genome.wustl.edu/packages/pindel/	WGS		X		
RDXplorer	2011	https://rdxplorer.sourceforge.net/	WGS	X			
ReadDepth	2011	https://github.com/chrisamiller/readDepth	WGS	X			
RSICNV	2013	https://github.com/yhwu/rsicnv	WGS	X			
SavvyCNV	2022	https://github.com/rdemolgen/SavvySuite	WES/panell	X			
SegSeq	2009	https://github.com/youngmook/SegSeq	WGs	X			
SVDetect	2010	http://svdetect.sourceforge.net/	WGS	X		X	
SVseq2		https://sourceforge.net/projects/svseq2/files/SVseq2_2/	WGS		X	X	
TARDIS	2017	https://github.com/BilkentCompGen/tardis	WGS	X	X	X	
TIGRA-SV	2013	https://bioinformatics.mdanderson.org/public-software/archive/tigra/	WGS				X
VarScan2	2012	https://github.com/Jeltje/varscan2	WES	X			
VisCap	2016	https://github.com/pughlab/VisCap	WES/Panell	X			
XCAVATOR	2017	http://sourceforge.net/projects/xcavator/	WGS	X			
XHMM	2017	https://github.com/RRafiee/XHMM	WES	X			

En la taula es pot trobar informació de l'any de creació, on trobar l'eina, per quines dades crues s'ha creat i quina estratègia utilitza per detectar les SVs. Només s'han incorporat les eines que actualment es troben accessibles (revisió a data de novembre 2024). Abreviatures: RD: profunditat de cobertura; SR: lectures dividides, PE: mapeig extrems aparellats; A: assemblatge , ML: aprenentatge automàtic (machine learning).

Publicacions addicionals

BARD1 pathogenic variants are associated with triple-negative breast cancer in a spanish hereditary breast and ovarian cancer cohort

Paula Rofes, Jesús Del Valle, Sara Torres-Esquius, Lídia Feliubadaló, Agostina Stradella , José Marcos Moreno-Cabrera, Adriana López-Doriga , Elisabet Munté, Rafael De Cid, Olga Campos, Raquel Cuesta, Álex Teulé, Èlia Grau, Judit Sanz, Gabriel Capellá, Orland Díez, Joan Brunet, Judith Balmaña, Conxi Lázaro

Genes (Basel). 2021 Jan 23;12(2):150.Doi: 10.3390/genes12020150.

Non-Lynch familial and early-onset colorectal cancer explained by accumulation of low-risk genetic variants

Pilar Mur , Nuria Bonifaci, Anna Díez-Villanueva, Elisabet Munté, Maria Henar Alonso , Mireia Obón-Santacana, Gemma Aiza, Matilde Navarro, Virginia Piñol, Joan Brunet , Ian Tomlinson, Gabriel Capellá, Victor Moreno, Laura Valle

Cancers (Basel). 2021 Jul 31;13(15):3857. Doi: 10.3390/cancers13153857.

Role of psychological background in cancer susceptibility genetic testing distress: It is not only about a positive result

Adrià López-Fernández , Guillermo Villacampa, Mònica Salinas , Elia Grau, Esther Darder, Estela Carrasco, Ares Solanes, Angela Velasco, Maite Torres, Elisabet Munté, Silvia Iglesias, Sara Torres-Esquius, Noemí Tuset, Orland Diez, Conxi Lázaro, Joan Brunet, Sergi Corbella, Judith Balmaña

J Genet Couns. 2023 Aug;32(4):778-787. Doi: 10.1002/jgc4.1687.

Risk of endometrial cancer after RRSO in *BRCA* 1/2 carriers: a multicentre cohort study

Helena Pla-Juher, Marta Pardo, Àngel J Izquierdo, Esther Darder, Anna Carbó, Elisabet Munté, Sara Torres-Esquius, Judith Balmaña, Concepción Lázaro, Joan M Brunet, Maria-Pilar Barretina-Ginesta

Clin Transl Oncol. 2024 Apr;26(4):1033-1037. Doi: 10.1007/s12094-023-03312-4. Epub 2023 Sep 8.

Altered chromatin landscape and 3D interactions associated with primary constitutional MLH1 epimutations

Paula Climent-Cantó, Marc Subirana-Granés, Mireia Ramos-Rodríguez, Estela Dámaso, Fátima Marín, Covadonga Vara, Beatriz Pérez-González, Helena Raurell, Elisabet Munté, José Luis Soto, Ángel Alonso, GiWon Shin, Hanlee Ji, Megan Hitchins, Gabriel Capellá, Lorenzo Pasquali,Marta Pineda

Clin Epigenetics. Acceptat novembre 2024. Doi: 10.1186/s13148-024-01770-3

DICER1 in cancer predisposition populations: prevalence, phenotypes and mosaics

Lluis Salvador, Jesús del Valle, Eduard Dorca, Elisabet Munté, José Camacho Valenzuela, Anne-Sophie Chong, MSc, Cristina Rioja, Laura Martí-Sánchez, Mónica Salinas, Esther Darder, Joan Brunet, Hector Salvador, Conxi Lázaro, Barbara Rivera.

Genet Med. Acceptat desembre 2024

TP53 Germline Testing and Hereditary Cancer: How Somatic Events and Clinical Criteria Affect Variant Detection Rate

Paula Rofes, Carmen Castillo-Manzano, Mireia Menéndez, Alex Teulé, Sílvia Iglesias, Elisabet Munté, Mireia Ramos-Muntada, Carolina Gómez, Eva Tornero, Esther Darder, Eva Montes, Laura Valle, Gabriel Capellá, Marta Pineda, Joan Brunet, Lidia Feliubadaló, Jesús del Valle, Conxi Lázaro

Genome Med. Pendent de decisió final.

A randomized study of two risk assessment models for individualized breast cancer risk estimation

Adrià López-Fernández, Laura Duran-Lozano, Guillermo Villacampa, Mónica Pardo, Eduard Pérez, Esther Darder, Anna Vallmajó, Rosa Alfonso, Mara Cruellas, Adriana Roqué5, Mireia Cartró, Adriana Bareas, Estela Carrasco, Alejandra Rezqallah, Ana Raquel Jimenez-Macedo, Sara Torres-Esquius, Maite Torres, Consol Lopez, Martín Espinosa, Alex Teulé, Elisabet Munté, Noemi Tuset, Orland Diez, Lidia Feliubadaló, Conxi Lázaro, Gemma Llort, Anna Mercadé, Antonis Antoniou, Joan Brunet, Teresa Ramon y Cajal, Judith Balmaña.

Journal of the National Cancer Institute. Revisió major.