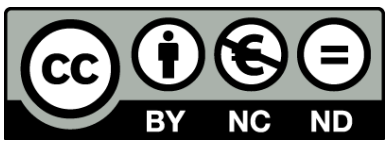
UNIVERSITAT DE BARCELONA

Anàlisi de dades de seqüenciació de nova generació pel diagnòstic molecular del càncer hereditari i per la recerca de les bases genètiques del càncer colorectal esporàdic

Adriana López-Dóriga Guerra



ANÀLISI DE DADES DE SEQÜENCIACIÓ DE NOVA GENERACIÓ PEL DIAGNÒSTIC MOLECULAR DEL CÀNCER HEREDITARI I PER LA RECERCA DE LES BASES GENÈTIQUES DEL CÀNCER COLORECTAL ESPORÀDIC

Adriana López-Dóriga Guerra

2016

ANÀLISI DE DADES DE SEQÜENCIACIÓ DE NOVA GENERACIÓ PEL DIAGNÒSTIC MOLECULAR DEL CÀNCER HEREDITARI I PER LA RECERCA DE LES BASES GENÈTIQUES DEL CÀNCER COLORECTAL ESPORÀDIC

Memòria presentada per
Adriana López-Dóriga Guerra

Per optar al Grau de
Doctora per la Universitat de Barcelona

Tesi realitzada sota la direcció de les doctores:
Conxi Lázaro García
i
Lídia Feliubadaló Elorza

A l'Institut Català d'Oncologia
Institut d'Investigació Biomèdica de Bellvitge
(ICO-IDIBELL)

Tesi adscrita a la línea de Càncer i Genètica Humana
Facultat de Medicina, Universitat de Barcelona (UB)
Tutor: **Dr. Víctor Moreno Aguado**

Conxi Lázaro Lídia Feliubadaló Víctor Moreno Adriana López-Dóriga

Barcelona, 2016

A la meva família

Agraïments

Els que em coneixeu sabeu que sóc persona de poques paraules, així en aquest apartat em limitaré a llistar els noms de tots aquells que m'heu ajudat d'una manera o d'una altra a realitzar aquesta tesi, amb cada nom em vénen al cap multitud de moments, de feina, de dinars i esmorzars, d'excursions, de sortides, però que sense dubte són els més importants i els que més recordaré, molt més que la tesi en si. Així,

Conxi i Nònia
Víctor
Gabi
Rebeca, Eli, Sussana, Isa, Henar, Xavis, Adrià, David, Gemma, Ferran, Fran, Carmen, Pili
Mireias, Dani, Marta, Evas, Juani, Esther, Jesús, Raquel, Gardenia, Laura, Fernando
Noemie, Leila
Cristina, Ramon

Papá, mamá, Sergio, Cristina, Cecilia, Tata, Nona

Arnau i Xesco

A tots, moltíssimes gràcies, perquè aquests anys durant els que he anat aprenent, analitzant i escrivint, no els podria recordar tan ben acompanyada.

ÍNDEX

ÍNDEX

ABREVIATURES

BAM: *Binary Alignment Map* (Alineament binari)
C3: *Chromosome conformation capture* (Captura de la conformació de cromosomes)
C4: *Chromosome conformation capture-on-chip* (Captura de la conformació del cromosomes sobre xip)
CCR: Càncer Colorectal
ChIP-Seq: *Chromatin immunoprecipitation sequencing* (Seqüenciació de la immunoprecipitació de la cromatina)
CNV: *Copy Number Variation* (Variació del nombre de còpies)
CRT: *Cyclic Reversible Termination* (Terminació reversible cíclica)
DNA: *Deoxyribonucleic Acid* (Àcid Desoxiribonucleic)
FAP: *Familial Adenomatous Polyposis* (Poliposi Adenomatosa Familiar)
GB: *Gigabyte*
HBOC: *Hereditary Breast and Ovarian Cancer* (Càncer hereditari de mama i ovari)
NPCC: *Hereditary Non-Polyposis Colorectal Cancer* (Càncer hereditari no polipòsic)
IGV: *Integrative Genomics Viewer* (Visor integrador de genòmica)
MAF: *Minor Allele Frequency* (Freqüència de l'al·lel minoritari)
MAP: *MUTYH-associated Polyposis* (Poliposi associada a *MUTYH*)
MeDIP-seq*: Methylated DNA immunoprecipitation sequencing* (Seqüenciació de la immunoprecipitació delDNA metilat)
Methyl-Seq: *Methylation Sequencing* (Seqüenciació de la metilació)
NGS: *Next Generation Sequencing* (Seqüenciació de Nova Genearació)
pa: Persones/any
pb: Parells de bases
PCR: *Polymerase Chain Reaction* (Reacció en cadena de la polimerasa)
RNA: *Ribonucleic acid* (Àcid Ribonucleic)
ROI: *Region of Interest* (Regió d'interès)
SAM: *Sequence Alignment Map* (Mapa d'Alineament de seqüències)
SNP: *Single Nucleotide Polymorphism* (Polimorfisme d'un únic nucleòtid)
SNS: *Single Nucleotide Substitution* (Substitució d'un únic nucleòtid)
SNV: *Single Nucleotide Variant* (Variant d'un únic nucleòtid)
UTR: *UnTranslated Region* (Regió no traduïda)
VCF: *Variant Calling Format* (Format de la detecció de variants)

INTRODUCCIÓ

1. Seqüenciació de nova generació (*Next Generation Sequencing*, NGS)

S'anomena seqüenciació de nova generació, seqüenciació de segona generació, seqüenciació d'alt rendiment o seqüenciació paral·lela massiva, a la tecnologia desenvolupada a partir del 2005 que permet la seqüenciació clonal (a partir de molècules aïllades) d'àcids nucleics en paral·lel amb un gran rendiment. En els últims deu anys, la seqüenciació massiva s'ha desenvolupat de forma exponencial aconseguint una relació qualitat/preu per base excel·lent, permetent projectes més ambiciosos o a major escala que la seqüenciació Sanger i a un preu més reduït. L'aparició de la NGS ha superat quasi totes les limitacions de les estratègies per a la seqüenciació utilitzades fins llavors en quant a capacitat, tot i que es produeixen errors en la resolució de seqüència que s'intenten pal·liar amb noves químiques en les reaccions i nous algoritmes bioinformàtics i diagnòstics.

1.1 Aplicacions de la NGS

La seqüenciació de nova generació és útil en aplicacions genòmiques, epigenòmiques i transcriptòmiques. En aquesta tesi s'utilitza la NGS per a la reseqüenciació del genoma, és a dir, la seqüenciació de regions del genoma conegudes per a comparar la seqüència resultant amb la de referència (Shankar 2011). A la seqüenciació de gens, exomes i genomes complets se'n fa referència com a **DNA-seq** (seqüenciació a partir de DNA). En aquest camp es poden dissenyar diferents experiments, la majoria d'ells enfocats a trobar variants en les mostres analitzades respecte a la seqüència coneguda de referència. En aquesta tesi s'aplica la DNA-Seq en projectes de diagnòstic genètic i de recerca relacionats amb càncer.

La NGS també s'utilitza per a la seqüenciació *de novo* de genomes, especialment genomes bacterians i virals. En aquests experiments es seqüencia el DNA del genoma desconegut i s'alineen les lectures sobreposant unes amb les altres per a formar una seqüència consens, que en la majoria d'estudis serà la nova seqüència d'interès (Harris et al. 2008).

Una altra de les aplicacions principals de la NGS és la seqüenciació d'RNA, coneguda com a **RNA-seq**, que permet caracteritzar el transcriptoma de cèl·lules, teixits i organismes (Cloonan et al. 2008, Wang et al. 2009).

Una altra de les aplicacions de la NGS és la seqüenciació de l'epigenoma, que estudia tots aquells processos que alteren l'expressió dels gens sense canviar la pròpia seqüència de DNA. Entre aquests processos es troben la modificació de l'estructura de la cromatina, o de la metilació del DNA. La NGS permet estudiar aquests processos mitjançant experiments de **ChIP-Seq** (*Chromatin immunoprecipitation sequencing*) (Kharchenko et al. 2008), **Methil-Seq** (*Methylation Sequencing*), **MedIP-Seq** *(Methylated DNA immunoprecipitation)* (Weber et al. 2005), **C3** (Chromosome conformation capture) (Dekker et al. 2002) o **C4** (*Chromosome conformation capture-on-chip*) (Simonis et al. 2006).

1.2 Seqüenciació de DNA amb la tecnologia NGS

Tot i que cada plataforma NGS té característiques específiques, el protocol general de seqüenciació és semblant en totes elles. Tal com s'esquematitza a la figura 1, els passos principals per a realitzar una seqüenciació massiva són: preparació de la mostra, preparació de la genoteca, fragmentació del DNA, afegint-li adaptadors i enriquint-lo en les regions d'interès (en anglès *Region of Interest*, ROI), amplificació clonal, immobilització, i seqüenciació (Metzker 2010).

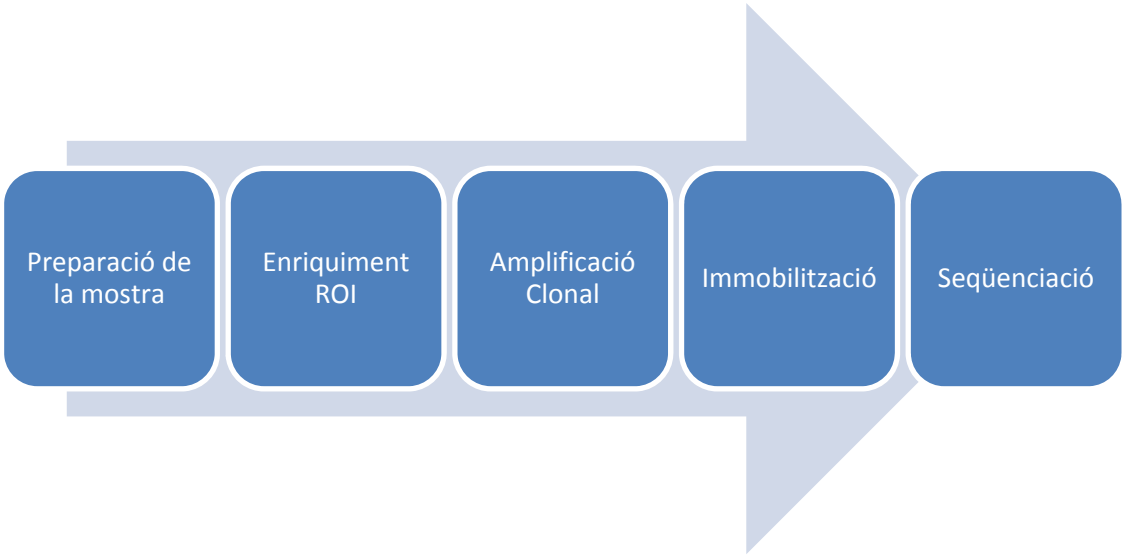


Figura 1. Passos de la NGS.

A continuació es comenten breument els passos de la seqüenciació massiva:

- Preparació de la mostra: Cal fragmentar, reparar i unir adaptadors al DNA o producte de PCR (*Polymerase Chain Reaction*) de bona qualitat amb una concentració mesurada amb precisió i una quantitat de l'ordre de micrograms (1 a 5 µg). Darrerament la quantitat inicial de DNA s'està aconseguint reduir fins a 50-500 ng, tot i que això sol anar en detriment de la diversitat de la genoteca.
- Enriquiment ROI: La seqüenciació del genoma complet és molt costosa i per molts projectes és innecessària ja que la funció de la majoria del genoma encara es desconeix. Per això, en molts casos és important realitzar un enriquiment del DNA per tal de seleccionar aquella o aquelles regions específiques d'interès. En general hi ha quatre maneres d'enriquir aquestes regions: la PCR, habitualment multiplex, la qual utilitza diverses parelles d'encebadors en una sola reacció generant múltiples amplicons, la captura per circularització, la selecció per hibridació en solució, i la selecció per hibridació en fase sòlida. Tot i que s'han descrit molts mètodes per capturar la regió d'interès per a seqüenciar, la més comuna per a capturar uns pocs gens específics sol ser la PCR multiplex, mentre que per capturar l'exoma humà, és a dir, totes les regions codificants del genoma humà, bàsicament s'utilitzen mètodes de captura per hibridació (Turner et al. 2009, Mamanova et al. 2010).

- Amplificació clonal: l'amplificació clonal es pot realitzar per diferents mètodes, així les tecnologies Roche o Ion Torrent realitzen una PCR en emulsió, mentre que Illumina realitza una PCR en fase sòlida.
- Immobilització: La immobilització dels clons es pot realitzar en una superfície de cristall amino revestida com es fa en la tecnologia Illumina, en pouets individuals en plaques *picotiter* com fa Roche, o per altres mètodes.
- Seqüenciació: la química de la seqüenciació també varia segons la tecnologia, així en les plataformes de *Roche* es realitza una piroseqüenciació, en plataformes *Illumina* una seqüenciació per terminació reversible, en plataformes de *Life Technologies* una seqüenciació per detecció de protons durant la polimerització, o en *Applied Biosystems* una seqüenciació per lligació.

1.3 Plataformes de NGS

Fins ara, diferents plataformes per a la seqüenciació de nova generació han sortit al mercat, les més comunes són el *Genome Sequencer FLX (454-Roche),* el *HiSeq (Illumina),* el *SOLID (Applied Biosystems),* l'*HeliScope (Helicos Biosciences)* i l'*Ion PGM (Life Technologies)*. Aquestes plataformes es diferencien en diversos aspectes, com la tecnologia utilitzada, la llargada de les lectures de seqüència, o el número de molècules de DNA seqüenciades, entre d'altres (Shendure and Ji 2008). S'han desenvolupat també plataformes de mitjà rendiment que s'adapten millor a les necessitats dels dissenys experimentals de laboratoris de diagnòstic. Així per exemple, el GS Junior de Roche és un seqüenciador amb la mateixa tecnologia que el GS FLX però amb menys capacitat, o el MiSeq o MiniSeq d'Illumina també són de dimensions i producció més petites que el HiSeq (Roche 2016) (Illumina 2016).
Seguint amb la comparació de les diferents tecnologies, podem distingir sis criteris principals pels quals es poden avaluar les plataformes de NGS i que estan molt relacionats entre ells (Mardis 2008, Shendure and Ji 2008, Metzker 2010):

- Rendiment: determina la quantitat de bases que es poden seqüenciar en un determinat temps i contempla el temps que requereix per a preparar i executar una carrera.
- Longitud de les lectures: La longitud mitjana dels fragments seqüenciats que determina la utilitat de la seqüència per a vàries aplicacions. Lectures més llargues tenen una probabilitat més alta de ser úniques en el genoma i per tant, són més fàcils d'utilitzar en la reconstrucció de DNA complet, especialment en les regions repetides. Les lectures llargues també ajuden a detectar delecions de diverses bases.
- Exactitud o qualitat: La qualitat de la seqüència generada es determina en funció de la freqüència d'errors. La qualitat té implicacions en la confiança per a la detecció de variants en el DNA i per tant, en la profunditat (en anglès *coverage*) requerida per a una anàlisi, això és el número de lectures que cobreix cada nucleòtid de la regió desitjada.

- Robustesa: L'habilitat de completar una carrera amb èxit i que depèn de la reproductibilitat dels protocols de laboratori, la robustesa de la química i de l'instrument. Els primers instruments amb els corresponents protocols fallaven bastant en aquest aspecte, i a base de l'experiència adquirida es van anar perfeccionant.
- Aplicacions: La diversitat d'aplicacions que es poden executar en un sol instrument determina la seva utilitat en estudis de recerca específics. Per exemple, la possibilitat de multiplexar, és a dir, de seqüenciar diferents mostres en una única carrera és un factor important per a la reseqüenciació de grups de gens.
- Cost: El cost per base de DNA, que en molt casos determina la possibilitat de la reseqüenciació d'alguns gens o de tot el genoma. Els costos també estan disminuint considerablement, i actualment ja es pot seqüenciar un genoma humà per $1000 (Illumina 2016).

Altres criteris a considerar són la disponibilitat de *software* per a l'anàlisi de les dades, la facilitat de compensar o corregir els errors de seqüència produïts i el cost computacional que això suposa, ja que els errors de substitució són més fàcils de detectar i corregir que les insercions o delecions. També cal valorar el potencial de la tecnologia per millorar qualsevol dels criteris anteriors.

A la taula 1 es mostren les principals tecnologies de NGS, amb els respectius instruments i les característiques més destacades de cadascun d'ells (Metzker 2010, Henson et al. 2012, Ozsolak 2012).

Taula 1. Comparació de les diferents tecnologies de NGS amb els corresponents instruments principals*.

Tecnologia	Química	Instrument	Llargada aprox. de les lectures	Rendiment per carrera	Temps aproximat de carrera	Error principal	Percentatge error/base
Roche 454	Piroseqüenciació	454 GS FLX Titanium XL+	1000	700 Mb	23 h	Insercions i delecions	0,50%
		454 GS Junior	700	70 Mb	18 h	Insercions i delecions	1%
Illumina	Seqüenciació per terminació reversible	HiSeq 2500	2x125	1000 Gb	11 dies	Substitucions	0,02%
		MiSeq	2x300	15 Gb	27 h	Substitucions	0,02%
		NextSeq500	2x150	120 Gb	11-29 h	Substitucions	*a*
Ion Torrent	Seqüenciació per detecció de protons durant la polimerització	PGM amb xip 316	200	100 Mb	2 h	Insercions i delecions	0,02%
		PGM amb xip 318	200	1 Gb	2 h	Insercions i delecions	0,02%
		Proton	200	> 1 Gb	2-4h	Insercions i delecions	0,02%
SOLiD	Seqüenciació per lligació	SOLiD 4	75	100 Gb	12 dies	Biaix A-T	0,06%
		SOLiD 4hq	75	300 Gb	14 dies	Biaix A-T	0,01%
		SOLiD PI	75	77 Gb	8 dies	Biaix A-T	0,01%

***** Adaptació de (Metzker 2010, Henson, Tischler et al. 2012) i actualitzada en base a les pàgines webs (Illumina 2016, Roche 2016, Thermo Fisher Scientific 2016, Thermo Fisher Scientific 2016).

a El format d'especificacions per a aquest instrument és: més del 75% de les bases amb una qualitat major de 30.

Dues de les tecnologies més utilitzades per a la seqüenciació del DNA i que s'han utilitzat en aquesta tesi són la de 454/Roche i la d'Illumina. A continuació s'explica amb més detall el seu funcionament:

- **Piroseqüenciació Roche/454:** A la tecnologia 454 es parteix de fragments de DNA que són amplificats clonalment mitjançant una PCR en emulsió sobre petites boles. Per a la piroseqüenciació, aquestes boles es dipositen individualment en pous separats per on iterativament flueixen els quatre nucleòtids. Quan un nucleòtid s'incorpora a la cadena creixent, el pirofosfat alliberat desencadena una cascada de reaccions que alliberen una quantitat de llum proporcional al nombre de nucleòtids incorporats. Aquesta tecnologia dóna la possibilitat de multiplexar, és a dir, seqüenciar diverses mostres juntes, les lectures de les quals es distingeixen gràcies a la incorporació d'uns fragments de DNA com a codi de barres, anomenats *Multiplex Identifiers (MIDs)*, entre els extrems universals i el fragment a llegir. Un dels desavantatges de la tecnologia és la poca precisió en determinar la llargada de les regions amb homopolímers, aquests són petits fragments de seqüència amb un sol nucleòtid repetit. Quan es troba un homopolímer, si el número de nucleòtids és alt, a partir de 6 aproximadament, es perd la proporcionalitat de la intensitat del senyal lumínic detectat respecte el número de nucleòtids repetits (De Leeneer et al. 2011, Loman et al. 2012). Comparant amb altres tecnologies de NGS, una de les fortaleses del sistema de Roche/454 és la longitud de les lectures seqüenciades. El Roche/454 GS FLX, amb la química GS FLX Titanium, pot generar més d'un milió de lectures amb longituds superiors a 1000 bases. Tot i que el cost per base és bastant superior a altres tecnologies com SOLiD o Illumina, el sistema de Roche/454 és molt útil en certes aplicacions com la seqüenciació *de novo* per a determinar nous genomes, o en la detecció de variants víriques, ja que la major llargada de lectura permet alineaments més robustos que toleren seqüències amb insercions o delecions mitjanes, on la llargada de les lectures és un factor crític. També permet veure si diverses variants es troben o no en el mateix al·lel, molt útil en casos de DNA hipervariables com les regions de l'HLA o les seqüències víriques. Aquestes característiques la converteixen en una eina eficient i sensible per a la detecció de variants, amb demostrades aplicacions en àrees de recerca clínica i diagnòstic (De Leeneer et al. 2011, Allard et al. 2012, Jiang et al. 2012, Danzer et al. 2013, Quer et al. 2015).

Els dos instruments que funcionen amb aquesta tecnologia són el "GS FLX" que produeix aproximadament 1.000.000 de lectures per carrera amb un rendiment de 700 Mb, i el "GS Junior" que produeix unes 100.000 lectures per carrera amb un rendiment total d'unes 70 Mb. El GS Junior és l'instrument amb tecnologia 454 per a la seqüenciació massiva pensat per a laboratoris amb projectes de mitjana escala, ja que els costos de maquinària i manteniment són menors, i la seva capacitat s'adapta millor a projectes que requereixen la seqüenciació de múltiples gens però sense arribar a exomes o genomes complets. El "GS Junior" va ser la primera plataforma de mitjà rendiment que va sortir al mercat l'any 2007.

- Seqüenciació Illumina: Illumina va treure al mercat la primera plataforma de seqüenciació de lectures curtes (Metzker 2010) i actualment domina el mercat de la NGS. La combinació de la terminació reversible i la immobilització en un pla permeten la seqüenciació massiva i en paral·lel de milions de fragments de DNA a un baix cost. Les mostres de DNA es fragmenten aleatòriament i els extrems dels

fragments es reparen per a generar terminacions 5' fosforilades per unir-hi adaptadors de PCR i de seqüenciació. Després, el DNA es desnaturalitza i els fragments de cadena simple s'amplifiquen immobilitzats sobre una superfície sòlida i transparent anomenada *flow cell*. Es tracta d'una PCR clonal en pont (en anglès *bridge PCR amplification*), que genera un conjunt (en anglès *cluster*) d'aproximadament 1000 molècules idèntiques de DNA per tal d'obtenir suficient intensitat de senyal lluminós per a una detecció fiable. La *bridge PCR* genera centenars de milions de *clusters,* permetent una densitat de seqüenciació molt elevada. Finalment, aquests fragments es seqüencien "per síntesi", mitjançant una reacció que utilitza els encebadors, la polimerasa i els quatre nucleòtids modificats com a terminadors reversibles, cadascun marcat amb una molècula fluorescent diferent, que es van incorporant en la *flow cell*. En cada cicle de seqüenciació s'incorpora un nucleòtid terminador reversible marcat amb fluorescència diferent, d'acord amb la complementarietat de bases de cada fragment de DNA. Després de la incorporació, la identitat i la posició dels terminadors incorporats a la *flow cell* es determinen d'acord amb la posició i longitud d'ona de la llum emesa per la molècula fluorescent. El senyal es grava utilitzant una càmera, i posteriorment es desxifra amb un *software* propi. Per poder començar el següent cicle, una reacció química modifica el nucleòtid terminador desprenent-ne la fluorescència i tornant-lo apte per a unir-se al següent nucleòtid. Així es va allargant (sintetitzant) la cadena fins la llargada pròpia de cada plataforma i *kit*.

Illumina té vàries plataformes amb aquesta tecnologia, i poc a poc en van sortint de noves per adaptar-se millor a les necessitats de cada aplicació; fins ara les dues més utilitzades són la HiSeq2500 i la MiSeq. De la mateixa manera que les plataformes de la tecnologia 454, la plataforma HiSeq2500 està pensada per a projectes de gran escala, per a obtenir fins a 1000 Gb de dades de seqüència. En canvi, la plataforma MiSeq està pensada per a projectes de mitjà rendiment, per a obtenir fins a 15 Gb de dades i per seqüenciar amb lectures més llargues que el HiSeq2500. El MiSeq va sortir al mercat a finals de l'any 2011.

2. Anàlisi bioinformàtica

El terme "bioinformàtica" fa referència específicament a la creació de teories, algoritmes, tècniques informàtiques i tècniques estadístiques per a resoldre problemes derivats del processament i l'anàlisi de dades biològiques. La disponibilitat de tècniques moleculars que generen grans quantitats de dades com la NGS, i el desenvolupament d'estratègies bioinformàtiques cada cop més complexes, han revolucionat la recerca biomèdica. La bioinformàtica és indispensable per moltes aplicacions de biologia molecular com poden ser: l'anàlisi de seqüències, la identificació de gens, l'anàlisi de l'expressió gènica, la biologia evolutiva, o l'anàlisi de mutacions en el càncer, entre d'altres.

La seqüenciació de nova generació genera grans quantitats de dades que un ordinador convencional en molts casos no pot processar (Richter and Sexton 2009). Tot i que hi ha instruments que inclouen *software* d'anàlisi per a usuaris no especialitzats, en la majoria de casos aquest tipus de dades requereix una persona amb coneixements d'informàtica per processar-les. Poc a poc es van generant diferents *software* comercials, cada cop més precisos, que permeten analitzar les dades de seqüenciació, tot i que

quan es tracta de fer anàlisis fora dels convencionals són molt poc flexibles. A més, la interpretació biològica dels resultats requereix la integració de dades de diversos experiments, així com d'informació disponible en bases de dades públiques que cal seleccionar.

2.1 Anàlisi bioinformàtica de dades de NGS

En la majoria d'anàlisis de NGS, els fitxers d'imatges produïts pels seqüenciadors es converteixen en sèries de bases que cal processar. Aquestes lectures es processen mitjançant assemblatge o alineament, i posteriorment una detecció de variants o quantificació de lectures, segons l'objectiu de l'experiment. Al llarg dels processos es creen molts resultats intermedis. La gran magnitud de les dades obliga a emmagatzemar només els resultats més necessaris i el codi d'anàlisi per si cal tornar a analitzar les dades. En termes de computació, l'anàlisi de les dades de NGS requereix una gran capacitat. Aquest fet crea la necessitat de dividir els processos d'anàlisi en subprocessos i després tornar a unir els resultats. Els algoritmes i programes s'han de modificar i es necessiten infraestructures informàtiques molt potents; els servidors amb diversos nodes per paral·lelitzar són imprescindibles (Pop and Salzberg 2008). L'anàlisi de les dades de NGS és molt similar en totes les plataformes, tot i que hi ha alguns aspectes particulars de cadascuna, a més aquesta anàlisi també depèn de l'experiment que es realitza i del seu objectiu. A la figura 2 es representa un esquema simplificat dels processos principals en l'anàlisi bioinformàtica de les dades de DNA-Seq amb NGS, seguida d'una breu explicació de cadascun dels procediments.

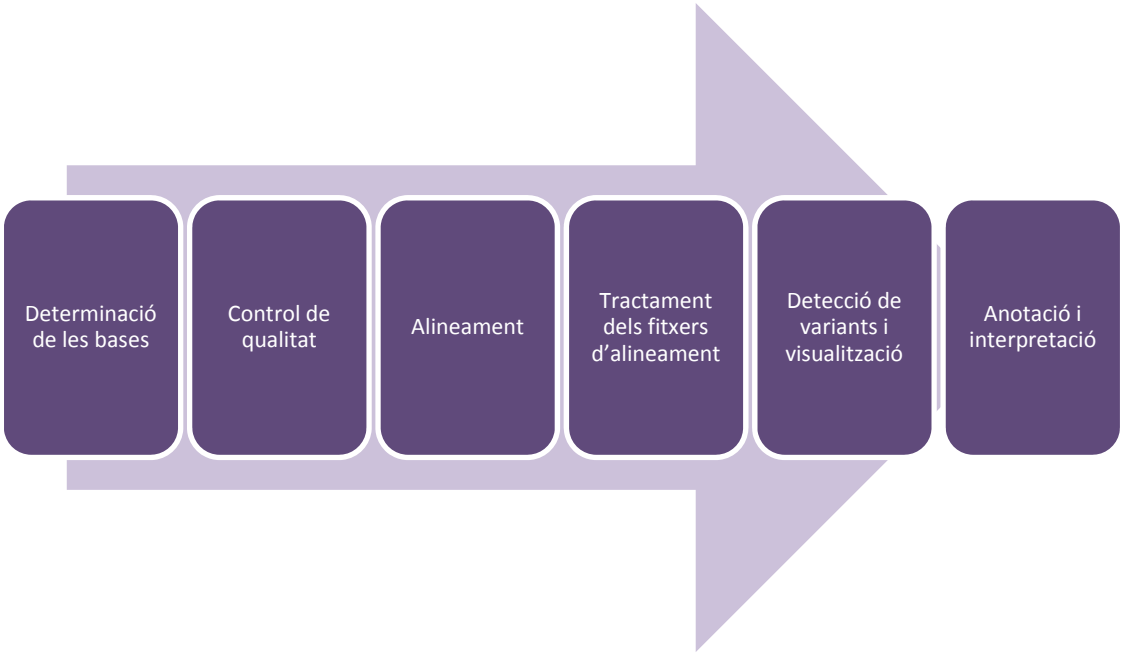


Figura 2. Protocol d'anàlisi bioinformàtica estàndard de dades de NGS per a la detecció de variants.

Qualsevol anàlisi de DNA-Seq inclou bàsicament sis passos: (1) la determinació de les bases, (2) el control de qualitat, (3) l'alineament, (4) el processament dels fitxers d'alineament, (5) la detecció de variants i (6) la interpretació funcional de les variants (Koboldt et al. 2010).

(1) **Determinació de les bases:** Es tracta d'una anàlisi automàtica que té lloc en els propis seqüenciadors i es basa en la conversió d'imatges digitals a dades significatives. Molt resumidament, les imatges generades per la seqüenciació s'alineen d'una manera determinada segons les seqüències d'interès i es van llegint. En la tecnologia Illumina, cada color representa un nucleòtid diferent i les intensitats del senyal s'utilitzen per al càlcul de les qualitats de les bases, normalment les qualitats de les bases es calculen en base a unes dades de referència obtingudes amb una seqüència coneguda que serveix per a calibrar. En el cas de Roche/454, les intensitats del senyal s'utilitzen per estimar el número de nucleòtids iguals contigus. Tant la tecnologia Roche/454 com la Illumina, tenen incorporat un pas de filtratge on filtren lectures molt curtes o amb mala qualitat abans de proporcionar el resultat definitiu. El format estàndard amb que s'extrauen les lectures amb les corresponents qualitats és el format "*fastq*". Les lectures resultants dels seqüenciadors Roche/454 també es poden obtenir en fitxers amb format "*.sff*", però a partir dels "*.sff*" es poden obtenir els fitxers ".fasta" (o també "*.fna"* i "*.qual")* equivalents.

(2) **Control de qualitat:** és important realitzar un control de qualitat de les lectures obtingudes després de la seqüenciació, en aquest pas s'obtenen unes estadístiques bàsiques sobre el nombre de lectures, la distribució de la llargada de les lectures, la qualitat mitjana per base i per seqüència, el contingut en GC, els nivells de duplicació de les seqüències, o el contingut d'adaptadors en les seqüències. Tots aquests són paràmetres que podrien indicar un problema en la seqüenciació i per tant, cal avaluar com poden afectar als resultats, i en alguns casos extrems podria ser recomanable repetir la seqüenciació.

(3) **Alineament:** el procés d'alineament té com a finalitat localitzar cadascuna de les lectures en la posició correcta de la seqüència de referència. Si no hi ha seqüència de referència perquè no es coneix, el que s'intenta és col·locar les lectures una a continuació de l'altra sobreposant les bases coincidents i construint una seqüència consens. Amb la NGS ha sigut necessari el desenvolupament de nous algoritmes que s'adaptessin millor als requeriments d'aquest tipus de dades de gran volum. Els nous algoritmes d'alineament han de contemplar aspectes com la capacitat d'alinear milions de seqüències, considerar que el mapatge pot no ser únic quan les lectures són curtes, i també cal considerar que la seqüenciació no és perfecta i existeix una petita probabilitat de que una base sigui incorrecta, a més de les variants reals de seqüència (Dohm et al. 2008). També cal tenir present que hi ha regions del DNA que són difícils de seqüenciar, com els homopolímers o les regions amb alt contingut de nucleòtids GC, i això pot augmentar la probabilitat d'error en la seqüenciació. Tot això fa que el procés d'alineament sigui complex, i que segons l'objectiu de l'estudi calgui avaluar quins aspectes prioritzar per a

triar l'algoritme més adient. En l'apartat de *software* (2.2) es comenten els algoritmes d'alineament que permeten processar dades de NGS amb un rendiment òptim.

(4) **Tractament dels fitxers d'alineament:** En general els *software* d'alineament generen un fitxer en format SAM (*Sequence Alignment Map*), que té el format binari equivalent BAM (Li and Durbin 2009). Es tracta d'un format genèric que conté les lectures i l'alineament corresponent respecte a la seqüència de referència, els fitxers (SAM/BAM) contenen molta informació, tant de les lectures mapades com de les no mapades, i fins i tot es poden recuperar les lectures originals a partir d'aquests fitxers. Les propietats de les lectures i del seu mapatge, és a dir, les propietats que indiquen si la lectura s'ha mapat correctament, si la seva parella ha mapat, o si està duplicada, entre d'altres, es descriuen amb una etiqueta (*flag*), aquesta etiqueta està formada per la suma de diversos números binaris que amb un 1 indiquen si té la propietat i amb un 0 si no la té. L'etiqueta és molt útil per a filtrar o seleccionar les lectures que ens interessa per a l'anàlisi. Aquest format s'ha estès entre els usuaris que analitzen dades de NGS. Els arxius d'alineament BAM han d'estar acompanyats sempre d'un arxiu BAM.BAI que és un índex que permet treballar més eficientment. Els arxius BAM són clau per a les anàlisis posteriors i cal que estiguin ordenats. Cal realitzar un filtratge d'aquells alineaments que no ens interessin segons l'objectiu de l'anàlisi, com per exemple lectures que mapen en més d'una posició pot ser que no ens interessin, o lectures la parella de les quals alineen en un altre cromosoma, entre d'altres opcions que variaran segons l'experiment.

(5) **Detecció de variants i visualització:** Aquest pas de l'anàlisi pretén detectar diferències en la seqüència d'un individu, formada per vàries lectures en cada posició, respecte a la seqüència de referència que pot ser un gen, un grup de gens, un cromosoma, o un genoma complet. Els resultats depenen molt de la qualitat de l'alineament, ja que errors en l'alineament donen lloc a falsos positius, i lectures no alineades poden comportar falsos negatius. La utilització de la NGS permet identificar quasi tot tipus de variants de seqüència del genoma. Tot i que des de la seva aparició fins a l'actualitat hi ha hagut una clara evolució, i tant les químiques com els *software* han millorat, encara cal evolucionar per detectar tots els tipus de mutacions amb una alta precisió. En especial cal treballar en la detecció de les variacions estructurals on els mètodes d'anàlisi estan millorant però els resultats encara no són molt precisos. Aquests mètodes depenen molt del promig i de la homogeneïtat de la cobertura obtinguda de la seqüenciació per tal que els resultats siguin fiables (Koboldt et al. 2012, Lee et al. 2014).

En aquest procés de detecció de variants, en molts casos interessa visualitzar l'alineament de les lectures al voltant de les variants detectades, per tal d'observar la localització i les regions que envolten a cada variant. Amb la visualització es pot comprovar si les variants es troben en regions amb bona cobertura i poc soroll, o bé les variants es troben en zones que no estan ben seqüenciades per la proximitat a homopolímers o a regions de seqüència complicades, en aquests últims casos la variant identificada podria ser un fals positiu.

(6) **Anotació i interpretació funcional de les variants:** En aquest pas és imprescindible utilitzar la informació disponible que contenen les bases de dades públiques per a classificar les variants, això és, determinar si es tracta de polimorfismes coneguts (variants de seqüència amb una freqüència igual o superior a l'1% de la població), si les variants ja han estat reportades i s'han descrit com a somàtiques (en cas de càncer, variants només presents en les cèl·lules tumorals), o en el cas de ser variants no reportades prèviament, com es poden classificar segons l'efecte que puguin produir en la proteïna en qüestió.

2.2 *Software* per a l'anàlisi de dades de NGS

Existeixen diferents programes enfocats a l'anàlisi de dades de NGS, algunes aplicacions són comercials, altres són lliures, algunes tenen una aplicació web, o també hi ha llibreries específiques de llenguatges com R o programes lliures. Alguns estan preparats per a l'anàlisi de dades d'una tecnologia i tipus d'experiment específics, i s'adapten força bé a les necessitats de la majoria dels usuaris, però la millora de l'anàlisi de les dades de NGS és encara un repte pels bioinformàtics.

Hi ha programes comercials que permeten dur a terme molts tipus d'anàlisi diferents i acostumen a ser d'ús relativament fàcil, uns exemples són el CLC-Workbench (CLCbio QIAGEN 2016) o el SeqNext (JSI Medical Systems 2016). Però a més de l'inconvenient econòmic que implica adquirir *software* comercial, també implica un desconeixement per part de l'usuari dels processos interns que s'empren per a l'anàlisi. Aquest últim fet és una limitació important, ja que en molts casos cal conèixer bé els procediments que es realitzen per a assegurar una bona qualitat i fiabilitat dels resultats.

En canvi, el *software* lliure permet accedir al codi font i conèixer els procediments que se segueixen, a més permet adaptar-lo a les necessitats dels usuaris. L'inconvenient en aquest cas és que el *software* lliure per a l'anàlisi de dades de NGS requereix uns coneixements bàsics d'informàtica que permetin executar els procediments i adaptar els paràmetres a les necessitats de l'anàlisi. En aquesta tesi s'ha treballat bàsicament amb *software* lliure, adaptant els paràmetres i desenvolupant funcions per a obtenir uns resultats òptims. Alguns dels programes lliures més utilitzats es resumeixen a continuació:

-*Software* per al control de qualitat: Per aquesta primera anàlisi el *software* més utilitzat és el FastQC, un *software* lliure que permet fer un control de qualitat exhaustiu mitjançant una interfície intuïtiva i fàcil per a tot tipus d'usuaris (Babraham Bioinformatics 2016). Alguns altres *software* que permeten realitzar un control de qualitat són el FaQCs (Lo and Chain 2014) que optimitza la velocitat paral·lelitzant els processos, o el NGS QC Toolkit (Patel and Jain 2012) que a més del control de qualitat, també inclou alguna eina per a retallar seqüències d'adaptadors o convertir els formats de les seqüències.

-*Software* per a l'alineament: Com s'ha comentat a l'apartat anterior, el primer pas clau en l'anàlisi de dades de reseqüenciació amb NGS és l'alineament de les lectures sobre la seqüència de referència. Els primers algoritmes d'alineament van sortir als anys vuitanta. Alguns com l'algoritme Smith-Waterman,

el FASTA o el BLAST, entre d'altres, es podien utilitzar en pàgines web i permetien alinear seqüències contra grans bases de dades. El nombre de seqüències va anar augmentant fins a tenir la necessitat d'alinear milers de seqüències. Va ser llavors que W. James Kent el 2002 va publicar una nova eina anomenada BLAT, que millorava BLAST especialment en la velocitat d'alineament, gràcies principalment a la indexació sobre la base de dades de referència enlloc d'indexar la seqüència per alinear (Kent 2002). El BLAT produeix uns alineaments molt específics i alinea lectures de llargades molt diverses, de fins a 200.000 bases, però el temps és una limitació quan la quantitat de lectures és de milions, ja que requereix dies i fins i tot setmanes de computació. En aquests últims anys s'han desenvolupat molts algoritmes per a millorar el rendiment en l'alineament de lectures curtes (menors de 70 bases). Els dos més utilitzats són el Bowtie2 (Langmead and Salzberg 2012) i el BWA (Li and Durbin 2009), que es basen en l'algoritme de Burrows-Wheeler i redueixen significativament el temps de processament. Degut a la ràpida evolució de les tecnologies, i en vista que la llargada de les lectures va en augment, els algoritmes han anat millorant per adaptar-se a lectures més llargues, així el BWA-MEM (Burrows-Wheeler Aligner 2016) permet alinear lectures de més de 70 bases de manera òptima i, a més dels alineaments globals del BWA, realitza també alineaments locals. Això fa que sigui lleugerament més lent que el BWA però més precís. Quan els "forats" (en anglès, *gaps*) són freqüents, es recomana l'algoritme BWA-SW (Li and Durbin 2010) que pot tenir una sensibilitat més alta però que es veu penalitzat amb el temps. Existeixen altres *software* per a l'alineament com el SHRIMP (Rumble et al. 2009) que es basa en l'algoritme de Smith-Waterman i permet alinear dades de SOLiD, o també l'alineador BFAST (Homer et al. 2009) que utilitza el mateix algoritme que BLAT però és més ràpid i permet forats en l'alineament. El Genome Multitool o GEM (Marco-Sola et al. 2012) és una de les últimes eines publicades que permet alinear dades de NGS, és molt potent i conté eines molt optimitzades per indexar i analitzar grans genomes. Però una gran limitació del GEM és el format amb què proporciona els resultats ja que és un format propi i això fa que el seu ús no sigui tan estès per incompatibilitats amb altres programes.

-*Software* per al tractament dels fitxers d'alineaments: Com s'ha comentat prèviament, és important tractar els fitxers d'alineament per a obtenir els formats i alineaments òptims, hi ha un *software* lliure anomenat SAMtools (Li et al. 2009) que té les eines necessàries per a manipular aquests tipus de fitxers, i permet ordenar, combinar, indexar, mostrar o crear arxius SAM/BAM . Altres eines com Picard (Broad Institute 2016) també permeten manipular aquest tipus de fitxers, així com adaptar el format als requeriments d'altres programes.

-*Software* per a la detecció de variants: S'han descrit nombrosos algoritmes per a detectar substitucions o insercions i delecions. Alguns utilitzen mètodes Bayesians com l'Atlas-SNP2 (Shen et al. 2010) o el SOAPsnp (Li et al. 2009), que estimen el genotip més probable basant-se en models de regressió logística entrenats a partir de molts grups de dades, i calculen la probabilitat que la substitució estimada sigui un error basats en la informació prèvia disponible sobre els errors de seqüència. Altres paquets com SAMtools (Li 2011), GATK (McKenna et al. 2010, DePristo et al. 2011), o VarScan2 (Koboldt

et al. 2012) inclouen utilitats per a detectar i filtrar variants basant-se en models heurístics i probabilístics, reforçats amb el coneixement empíric de les diferents plataformes de seqüenciació, i filtrant les variants en base a paràmetres com la cobertura, la qualitat de la base o de l'alineament, o la freqüència. Per a qualsevol dels programes anteriors, la detecció d'insercions i delecions és més costosa i produeix més falsos positius, això s'explica bàsicament perquè a nivell de computació és més fàcil alinear lectures amb substitucions que lectures amb insercions o delecions. En teoria, amb lectures llargues com les que generen tecnologies com 454 no seria tan problema detectar insercions i delecions, però aquestes tecnologies presenten la dificultat al voltant dels homopolímers, on s'acumulen la majoria de falsos positius en forma de delecions i insercions. Tot i això, *software* com VarScan2 , GATK o Pindel (Ye et al. 2009) són cada vegada més precisos detectant insercions i delecions a partir d'alineaments amb forats (Xu et al. 2014). La majoria de *software* que detecta variants genera els resultats en format *.VCF* (en anglès *Variant Calling Format*)(Danecek et al. 2011). Aquest format conté una capçalera amb la informació de tots els paràmetres utilitzats per a descriure les variants, i després cada línia correspon a una variant on s'indica, amb diferents variables, la posició, el canvi de nucleòtid o inserció/deleció, la cobertura, o nombre de lectures en aquella posició amb la base de referència i amb l'alternativa, també desglossat segons el sentit de les lectures, i la qualitat del mapatge i de les bases, entre d'altres paràmetres. No tots els *software* generen els resultats amb les variants en format *.vcf,* un exemple és el VarScan, que reporta les variants en un fitxer de text seguint un esquema similar al *VCF*. La sensibilitat i especificitat en la detecció de variants encara són millorables, en part degut als errors de seqüència, i és necessari un filtratge acurat de les variants, amb paràmetres específics segons la tecnologia utilitzada i el poder estadístic requerit, per a obtenir uns resultats òptims.

-*Software* per a la visualització: Com hem comentat prèviament, en molts casos és útil visualitzar els alineaments i les variants per a determinar a cop d'ull si es tracta d'una variant probablement real, o si clarament sembla un fals positiu degut a la proximitat d'homopolímers o a un alineament confús. Hi ha *software* que permet carregar els alineaments en format SAM/BAM i dirigir-se a la posició d'interès a visualitzar. Dos dels programes no comercials més utilitzats són l'IGV (Integrative Genomics Viewer) (Robinson et al. 2011, Thorvaldsdottir et al. 2013) (Figura 3) i el LookSeq (Manske and Kwiatkowski 2009). Altres programes com l'Ugene (Okonechnikov et al. 2012) permeten visualitzar els alineaments, a més de tenir altres eines per a l'anàlisi de seqüències.

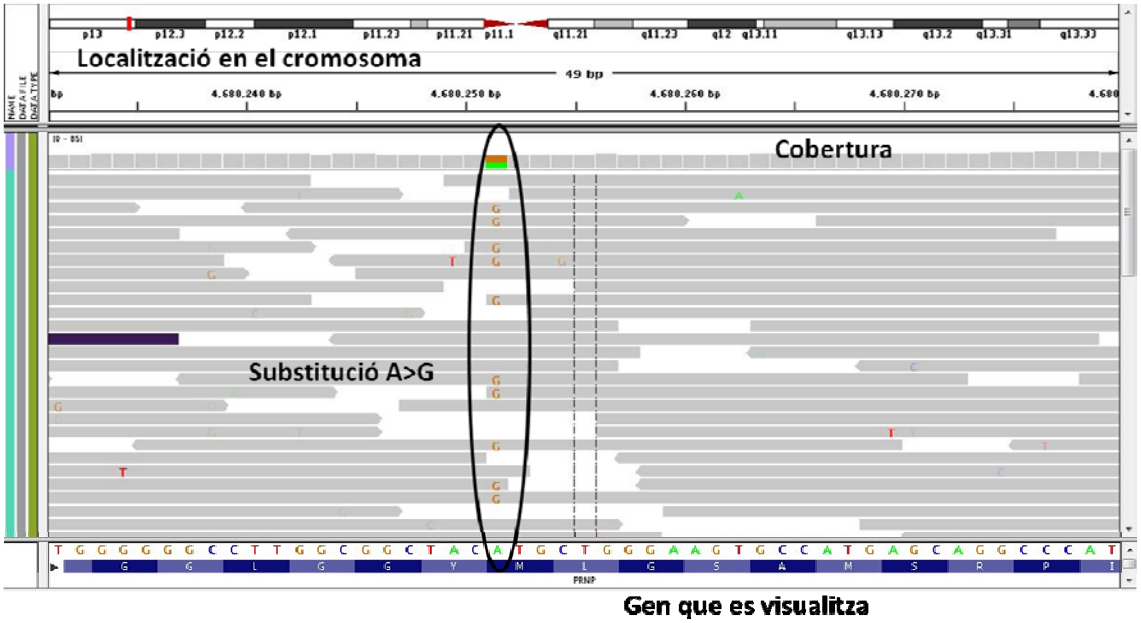


Figura 3. Visualització d'una substitució A>G mitjançant l'IGV. En la imatge s'indica la localització del gen i la variant en el cromosoma, la cobertura, el gen i els aminoàcid on està situada la variant, i la substitució en diferents colors que està present aproximadament en el 50% de les lectures.

-*Software* per a l'anotació de variants: Un cop tenim els llistats de variants candidates és imprescindible anotar-les, és a dir, identificar el gen on estan, en quina posició es troben, el canvi de codó que produeixen en casos de variants exòniques, o reportar si les variants ja han estat descrites prèviament com a polimorfismes o com a variants somàtiques, entre d'altres. Anotar les variants és un procés laboriós, ja que hi ha molta informació a les bases de dades públiques i és necessari filtrar allò que ens interessa. L'anotació es pot fer variant a variant, però també hi ha programes que permeten anotar les variants de forma sistemàtica, un exemple és l'eina Annovar (Wang et al. 2010). També hi ha eines en format web que fan aquesta tasca com el SeattleSeqAnnotation (Ng et al. 2009, SeattleSeq Annotation 137 2016). Tot i això, un cop anotades les variants, és necessari seleccionar la informació més adequada, com per exemple el transcrit d'interès, per a concloure els resultats biològics més rellevants.

A la taula 2 es resumeix el *software* lliure més utilitzat per a l'anàlisi de NGS amb una breu descripció.

Taula 2. Alguns dels programes lliures més utilitzats per a l'anàlisi de dades de NGS segons la seva funció.

Funció	Nom	Descripció	URL
Control de qualitat	**FastQC**	Analitza la qualitat de les dades de NGS. Proporciona els resultats organitzats en mòduls que indiquen de manera ràpida els aspectes on les dades podrien tenir algun problema.	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	FaQCs	Processa grans quantitats de dades de NGS i proporciona resultats sobre la qualitat de les dades, però també permet convertir formats FASTQ. Una de les seves fortaleses és la capacitat de processar les dades en paral·lel en diversos nodes.	http://faqcs.readthedocs.org/en/latest/
	NGS QC Toolkit	Inclou eines amigables per a l'usuari per al control de qualitat de les dades de Roche 454 i d'Illumina proporcionant els resultats en forma de taules i gràfics, i filtrant les dades amb bona qualitat.	http://www.nipgr.res.in/ngsqctoolkit.html
Alineament	**BLAT/BLAST/BFAST**	Eina per a l'alineament local de seqüències de fins a 200000 bases. BLAST és una versió millorada de BLAT més sensible i més ràpida que permet alinear major nombre de seqüències, i BFAST és la versió adaptada a dades de NGS.	http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
	Bowtie2	Molt ràpida i eficient en l'ús de memòria per alinear lectures sobre una seqüència de referència. Permet alineaments amb forats, alineaments locals i alineaments amb lectures aparellades.	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
	BWA/BWA-SW/BWA-MEM	BWA és un *software* per a alinear seqüències poc divergents contra un genoma de referència gran com l'humà. Conté tres algoritmes diferents, el BWA per a lectures de fins a 70 bases, el BWA-SW per a lectures llargues i amb forats, i el BWA-MEM per a lectures llargues i amb alineament local.	http://bio-bwa.sourceforge.net/
	SHRIMP	Alinea lectures genòmiques contra un genoma de referència. Es va desenvolupar per a lectures curtes de NGS i per a lectures de la tecnologia SOLID en format d'espai-color (*colorspace*).	http://compbio.cs.toronto.edu/shrimp/
	GEM	Permet buscar el millor alineament jugant amb la precisió, ja que reporta totes les possibles opcions segons els paràmetres requerits. És més ràpid que altres alineadors com Bowtie2 o BWA.	http://big.crg.cat/services/gem_genome_multi_tool_library
Tractament de fitxers d'alineament	**SAMTools**	Conjunt d'eines per a manipular els alineaments en format SAM, incloent l'ordenació, la combinació o la indexació, entre d'altres.	http://samtools.sourceforge.net/
	Picard	Programat en Java, permet manipular els fitxers SAM i permet crear nous programes que escriguin i llegeixin fitxers SAM.	http://picard.sourceforge.net/

Detecció de variants	**SAMTools**	Conjunt d'eines per a manipular els alineaments en format SAM, que entre d'altres funcions també permet detectar variants a partir de l'alineament.	http://samtools.sourceforge.net/
	GATK	Desenvolupat al Broad Institute per analitzar dades de reseqüenciació de NGS. Enfocat principalment a detectar variants i a genotipar, així com al control de qualitat de les dades de NGS.	http://www.broadinstitute.org/gatk/
	VARScan	Detecta variants en dades de NGS. Utilitza un model heurístic i estadístic molt robust reportant paràmetres com el *coverage*, la qualitat de les bases i la significació estadística.	http://varscan.sourceforge.net/
	Pindel	Identifica insercions, delecions i variacions estructurals en dades de NGS.	http://gmt.genome.wustl.edu/pindel/current/
Visualització	**IGV**	Eina de visualització per a l'exploració interactiva de grans bases de dades genòmiques. Permet visualitzar dades de diversos tipus, entre elles dades d'arrays, dades de NGS i dades d'anotacions genòmiques.	http://www.broadinstitute.org/igv/
	LookSeq	Aplicació web que permet la visualització d'alineaments i l'anàlisi de dades de seqüenciació de genomes.	http://www.sanger.ac.uk/resources/software/lookseq/
Anotació de variants	**Annovar**	Utilitza eficientment informació actualitzada de diverses bases de dades per a anotar variants genètiques detectades en diferents genomes com l'humà, el de ratolí, el de cuc o el de mosca, entre molts altres.	http://www.openbioinformatics.org/annovar/

2.3 Requeriments computacionals per a l'anàlisi de dades de NGS

Com hem comentat prèviament, el procés computacional de l'anàlisi de dades de NGS és intensiu. La quantitat de memòria que ocupen els arxius, i el temps d'anàlisi depèn principalment de la regió del genoma seqüenciada i de la cobertura amb què s'ha seqüenciat, també depèn de la llargada de les lectures. Per tal d'estalviar espai en els discos s'acostumen a emmagatzemar els arxius d'alineament BAM, ja que a partir d'aquests, es poden recuperar les dades originals així com reproduir totes les anàlisis posteriors. A la taula 3 es mostren alguns exemples de la mida que tenen els arxius BAM d'alguns experiments en funció de diferents paràmetres.

Taula 3. Relació de les mides dels arxius d'alineament per a diferents experiments.

Proporció del genoma seqüenciada	Cobertura	Número de lectures	Llargada de les lectures	Mida de l'arxiu d'alineament
Tot el genoma	28x	3200000000	75	88 Gb
Exoma	40x	65000000	75	8 Gb
Gens *BRCA1* i *BRCA2*	500x	114300	400	97 Mb

De la mateixa manera, el temps per a analitzar les dades d'un experiment depèn de la mida, però també depèn del tipus de processador de l'ordinador, del *software* que s'utilitzi, i de si és possible paral·lelitzar la tasca en diferents processadors.

Així, els factors més limitants a l'hora d'analitzar dades de NGS són: la capacitat del processador que fa variar el temps d'anàlisi, la memòria RAM que permet analitzar més o menys quantitat de dades, i la capacitat d'emmagatzematge. Tot i que aquests factors són més o menys limitants depenent del tipus d'experiment que es vulgui analitzar. Si l'objectiu és analitzar un genoma sencer, el processador i la memòria RAM haurà de ser molt superior que si es volen analitzar un conjunt de pocs gens, de la mateixa manera, el temps d'anàlisi no serà el mateix. Tampoc s'utilitza el mateix temps si es volen analitzar 10 mostres o 1000, ja que el temps es multiplica. Per tal d'emmagatzemar les dades, és important disposar de discos amb prou capacitat per guardar especialment les dades generades pels seqüenciadors. A més, és important que aquests discos d'emmagatzematge tinguin còpies de seguretat i estiguin protegits ja que el contingut són dades genètiques susceptibles i confidencials en la majoria de casos. Abans de realitzar experiments cal plantejar bé quins requeriments informàtics seran necessaris.

3. Anàlisi de variants i càncer

Les variants en la seqüència de DNA implicades en l'oncogènesi poden ser somàtiques, germinals o ambdues (Futreal et al. 2004). Les mutacions somàtiques en el cas de tumors són aquelles que estan presents només a les cèl·lules canceroses, aquestes mutacions no es transmeten a la descendència. En canvi, les mutacions germinals són aquelles que afecten a les cèl·lules productores dels gàmetes, de manera que aquestes mutacions sí que es transmeten a la següent generació, que les portarà a totes les cèl·lules. Aproximadament s'estima que el 80% dels gens relacionats amb el càncer presenten

mutacions somàtiques, el 10% presenten mutacions germinals, i el 10% restant presenten mutacions tant somàtiques com germinals (Futreal et al. 2004).

3.1 Tipus de variants genètiques

Existeixen diferents tipus de variants de seqüència, i malgrat encara no existeix una metodologia de detecció única que pugui cobrir tot l'espectre mutacional en un gen concret, actualment la NGS permet detectar la majoria dels tipus de variants.
Les variants de seqüència es poden classificar en funció del tipus de canvi en la seqüència segons siguin: substitucions d'un únic nucleòtid, petites insercions i delecions, o variacions estructurals.

- **Substitucions d'un únic nucleòtid:** canvis d'una sola base (en anglès *single nucleotide substitution*, SNS), es reemplaça un nucleòtid per una altre de diferent.
- **Petites insercions i delecions:** A més de les substitucions, un nucleòtid també pot estar deleccionat, això és que desapareix, o bé es pot insertar en el DNA. Les insercions o delecions afegeixen o treuen un o més nucleòtids al DNA. A les petites insercions i delecions acostuma a fer-se'n referència com a *INDELs*.
- **Variacions estructurals:** els genomes no només varien per aquestes diferències d'un o pocs nucleòtids, també hi poden haver insercions i delecions de centenars de milions de bases consecutives, inversions o translocacions. Les grans duplicacions o insercions de més d'una kilobase es coneixen com a Variacions en el Nombre de Còpies (*Copy Number Variations*, CNV). Estudis recents estimen que les CNV composen fins el 20% de totes les variants genètiques en humans. Així, per a un individu qualsevol, les variacions estructurals poden constituir del 0,5 a l'1% del genoma, això és entre 9 i 25Mb (Redon et al. 2006). A més podem trobar inversions genòmiques, un tipus de variant estructural que fa canviar el sentit d'un segment en el cromosoma, o també les translocacions, que intercanvien parts de cromosomes no homòlegs.

Les variants poden estar en zona intrònica o en zona codificant, quan estan en zona codificant, si el canvi que produeixen en la seqüència de DNA del gen fa variar l'aminoàcid codificat reben el nom de no-sinònimes (en anglès *nonsynonimous*). Les substitucions no-sinònimes poden ser mutacions amb error de sentit (en anglès *missense mutations*) si el canvi fa que el DNA resultant codifiqui per un altre aminoàcid, o poden ser mutacions sense sentit (en anglès *nonsense mutations*) si codifica per a un codó *stop* que truncarà la proteïna. Les petites insercions o delecions no múltiples de tres provoquen canvis en la pauta de lectura (*frameshift*) provocant la generació d'un codó de terminació prematur (*stop codon*) que donarà lloc a una proteïna truncant.

Les variants de seqüència identificades en qualsevol gen també es poden classificar segons la seva significació clínica en: Polimorfismes, Variants Patogèniques i Variants de Significat Desconegut.

- **Polimorfisme:** A nivell general es considera que una variant de seqüència de DNA amb una freqüència igual o superior a l'1% de la població és un polimorfisme genètic. En la majoria de síndromes de càncer hereditari, els polimorfismes en gens d'alt risc són considerats sense significació clínica. També es poden considerar com a variants sense efecte clínic les variants amb una freqüència inferior a l'1% però on estudis clínics, genètics o bioquímics descarten la seva significació clínica. Si un polimorfisme només afecta a una base s'anomena polimorfisme d'un únic nucleòtid (*Single Nucleotide Polymorphism*, SNP) i és la classe de variacions genètiques més prevalent entre els humans (Frazer et al. 2009). El genoma humà conté uns 15 milions d'SNP amb una freqüència de l'al·lel menor (Minor Allele Frequency, MAF) per sobre de l'1%. Això significa que dos genomes humans qualssevol es diferencien aproximadament en 1 nucleòtid per cada 1331 pb (Kruglyak and Nickerson 2001), tot i que el nombre de variacions és diferent segons la població.
- **Variants Patogèniques:** Es classifiquen com a patogèniques les variants que afecten a la funcionalitat de la proteïna codificada.
- **Variants de Significat Desconegut:** En general quan un canvi genètic no està present en més de l'1% de la població i no existeix prou evidència per a determinar el seu efecte a nivell proteic i funcional, aquest canvi es classifica com una variant de significat desconegut (VSD). La rellevància clínica de les VSD és un dels grans reptes que presenta la realització d'estudis genètics. Normalment, per a classificar aquest grup de variants, s'utilitzen algoritmes que integren diferents graus d'informació sobre la variant en qüestió, com estudis de cosegregació de la variant, anàlisi en població general o estudis funcionals, entre d'altres.

3.2 Anàlisi genètica del càncer hereditari

Es considera que la majoria dels casos de càncer són esporàdics degut a factors ambientals i mutacions somàtiques. Tanmateix, en un petit percentatge de casos s'observa un cert grau d'agregació familiar i se'n fa referència com a càncer familiar; en una petita fracció d'aquests casos familiars s'observa una herència clarament mendeliana, aquest casos es classifiquen com casos de càncer hereditari. Els casos de càncer hereditari solen presentar una sèrie de criteris clínics i/o moleculars específics com són: edat d'aparició primerenca, multifocalitat de les lesions o lesions bilaterals, aparició de més d'un tumor primari en el mateix individu, antecedents familiars de la mateixa neoplàsia, alta incidència de càncer dins de la mateixa família, o associació de tumors amb malformacions congènites o retard mental. Qualsevol d'aquestes situacions pot alertar sobre la possibilitat de trobar-se davant d'un cas de càncer hereditari i sobre la conveniència de dirigir el pacient i la seva família a una unitat de consell genètic (Lemke et al. 2012). La Societat Americana d'Oncologia Clínica (*American Society of Clinical Oncology,*

ASCO) recomana que la determinació genètica en el càncer hereditari s'ofereixi quan els resultats del test genètic es puguin interpretar amb fiabilitat i puguin tenir repercussió en el seguiment i tractament del pacient o dels seus familiars (American Society of Clinical Oncology 2016).

Les mutacions responsables del càncer hereditari són generalment mutacions germinals (existeixen alguns casos de mosaïcisme genètic), afecten a les cèl·lules productores dels gàmetes transmetent-se a la següent generació, i estan presents a totes les cèl·lules de l'individu. Així, per exemple, mutacions germinals en *BRCA1* i *BRCA2* són la principal causa hereditària del càncer de mama i donen lloc a la síndrome de càncer de mama i ovari hereditari (*en anglès Hereditary Breast and Ovarian Cancer, HBOC*), caracteritzada per la presència de casos de càncer de mama i ovari en una mateixa família, o de múltiples casos de càncer de mama precoç. De la mateixa manera, mutacions germinals en els gens *APC* o *MUTYH* estan relacionades amb la poliposi adenomatosa familiar (en anglès, *Familial Adenomatous Polyposis, FAP*), o amb la poliposi associada a *MUTYH* (*en anglès MUTYH associated polyposis, MAP*), respectivament, que tenen associat un alt risc de desenvolupar càncer colorectal. El càncer colorectal hereditari no polipòsic (*Hereditary non-polyposis colorectal cancer, HNPCC*) normalment s'associa a la identificació de mutacions germinals en els gens reparadors *MLH1, MSH2, MSH6* o *PMS2*.

En general, els tipus de càncer associats a mutacions germinals en un gen particular són similars a aquells càncers causats per mutacions somàtiques en el mateix gen. Tot i això hi ha excepcions, com per exemple, el gen *TP53*, les mutacions somàtiques del qual es troben en la meitat dels càncers colorectals, però on les mutacions germinals no sembla que causin predisposició hereditària al càncer colorectal. També hi ha gens amb mutacions germinals que predisposen al càncer i que gairebé no presenten mutacions somàtiques en càncers esporàdics del mateix tipus, un exemple són els gens *BRCA1 i BRCA2* en càncer de mama (Futreal et al. 2004).

Tot i que les síndromes de càncer hereditari es consideren malalties rares (afecten a menys d'un individu de cada 2000), suposen el 5-10% dels casos de càncer i el seu estudi ha permès grans avenços en el coneixement de l'etiologia de formes més comunes de càncer (Figura 4).

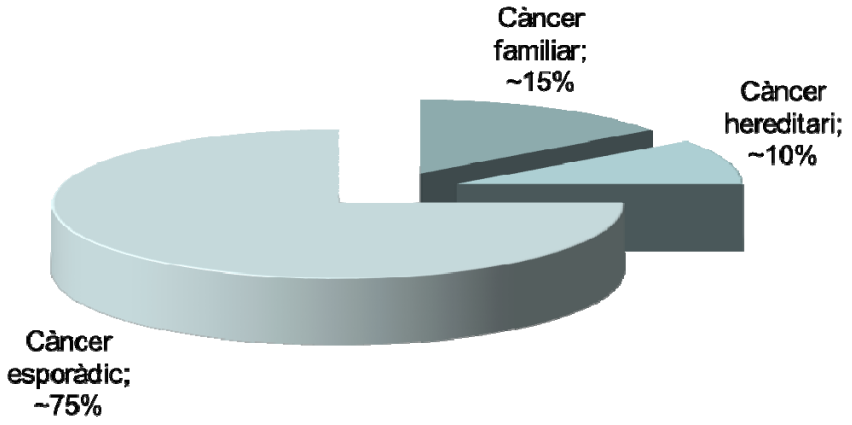


Figura 4. Distribució dels tipus de càncer segons el seu origen.

Fins ara, les unitats de diagnòstic de càncer hereditari a Catalunya solen oferir l'anàlisi mutacional en alguns gens concrets, dels que se'n coneix una clara associació entre mutació i probabilitat de desenvolupar càncer. Amb els ràpids avenços de les tecnologies i la reducció de costos de la NGS, s'estan implementant panells de gens més extensos, del voltant de 100 gens, que permetran analitzar més quantitat de gens i associacions a costos similars (LaDuca et al. 2014, Petric et al. 2015).

3.3 Genomes i exomes per a la recerca del càncer

Com s'ha comentat prèviament, la majoria dels càncers en humans sorgeixen de tumors esporàdics on s'acumulen alteracions somàtiques en el DNA. Per tant, la identificació d'aquests canvis genètics que condueixen al desenvolupament del càncer donen coneixement de la malaltia i acceleren el descobriment de noves tècniques per al diagnòstic o la predicció de risc, o fins i tot de noves dianes terapèutiques. Projectes com el "Human Genome Project", que el 2003 va proporcionar una seqüència consens del genoma humà, han facilitat molt l'estudi de la biologia del càncer. Amb la seqüenciació de nova generació ha sigut possible seqüenciar genomes sencers de teixits normals i tumorals aparellats, i caracteritzar la malaltia en trobar les diferències. Per exemple, el projecte "The Cancer Genome Atlas" (TCGA) (NIH 2016), entre d'altres tasques, està seqüenciant centenars de genomes de tumors diferents per a identificar gens relacionats amb diferents tipus de càncer. En molts casos no cal reseqüenciar tot el genoma, doncs malgrat que les regions codificants constitueixen només l'1-2% del genoma humà, s'estima que contenen el 85% de les mutacions implicades en malalties (Choil et al. 2009). Per tant, combinant les tecnologies de seqüenciació amb la captura de regions específiques com l'exoma, es redueix molt el cost i es poden descobrir eficientment mutacions relacionades amb la malaltia (Thompson et al. 2012). Actualment hi ha grans consorcis internacionals, com l'"*International Cancer Genome Consortium*" (ICGC 2016), que tenen l'objectiu de crear un catàleg amb els gens responsables del desenvolupament i la progressió del càncer. El principal projecte espanyol d'aquest consorci està relacionat amb la recerca d'un grup de leucèmies anomenat "*Chronic Lymphocytic Leukemia - CLL with mutated and unmutated IgVH*" (Puente et al. 2011). Aquest consorci ha pogut identificar mutacions recurrents en leucèmies limfàtiques cròniques mitjançant la seqüenciació de genomes i exomes.

OBJECTIUS

Objectiu general

L'objectiu principal d'aquesta tesi és el desenvolupament i l'aplicació d'estratègies bioinformàtiques per a l'anàlisi de dades de NGS en el diagnòstic molecular del càncer hereditari i en la recerca de les bases genètiques del càncer colorectal esporàdic.

Objectius específics

- Dissenyar un algoritme d'anàlisi de les dades provinents de la seqüenciació de diferents gens amb la plataforma de NGS anomenada GS Junior i realitzar una prova de concepte. (*Article 1*)

- Crear una aplicació per a l'anàlisi de les dades provinents del GS Junior per detectar variacions, amb la nomenclatura de la HGVS (Human Genome Variation Society) i la posterior interpretació de les variants, en els gens de susceptibilitat al càncer hereditari que s'estudien a la Unitat de Diagnòstic Molecular del Programa de Càncer Hereditari de l'ICO (*BRCAs, APC, MUTYH, MLH1, MSH2, MSH6 i PMS2*). (*Article 2*)

-Avaluar el rendiment dels panells de múltiples gens com a estratègia en el diagnòstic del càncer hereditari. (*Apartat Altres contribucions*)

- Aplicar les tècniques d'anàlisi bioinformàtiques a les dades de seqüenciació d'exomes per a la recerca de les bases genètiques del càncer colorectal esporàdic. (*Article 3*)

RESULTATS

RESULTATS

Els resultats obtinguts del treball realitzat en aquesta tesi estan inclosos en diversos articles científics. En aquesta memòria els resultats es descriuen breument i seguidament s'adjunten els articles.

NGS per al diagnòstic genètic del càncer hereditari

Article 1

"Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of *BRCA1* and *BRCA2* genes"

Article 2

"ICO Amplicon NGS Data Analysis: A Web Tool for Variant Detection in common High-Risk Hereditary Cancer Genes Analyzed by Amplicon GS Junior Next-Generation Sequencing"

Altres contribucions

"Avaluació de l'eficiència del panell de gens Trusight Cancer d'Illumina com a eina de diagnòstic genètic"

NGS per a la recerca del CCR esporàdic

Article 3

"Exome sequencing reveals *AMER1* as a frequently mutated gene in colorectal cancer"

NGS per al diagnòstic genètic del càncer hereditari

ARTICLE 1

Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of *BRCA1* and *BRCA2* genes

Lídia Feliubadaló*, **Adriana Lopez-Doriga***, Ester Castellsague*, Jesus del Valle, Mireia Menéndez, Eva Tornero, Eva Montes, Raquel Cuesta, Carolina Gómez, Olga Campos, Marta Pineda, Sara Gonzalez, Víctor Moreno, Joan Brunet, Ignacio Blanco, Eduard Serra, Gabriel Capellá and Conxi Lázaro

(*) Aquests autors han contribuït equitativament a aquest treball

Resum del treball: La seqüenciació de nova generació (*Next Generation Sequencing*, NGS) està canviant el diagnòstic genètic gràcies a la gran capacitat de seqüenciació i el seu cost-eficiència. L'objectiu d'aquest estudi ha estat desenvolupar un protocol basat en la NGS per al diagnòstic rutinari de la síndrome de càncer de mama i ovari hereditari (*Hereditary Breast and Ovarian Cancer Syndrome*, HBOCS), per millorar el diagnòstic genètic dels gens *BRCA1* i *BRCA2*. Es va dissenyar un protocol basat en NGS utilitzant les llibreries d'amplicons del MASTR kit de Multiplicom, seguit de la piroseqüenciació amb el GS Junior. Per a l'anàlisi de les dades s'utilitzava un *software* lliure, el *Variant Identificantion Pipeline* (VIP), combinat amb funcions pròpies programades en R, on es passaven un seguit de filtres i es generaven informes de la cobertura i de les variants. En paral·lel es realitzava un assaig per als homopolímers dels *BRCA*. Es va utilitzar un grup d'entrenament amb 28 mostres de DNA que contenien 23 mutacions úniques patogèniques i 204 variants (33 úniques) d'altres tipus. El protocol es va validar en un grup de 14 mostres de famílies amb HBOCS que es van analitzar en paral·lel juntament amb el protocol utilitzat en aquell moment. El nou protocol desenvolupat basat en NGS va permetre identificar totes les mutacions patogèniques i totes les altres variants genètiques, incloses les variants contingudes o properes a homopolímers. El nou protocol demostra una sensibilitat i especificitat màximes, necessàries per al diagnòstic genètic de HBOCS i millora el protocol utilitzat fins ara en cost-eficàcia.

Contribució de la doctoranda: En aquest estudi la doctoranda ha participat realitzant les anàlisis bioinformàtiques, des de l'estudi dels diferents programes disponibles fent les proves pertinents amb cadascun d'ells per a l'anàlisi de les dades, fins a triar el VIP com al millor *software* i realitzar les anàlisis bioinformàtiques de totes les carreres. Ha realitzat un estudi molt acurat dels filtres més adients per a descartar falsos positius i per a determinar els passos del protocol d'anàlisi. Ha programat unes funcions amb R per tal de recollir la informació que genera el VIP i crear els informes, un informe amb els resultats de la cobertura i amb els gràfics associats, i un altre amb el llistat de variants amb la

nomenclatura en CDS corresponent. Finalment ha participat en l'escriptura de l'article i la preparació de les taules i figures.

European Journal of Human Genetics (2012), 1–7

www.nature.com/ejhg

ARTICLE

Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of *BRCA1* and *BRCA2* genes

Lídia Feliubadaló[1,6], Adriana Lopez-Doriga[2,6], Ester Castellsagué[1,6], Jesús del Valle[1], Mireia Menéndez[1], Eva Tornero[1], Eva Montes[1], Raquel Cuesta[1], Carolina Gómez[1], Olga Campos[1], Marta Pineda[1], Sara González[1], Victor Moreno[3], Joan Brunet[4], Ignacio Blanco[1], Eduard Serra[5], Gabriel Capellá[1] and Conxi Lázaro*[,1]

Next-generation sequencing (NGS) is changing genetic diagnosis due to its huge sequencing capacity and cost-effectiveness. The aim of this study was to develop an NGS-based workflow for routine diagnostics for hereditary breast and ovarian cancer syndrome (HBOCS), to improve genetic testing for *BRCA1* and *BRCA2*. A NGS-based workflow was designed using BRCA MASTR kit amplicon libraries followed by GS Junior pyrosequencing. Data analysis combined Variant Identification Pipeline freely available software and *ad hoc* R scripts, including a cascade of filters to generate coverage and variant calling reports. A *BRCA* homopolymer assay was performed in parallel. A research scheme was designed in two parts. A Training Set of 28 DNA samples containing 23 unique pathogenic mutations and 213 other variants (33 unique) was used. The workflow was validated in a set of 14 samples from HBOCS families in parallel with the current diagnostic workflow (Validation Set). The NGS-based workflow developed permitted the identification of all pathogenic mutations and genetic variants, including those located in or close to homopolymers. The use of NGS for detecting copy-number alterations was also investigated. The workflow meets the sensitivity and specificity requirements for the genetic diagnosis of HBOCS and improves on the cost-effectiveness of current approaches.
European Journal of Human Genetics advance online publication, 19 December 2012; doi:10.1038/ejhg.2012.270

Keywords: Next-generation sequencing; hereditary breast and ovarian cancer syndrome; *BRCA1*; *BRCA2*; genetic testing; molecular diagnostics

INTRODUCTION

Next-generation sequencing (NGS) is an increasingly used technology that generates up to gigabases of DNA reads at high speed and with low cost per base. This high-throughput technology, based on massively parallel sequencing of spatially separated DNA molecules, is currently used with several available platforms, such as the Genome Sequencer (Roche-454 Life Sciences, Indianapolis, IN, USA), the Genome Analyzer/HiSeq/MiSeq (Illumina-Solexa, San Diego, CA, USA), the SOLiD System, Ion PGM/Ion Proton (Ion Torrent-Invitrogen, Carlsbad, CA, USA), and the HeliScope from Helicos BioSciences (Cambridge, MA, USA).[1,2] In Roche-454 technology, bead-attached DNA fragments clonally amplified in a water-in-oil emulsion (emulsion PCR) are deposited in single-bead capacity wells of a plate over which nucleotides flow sequentially, releasing chemiluminescence only when a nucleotide is correctly incorporated (pyrosequencing). In molecular diagnostics, targeted genomic resequencing of pooled samples from different individuals benefits from the high throughput achieved by using NGS technology. To enrich the target fragments to be resequenced in this type of gene-centric approach, PCR-based methods are generally used.[3,4] *BRCA1* and *BRCA2* are the two main highly penetrant genes that predispose to hereditary breast and ovarian cancer syndrome (HBOCS).[5] Molecular diagnosis of HBOCS is essential for the provision of genetic counseling and to establish preventive screening and therapeutic strategies.[6] Although direct Sanger sequencing is considered the gold standard for the analysis of *BRCA1* and *BRCA2* mutations, their large size (5592 bp and 10257 bp, respectively), and lack of mutation hot spots (see Breast Cancer Information Core database: http://www.research.nhgri.nih.gov/bic/) mean useful prescreening strategies.[7–9] Moreover, large genomic rearrangements (LGRs) of these genes require the use of other complementary techniques.[10,11] The development of cost-effective *BRCA* mutation detection workflows will not only benefit the genetic counseling process for patients with HBOCS but will also enhance the process of selecting patients for personalized treatments, as could be the case of PARP inhibitors, for example. Mutation analyses of *BRCA1* and *BRCA2* using NGS have been already performed for high-capacity NGS platforms, such as the 454 FLX (Roche),[12] the Helicos (Heliscope),[13] the Genome Analyzer (Illumina)[4] and, very recently, the GS Junior instrument.[14] Most of these studies used large-capacity

[1]Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [2]Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [3]Prevention Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [4]Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IdIBGi), Girona, Spain; [5]Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Badalona, Barcelona, Spain
[6]These authors contributed equally to this work.
*Correspondence: Dr C Lázaro, Hereditary Cancer Program, Molecular Diagnosis Unit, Laboratori de Recerca Translacional, Institut Català d'Oncologia (ICO-IDIBELL), Hospital Duran i Reynals, Gran Via 199-203, L'Hospitalet de Llobregat, 08908 Barcelona, Spain. Tel: +34 932607342; Fax: +34 932607466; E-mail: clazaro@iconcologia.net
Received 9 July 2012; revised 28 September 2012; accepted 13 November 2012

platforms that generally exceed the demand of most mid-sized genetic testing laboratories and whose approaches are difficult to translate to benchtop next-generation sequencers. Only one of the studies used small-scale equipment, the GS Junior, but the number of samples tested is very small and no discussion is offered regarding how to overcome the main problem associated with pyrosequencing, that is, DNA lectures in homopolymeric regions.[14] Here, we present a rigorous sensitivity and specificity analysis of our newly established HBOCS workflow for genetic testing of *BRCA* genes using a small-capacity next-generation instrument. We present data from a Training Set and from a Validation Set of samples. We demonstrate that a combined approach using the GS Junior platform and an specific assay for homopolymeric tracts with a custom bioinformatics pipeline provides accurate results that can be used for genetic diagnosis.

MATERIALS AND METHODS

Samples analyzed

In our unit, a multistep workflow including conformation-sensitive capillary electrophoresis[9] as a prescreening method for analysis of *BRCA* mutations was used (Supplementary Figure 1). A total of 28 DNA samples previously characterized by this workflow were used as a Training Set to setup our NGS workflow, and 14 new DNAs were used as a Validation Set (see Experimental design in the Results section). To properly compare NGS with our workflow, only variants in heterozygosity were considered (as homozygous variants are not detected by conformation-sensitive capillary electrophoresis). This study was approved by our Institutional Review Board.

Multiplex PCR-based target amplification and resequencing

Target amplification of *BRCA1* and *BRCA2* was achieved using BRCA MASTR assays following manufacturer's instructions (http://www.multiplicom.com). Several versions of the kit were used as they were released. Briefly, the assay generates a library of specific amplicons in two rounds of PCR: a first multiplex PCR that amplifies the target sequences; and a second PCR to attach MID (Multiplex Identifier) barcodes and 454 adapters to each amplicon. The barcoded multiplex products were assessed by fluorescent labeling and capillary electrophoresis, and quantified using Quant-iT PicoGreen (Invitrogen). Then, PCRs from different individuals were equimolarly pooled and purified using AgencourtAMPure XP (Beckman Coulter, Beverly, MA, USA) and PicoGreen quantified. Emulsion PCR of the combined purified libraries was carried out using the GS Junior Titanium emPCR Kit (Lib-A) and pyrosequenced on GS Junior following manufacturer's instructions (Roche).

Data analysis

Reads from the GS Junior sequencer were analyzed with the open source software Variant Identification Pipeline (VIP) version 1.4.[15] Using VIP, the reads from each sample were demultiplexed and then aligned against *BRCA1* NG_005905.2 and *BRCA2* NG_012772.1 reference sequences using the BLAT algorithm.[16] Results from VIP were then processed using R (A Language and Environment for Statistical Computing) commands. Specific primers from each amplicon were trimmed and identified variants were annotated according to the Human Genome Variation Society (HGVS) nomenclature recommendations version 2.0 (http://www.hgvs.org/mutnomen/). Two reports were obtained: a coverage report, listing low-coverage fragments indicated for further Sanger sequencing; and a variant report. Intronic variants located deep inside introns (after position +20 of the donor site and before position −50 of the acceptor site) were not included in the variant report. Multiple alignments of reads for each MID and amplicon were visualized with the GS Amplicon Variant Analyzer v2.7 (AVA) software (Roche). Scripts are available upon request (Lopez-Doriga *et al*, manuscript in preparation).

We also evaluated the capacity to detect LGRs. Eight samples with known rearrangements were tested in three different runs. One of the samples was included in the Validation Set, and the other seven were added later. The known LGRs consist of: deletion of exons 1–2, deletion of exons 1–13, deletion of exon 14, deletion of exon 20, deletion of exon 22, and duplication of exons 9–24 in *BRCA1*, and deletion of exons 1–24 and deletion of exon 2 in *BRCA2*. To assess copy number for each amplicon, a methodology described elsewhere was applied.[3] Briefly, the relative read count of an amplicon was determined as the ratio of the read count for that amplicon over the sum of all gene amplicons for the other gene in the specific multiplex to which the amplicon belongs. Hence, to analyze *BRCA1* amplicons, we used the sum of *BRCA2* amplicons from the same multiplex, and vice versa. Next, intersample normalization was performed, dividing each ratio by the average of the control samples in the same experiment (at least three controls were used).

Homopolymer analysis

To treat homopolymers, the BRCA HP v2.0 (Multiplicom, Niel, Belgium) assay was used. This kit targets all *BRCA1*- and *BRCA2*-coding homopolymer stretches of 6 bp or longer by producing 29 PCR products in two multiplex reactions. Fragment length was assessed by capillary electrophoresis (3730 ABi sequencer, Applied Biosystems, Foster City, CA, USA) and visualized with the MAQ-S software (Multiplicom).

Sanger sequencing

All fragments with coverage under 38× and all non-polymorphic DNA variants identified were sequenced by Sanger.

RESULTS

Experimental design

The Training Set (28 samples analyzed in two experiments) contained 23 unique pathogenic mutations and 204 (33 unique) non-pathogenic mutations or mutations with unknown significance DNA variants (Supplementary Table 1) (Figure 1). In the Validation Set, 14 samples were blindly sequenced together with a sample containing a multi-exon duplication in *BRCA1* (Figure 1). To better assess the usefulness of this approach to detect LGR, a set of seven positive samples showing LGRs were also analyzed.

Workflow setup

In experiment 1, 28 samples were amplified with the BRCA MASTR v1.2 kit (170 amplicons, Multiplicom) in four GS Junior runs (R1-R4) (7 patients per run). Only 0.5% of the passed reads was lost, due to short length, low quality or incorrect MIDs or primer sequences, and did not map in the reference sequence. While experiment 1 was being conducted, Multiplicom released a new kit (v2.0, 94 amplicons), which was used in experiment 2 to reanalyze 14 samples from experiment 1 in two runs (R5–R6).

Coverage analysis of the Training Set

The coverage of each run was evaluated (Table 1). In experiment 1, the average mean base coverage was 69 ± 27. The coverage for the various MIDs used (MID1–MID15) did not exhibit any significant difference (data not shown). The number of mapped reads in R5 and R6 was similar to the runs in experiment 1, but coverage was substantially increased (127 ± 53) due to the lower number of amplicons. Of the 24 undercovered amplicons (coverage <38), 14 belonged to amplicon *BRCA1*_exon7 from different patients (Supplementary Figure 2A).

Filters and variant calling in the Training Set

Next, identification of all the variants was investigated. First, each experiment was analyzed alone (data not shown), then the results were combined as the Training Set, incorporating into experiment 2 samples not repeated from experiment 1 (to avoid bias due to duplication of samples). In total, 4260 variants were identified, of

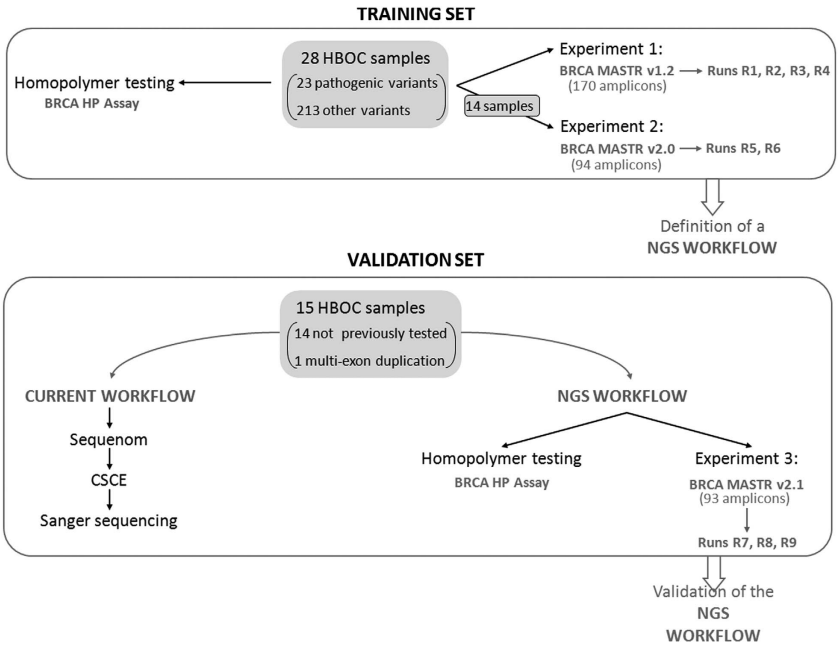


Figure 1 Experimental design. Our study was divided into two parts: the Training Set and the Validation Set. In the Training Set, 28 HBOCS samples, already analyzed by our current diagnostic workflow, were assessed (Supplementary Figure 1). Of this group, 23 samples contained a variety of pathogenic mutations, including challenging insertions and deletions, inside and outside homopolymeric regions, as well as a subset of non-pathogenic variants. The remaining 5 samples belonged to affected individuals from high-risk HBOCS families, in whom no pathogenic mutation had been found after applying our current multistep protocol. In total, this subset of 28 samples contained 23 unique pathogenic mutations and 213 (33 unique) non-pathogenic DNA variants (Supplementary Table 1). The Training Set was subjected to two different experiments: in experiment 1, all 28 samples were amplified using BRCA MASTR v1.2 and sequences in 4 runs; in experiment 2, 14 of the DNAs from experiment 1 were used but they were amplified with the newly released kit (v2.0) and sequenced in two runs. In parallel, homopolymeric regions of all samples were studied with the BRCA HP kit. Thanks to the Training Set experiment, we were able to define an NGS workflow for the genetic analysis of *BRCA* genes in the HBOCS diagnostic routine. In the Validation Set, we assessed a total of 15 HBOC samples, 14 not previously tested and the remaining 1 containing a multi-exon duplication. These samples were analyzed in parallel with our current diagnostic workflow and with the newly designed NGS workflow. In this case, experiment 3 was carried out using the most recent version (v2.1) of the BRCA MASTR kit and samples were sequenced in three runs.

Table 1 Overall coverage results

	Experiment 1 (BRCA MASTR v1.2)				*Experiment 2 (BRCA MASTR v2.0)*		*Experiment 3 (BRCA MASTR v2.1)*		
Run	*R1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*	*R8*	*R9*
Samples	7	7	7	7	7	7	5	5	5
Passed reads	106 699	71 391	77 696	98 227	76 860	91 653	89 102	111 668	83 076
BRCA-mapped reads	106 303	70 953	77 339	97 778	76 559	91 421	88 699	110 724	82 718
(% of passed)	(99.6%)	(99.4%)	(99.54%)	(99.54%)	(99.6%)	(99.75%)	(99.5%)	(99.15%)	(99.5%)
Coverage, mean	81.8	50.9	62.7	81.6	115	138	216	269	202
[min, max]	[5,201]	[0,133]	[8,157]	[0,200]	[5,498]	[6,494]	[43,595]	[51,807]	[47,610]
Coverage SD	31.3	21.6	23.77	31.7	49.5	55.8	91.36	107.8	85.26
Coverage fold difference to mean ratio 90%/95%	1.98/2.77	1.95/2.39	1.86/2.34	1.91/2.23	1.81/2.11	1.82/2.43	1.68/2.09	1.49/1.74	1.69/2.04
No. of bases < 38	9430	28 947	15 238	5680	2895	3696	0	0	0
(% of mapped)	(8.2%)	(25.2%)	(13.3%)	(4.9%)	(1.78%)	(2.28%)			
No. of fragments < 38	106	318	178	74	10	14	0	0	0

which 223 were true positives (TP) and 4037 were false positives (FP). The high proportion (95%) of FPs identified by the NGS platform after alignment and raw variant calling means that filters are required. To discard false positives, six filters were assessed as follows (Table 2):

(1) Insertions and deletions covered by the BRCA HP assay. This filter is used to reduce the number of FP of insertions or deletions, caused by HP of 6 bp or longer (targetted by the assay), but also by HP of 5 bp (many of them covered by the BRCA HP assay PCRs). This filter discarded 1730 FP and 11 TP. All these 11 TP, plus one variant not detected by VIP (*BRCA1* c.1961delA, in a homopolymer of 8 As), were found by the HP kit, which demonstrated to be clear and completely reliable detecting length changes.

(2) Variants in regions with coverage below 38 × were considered undercovered and thus Sanger sequenced. This coverage threshold was based on De Leeneer's calculations, according to which this number of reads would allow to find an heterozygous variant for a minimum frequency of 25% with a power of 99.9%. This sensitivity is equivalent to a Phred score of 30.[17] This filter discarded 97 FP and 10

Table 2 Cumulative application of filters

		1[a]*: Ins/del BRCA HP*		*1→2*[b]*: Cov<38*		*(1+2)→3*[c]*: VAF<0.25*		*(1+2+3)→4*[d]*: Fcov=0 or Rcov=0*		*(1+2+3)→5*[e]*: FQ<30 & RQ<30*		*(1+2+3)→6*[f]*: Total Q <30*	
	Before filters	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*
Training Set													
FP	4037	2307	1730	2210	97	512	1698	9	503	228	284	227	285
TP	223	212	11	202	10	202	0	200	2	201	1	200	2
Sensitivity		0.951		0.953		1.000		0.990		0.995		0.990	
Specificity		0.429		0.042		0.769		0.982		0.555		0.557	
Validation Set													
FP	1471	872	599	872	0	168	704	3	165	59	109	59	109
TP	123	122	1	122	0	122	0	121	1	122	0	122	0
Sensitivity		0.992		1.000		1.000		0.992		1.000		1.000	
Specificity		0.407		0.000		0.807		0.982		0.649		0.649	

Variants retained (In) and discarded (Out) by the application of:
[a]filter 1: insertion or deletion covered by the BRCA HP assay.
[b]filter 2: coverage below 38, to variants retained by filter 1.
[c]filter 3: variant allele frequency below 0.25, to variants retained by filters 1 and 2.
[d]filter 4: variant forward coverage or variant reverse coverage equal to 0, to variants retained by filters 1, 2, and 3.
[e]filter 5: variant forward quality and variant reverse quality below 30, to variants retained by filters 1, 2, and 3.
[f]filter 6: total variant quality below 30, to variants retained by filters 1, 2, and 3.

TP in the Training Set, all of them were confirmed by the subsequent Sanger sequencing.

(3) Variants with an allele frequency <25% were disregarded. This filter discarded 1698 additional FP for the Training Set but not any TP.

(4) Variants detected in only one strand. This filter, indicated by VIP as the variant having forward coverage or reverse coverage equal to 0, discarded 503 FP and 2 TP (additionally to filters 1+2+3).

(5) Variants with forward and reverse variant mean qualities below 30.[12] This filter discarded 284 FP and 1 TP (additionally to filters 1+2+3).

(6) Variants with total quality below 30. This filter was very similar to filter 5 but differed in some variants, so it was tested to compare with filters 4 and 5. It discarded 285 FP and 2 TP (additionally to filters 1+2+3).

We observed that the application of the first three filters did not lead to the loss of any true mutation. These filters also lowered the number of FP from 4037 to 512 (Supplementary Figure 3). Filters 4–6 (variants detected in only one strand; variants with variant mean quality in forward and reverse below 30; variants with total quality below 30) resulted in the loss of 1 or 2 TP out of 28 samples, which is not acceptable in a *BRCA* diagnostic setting. If these filters were not used, Sanger sequencing of 512 FP and the 29 TP (23 pathogenic and 6 unknown significance variants, see Supplementary Table 1) would be needed to provide robust results, considerably increasing the cost and time of the workflow. Consequently, we opted for an intermediate strategy that consisted in using filter 4 (variants detected in only one strand) to generate a list of variants for which visual inspection of the aligned region was required. Filter 4 was chosen because it filtered most of the remaining FP (Table 2). Supplementary Figure 4 uses Venn diagrams to show the common and different FP and TP that filters 4, 5 and 6 would discard. Visualization was performed using the Amplicon Variant Analysis (AVA, Roche) software, permiting to discard artifactual variants present only in one strand, while keeping real variants that were wrongly aligned in different positions in both strands. This manual analysis discarded 501 FP and 0 TP, leaving 2FP and 2TP for Sanger sequence analysis (Supplementary Figure 3). Analysis of the HP assay detected all of the insertions and deletions that fall between its primers. Sanger sequencing confirmed that all FPs were pyrosequencing errors.

To summarize, in the Training Set we expected to find 227 heterozygous variants. Considering only the variant calling results from GS Junior with the application of 3 filters, we found 202 TP (none of which were discarded by the blind visual inspection); the HP assay detected 12 more, and Sanger sequencing of low-coverage regions identified the remaining 13 TP variants. As expected, FPs decreased with the correlative application of filters and visualization in our workflow design. Only 11 FP required Sanger sequencing to be discarded. These numbers would correspond to an experimental sensitivity and specificity for point mutations of 100% at the last step of our workflow (Table 3). Consequently, complete analysis of the Training Set enabled us to generate a new NGS-based workflow for genetic testing of *BRCA* genes (Figure 2).

Variants in homopolymer sequences

Pyrosequencing of homopolymers presented a technical limitation, as it was difficult to distinguish FP from TP deletions in homopolymer stretches of 6 bp or longer. Therefore, an HP assay is needed. Examples of homopolymer difficulties are shown in Supplementary Figure 5. Some variants in HP of 6 bp or longer are also detected by VIP but the BRCA HP assay is more reliable.

Validation Set

To validate the usefulness and readiness of the pipeline, 14 consecutive samples received for diagnosis of HBOCS were simultaneously analyzed by separate teams using NGS and our current workflow. A fifteenth sample, which bears a pathogenic *BRCA1* mutation as well as a duplication of exons 9–24 of *BRCA1*, was added to test whether copy-number variation could be detected at this coverage. The library for this Validation Set was created using a new version of the BRCA MASTR kit (v2.1), in which the problem of coverage of *BRCA1* exon 7 was solved. To increase coverage, the 15 samples were sequenced in 3 GS Junior runs (R7–R9), 5 samples per run.

The average mean base coverage was 229 ± 95. The average fold difference to mean ratio was 1.62 at the 10th percentile and 1.96 at the

Table 3 Variant calling results

	Training Set				*Validation Set*			
	GS Junior[a]	*+ Visual review*	*+ HP Kit*	*+ Sanger*	*GS Junior*[a]	*+ Visual review*	*+ HP Kit*	*+ Sanger*
True +	202	202	214	227	122	122	123	123
True −	613 161	613 662	613 662	613 673	347 619	347 783	347 783	347 787
False +	512	11	11	0	168	4	4	0
False −	12	12	0	0	1	1	0	0
Variants in low coverage	13	13	13	0	0	0	0	0
Sensitivity	0.88987	0.88987	0.94273	1.00000	0.99187	0.99187	1.00000	1.00000
Specificity	0.99917	0.99998	0.99998	1.00000	0.99952	0.99999	0.99999	1.00000

[a]After applying filters 1 + 2 + 3.

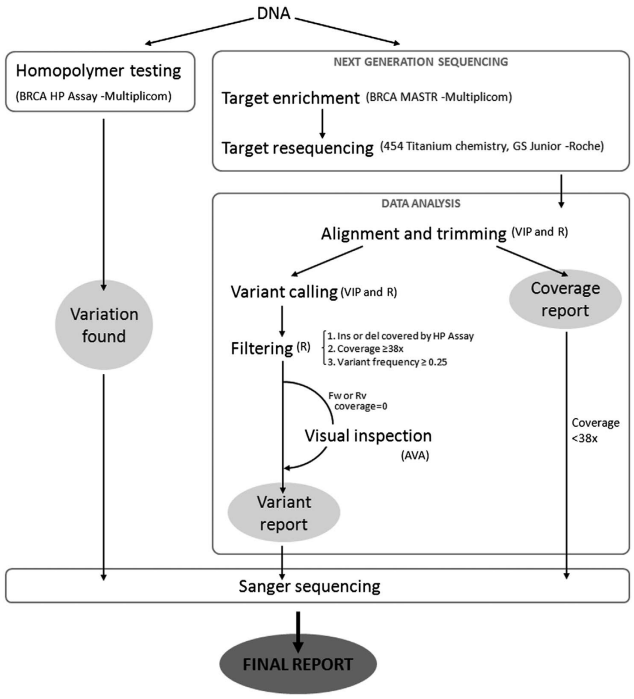


Figure 2 Proposed workflow for analyzing *BRCA1* and *BRCA2* using NGS. A screening using the BRCA HP kit (Multiplicom) allows detection of insertions or deletions located in homopolymers of 6 bp or longer and their surroundings. Sanger sequencing confirms any aberrant pattern found. Simultaneously, DNA samples are analyzed by NGS. *BRCA1* and *BRCA2* coding regions and their intron–exon boundaries are amplified using the BRCA MASTR kit (Multiplicom), adding specific identifiers (MIDs) for each sample to pool them. Sequencing of the enriched regions from pooled samples is performed by using 454 Titanium chemistry in a GS Junior platform (Roche). Data generated by the sequencer are analyzed using the public software VIP and R instructions, which allows us to align all of the sequences generated, trim the surrounding regions of each amplicon (adapters, MIDs and primers) and call putative variants. After filtering the initial variants with filters 1, 2, 3 and 4, a subset (variants with null forward or reverse coverage) is selected for visual inspection of their alignment with AVA, which will discard obvious FPs. All remaining variants are confirmed by Sanger sequencing. As our aim was to integrate this approach into the diagnostic routine, this revision was performed independently by two qualified technicians to generate a common list indicating the decision for any variant under analysis. If a discrepancy arose between the two referees, the most conservative decision was adopted. Regions with low coverage ($<38\times$) are also Sanger sequenced.

5th percentile (Table 1). No bases with coverage under $38\times$ were observed, meaning that Sanger resequencing was unnecessary for low coverage. For example, in experiment R7, all amplicons produced coverage over $50\times$ except amplicon *BRCA1*_ex20.1 in MID1 (Supplementary Figure 2B).

Our analysis algorithm detected 123 heterozygous variants in this set of samples (2 of which were pathogenic). In all, 122 TP (none of which were discarded by the blind visual inspection) were identified by NGS plus filtering, and the remaining TP were detected by the BRCA HP assay. The first three filters reduced FP from 1471 to 168. After the visual alignment review, four FP remained, which were adequately classified after Sanger sequencing. Also for the Validation Set, an experimental sensitivity and an experimental specificity of 100% were achieved by the workflow (Table 3). However, as explained thoroughly in Mattocks *et al*,[18] when the measured sensitivity in the validation of a qualitative test is 100%, a good estimation of the 95% confidence interval should be calculated by the rule of three. As our sample size consists in 123 mutations tested in the Validation Set, our statistical power corresponds to a confidence interval $\geq 97.5\%$.

Large rearrangements detection

A large genomic duplication comprising exons 9–24 of *BRCA1*[19] was included in the Validation Set in run R9. A total of 27 out of 30 amplicons involved in the duplication yielded a dosage quotient value above 1.35, similar to the MLPA results. In addition, the borders of the duplication were quite well defined. To explore the limitations of this analysis in greater depth, we decided to add seven previously identified LGRs showing different deletions and duplications.[19,20] These samples were analyzed in subsequent runs mixed with samples without LGRs. In summary, all LGRs were detected (Figure 3 and Supplementary Figure 6B), duplications showed normalized amplicon values above 1.3 and deletions showed values below 0.7. However, many other amplicons showed values outside these limits (0.7–1.3) representing FPs, which were identified both in control samples (Supplementary Figure 6A) and in other regions of samples showing LGRs. In addition, when very large rearrangements were present in one gene, amplicons from the other gene were affected in the opposite direction due to a bias produced in the normalization process, making it difficult to discriminate real deletions/duplications from FP amplicons.

Cost efficiency

A study of all the consumables and time used, from DNA extraction to obtain the final report, was performed with the aim of comparing

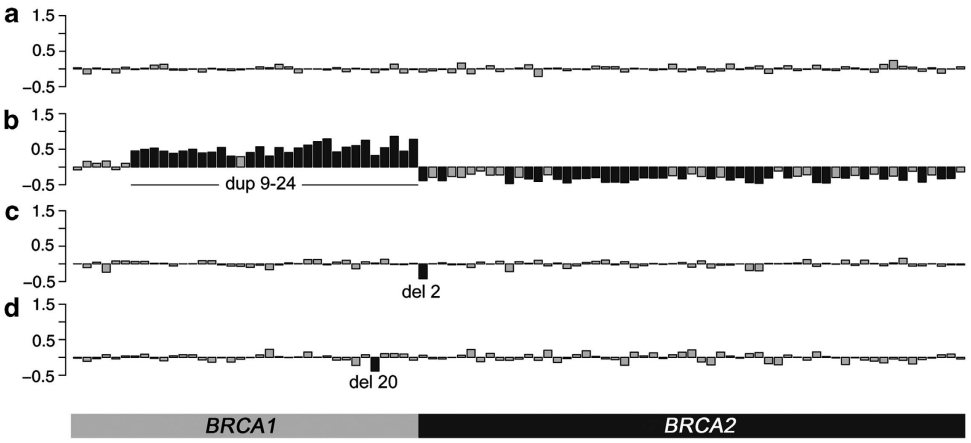


Figure 3 Detection of LGRs using NGS results. Bar plots of the dose of NGS amplicons after normalization. X-axis: NGS amplicons. Y-axis: Count ratio minus 1. Fragments with normalized ratios above 1.3 and below 0.7 are highlighted in black, indicating putative duplications and deletions, respectively. (**a**) Control sample with no alterations. (**b**) Sample with a duplication of the region comprising *BRCA1* exons 9–24. (**c**) Sample with a deletion comprising *BRCA2* exon 2. (**d**) Sample with a deletion comprising exon 20 of *BRCA1*.

our former genetic testing strategy with the new strategy. We found that the overall price of consumables was similar for both approaches (conformation-sensitive capillary electrophoresis + Sanger sequencing *vs* NGS + HP assay + Sanger sequencing), with an estimated cost of €325 in each case. However, the hands-on time and turnaround time were substantially different. By using our proposed NGS workflow, we save 57% of the time cost per technician (down from 14 h/sample to 6 h/sample) and obtain a reduction of ~25% in turnaround time (down from 20 days for 13 samples to 15 days for 14 samples).

DISCUSSION

Here we present a complete workflow for the analysis of the *BRCA1* and *BRCA2* genes, based on the use of a multiplex PCR strategy (Multiplicom) to generate the patient's DNA library followed by pyrosequencing using a benchtop NGS platform (GS Junior) and subsequent bioinformatic analysis based on a combination of three software (VIP, R, and AVA). The analysis of insertions and duplications in homopolymeric regions was performed by an HP assay (Multiplicom). Our results indicate that this workflow achieves an excellent performance for point mutations, with a specificity of 100% and a sensitivity ≥97.5% (95% CI) (Figure 2, Table 2).

Our approach improves previous studies using NGS for BRCA genetic testing in different aspects including: 1) the combination of a Training and a Validation Set, which is the best way to accurately assess the sensitivity of a given approach; 2) the development of a complete algorithm, incorporating the use of the BRCA HP kit, allows us to reach a sensitivity of 100% (≥97.5% with a 95% confidence interval), keeping with an excellent specificity (100%; ≥99.9991% with a 95% confidence interval); and 3) the cost-effective analysis for BRCA analysis in a benchtop NGS platform. Although it seems that improvements on analysis are still needed, the presented results open the door to the identification of large rearrangements, especially those affecting several exons.

The first step when using any NGS platform is to obtain the patient's DNA library for the region/s of interest. We selected a commercial multiplex PCR assay (Multiplicom) because it offers better reproducibility, more straightforward setup and better performance than in-house methods. This assay showed increased efficiency and homogeneity in the amplification of *BRCA* fragments with every new version of the kit released. A crucial step in preparing a DNA library for sequencing is to obtain equimolar proportions of all studied fragments to prevent undercovered regions and avoid the need for high mean coverage, which would generate higher costs. The latest version of the kit achieves an excellent ratio (1.96) between mean coverage and the 5th percentile of coverage (Table 1). This result outperforms the homogeneity previously reported by other groups describing next-generation *BRCA* testing using either long-range PCR,[4] primer-specific direct capture for single-molecule sequencing,[13] or in-house single/multiplex PCR.[12,14] It is also important to note that all of the MIDs used in the present study showed similar coverage results. Overall, this commercial assay allows the generation of a robust library for all the patients under study, maximizing the number of samples analyzed in a run.

Pyrosequencing performance with the GS Junior has been found to be similar to that of the GS FLX system,[12] which also uses Roche-454 technology. The GS Junior offers a more convenient scale for a mid-sized genetic testing laboratory, where the need to pool a large number of samples to use the whole capacity of a GS FLX device would increase waiting lists and, as a result, diagnostic turnaround times. GS Junior offers low entry and operating costs, providing conventional molecular diagnostics laboratories with a means of using NGS. Compared with other NGS technologies, Roche pyrosequencing currently offers the longest reads. This is advantageous for aligning possible mid-size insertions and deletions. In this study, the longest deletion tested (19 bp) was detected without a decrease in the expected allele frequency. The main disadvantage of pyrosequencing relative to other NGS technologies is the accuracy of length determination in homopolymers.[12,17,21] In pyrosequencing, the light-intensity signal observed in each cycle is proportional to the actual number of incorporated nucleotides, which is the base for homopolymer length calling. The accuracy of this method decreases with homopolymer length, which may eventually generate artefactual insertions and deletions in long homopolymers.[22,23] Our workflow circumvents this problem by using the BRCA HP assay.

To analyze the results we designed our own bioinformatic analysis pipeline using a combination of different software. VIP proved to find every variant, when enough coverage, but one deletion in a HP of 8 and has the advantage of being open source, making it preferable to other commercial software packages, which have only a limited capacity for adaptation to particular genes or laboratory needs. The generation of a reliable variant list is one of the most complex parts of the analysis and a key stage in the implementation of all next-generation platforms. The systematic application of a set of

evaluated filters is needed.[12] Ours is a four-filter approach: three run automatically and a fourth filter generates a list of variants that require visual examination or Sanger confirmation. Visual examination took about 3 h per run per revisor, and both revisions provided concordant results. Application of this four-filter approach left 16 fragments per patient requiring visual inspection, after which only 1% of them required Sanger confirmation. The fourth filter was able to remove a substantial proportion of the FPs without losing any TP when compared with other series.[12] The use of the commercial homopolymer kit was paramount for correctly reading sequences containing homopolymer stretches, which often require visual inspection and/or Sanger sequencing. Nevertheless, further development of tools for analysis of HP regions in NGS is needed to improve performance and to reduce the number of results requiring visual inspection.

In relation to the number of samples to be placed in each run, our results indicate that 5–7 is optimal with the new version of the kit. The latest version was experimentally tested using five samples and none of the fragments required resequencing for low coverage. We also carried out an *in silico* simulation of the same test with seven samples in each run instead of the five samples tested experimentally. The simulation was performed by randomly selecting 71% (five sevenths) of reads from each run and following the same analysis pipeline as for the Validation Set. The simulation results indicate that four fragments would have required Sanger sequencing due to low coverage (2 for R7, 0 for R8 and 2 for R9; that is, $\sim$0.2 fragments per sample), maintaining the same specificity and sensitivity as observed in the Validation Set (data not shown).

Although we have been able to detect LGRs, FPs have also been identified both in control and in patient samples, indicating that the specificity is too low for this method to be considered as an alternative strategy for detecting this type of mutations with the current software, kit protocol, and normalization procedures. Hopefully, in the near future, improvements to methodologies will lead to better specificity, allowing this approach to be used for the identification of LGRs in a diagnostic setting.

In a typical clinical setting, it is necessary to study a small number of genes comprehensively with the certainty of covering the whole coding region without any exception, with a sensitivity equal to or greater than that of conventional Sanger sequencing. Few studies have tackled a comprehensive assessment of specificity and sensitivity of NGS in the context of the requirements needed for a clinical diagnosis laboratory. To our knowledge, this is the first time that a NGS-based approach has been developed to perform comprehensive genetic testing of *BRCA* genes, including homopolymer regions, in a benchtop platform. We propose here a workflow that, using the GS Junior platform, allowed the identification of all DNA variants previously detected. A complete methodological process together with a detailed bioinformatic pipeline and validation of filters using open access programs has been critical to this achievement. Our custom-designed NGS workflow for genetic testing of *BRCA* genes meets the sensitivity and specificity requirements for the genetic diagnosis of HBOCS, making it feasible and cost-effective in comparison to current standards.

ACKNOWLEDGEMENTS

We thank Bernat Gel and Anna Ruiz for critical advice and corrections of the manuscript, and Toni Berenguer for statistical advice. We would also like to thank the Spanish Association Against Cancer (AECC) for recognizing our group with one of its awards. Finally, we would like to thank the teams from Multiplicom and Roche for their constant support. We thank contract grant sponsors: Spanish Health Research Fund; Carlos III Health Institute; Catalan Health Institute and Autonomous Government of Catalonia. Contract grant numbers: ISCIIIRETIC: RD06/0020/1051, RD06/0020/1050; 2009SGR290; PI10/01422; CA10/01474.

AUTHOR CONTRIBUTIONS

The project was conceived and the experiments and data analyses coordinated by LF, EC, CL, ES, GC. Samples were genetically characterized by JDV, MM, ET, EM, RC, CG, OC, MP, SG. Bioinformatic analysis was performed by ALD and VM. Samples from patients were obtained from JB and IB. The manuscript was written by LF, ALD, EC, JDV and CL and was discussed and improved by all the authors.

1 Rothberg JM, Hinz W, Rearick TM *et al*: An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011; **475**: 348–352.
2 Voelkerding KV, Dames SA, Durtschi JD: Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009; **55**: 641–658.
3 Goossens D, Moens LN, Nelis E *et al*: Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. *Hum Mutat* 2009; **30**: 472–476.
4 Morgan JE, Carr IM, Sheridan E *et al*: Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum Mutat* 2010; **31**: 484–491.
5 King MC, Marks JH, Mandell JB: Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* 2003; **302**: 643–646.
6 Bermejo-Perez MJ, Marquez-Calderon S, Llanos-Mendez A: Effectiveness of preventive interventions in BRCA1/2 gene mutation carriers: a systematic review. *Int J Cancer* 2007; **121**: 225–231.
7 De Leeneer K, Coene I, Poppe B, De Paepe A, Claes K: Rapid and sensitive detection of BRCA1/2 mutations in a diagnostic setting: comparison of two high-resolution melting platforms. *Clin Chem* 2008; **54**: 982–989.
8 Marsh DJ, Howell VM: The use of denaturing high performance liquid chromatography (DHPLC) for mutation scanning of hereditary cancer genes. *Methods Mol Biol* 2010; **653**: 133–145.
9 Mattocks CJ, Watkins G, Ward D *et al*: Interlaboratory diagnostic validation of conformation-sensitive capillary electrophoresis for mutation scanning. *Clin Chem* 2010; **56**: 593–602.
10 Ewald IP, Ribeiro PL, Palmero EI, Cossio SL, Giugliani R, Ashton-Prolla P: Genomic rearrangements in BRCA1 and BRCA2: a literature review. *Genet Mol Biol* 2009; **32**: 437–446.
11 Sluiter MD, van Rensburg EJ: Large genomic rearrangements of the BRCA1 and BRCA2 genes: review of the literature and report of a novel BRCA1 mutation. *Breast Cancer Res Treat* 2011; **125**: 325–349.
12 De Leeneer K, Hellemans J, De Schrijver J *et al*: Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. *Hum Mutat* 2011; **32**: 335–344.
13 Thompson JF, Reifenberger JG, Giladi E *et al*: Single-step capture and sequencing of natural DNA for detection of BRCA1 mutations. *Genome Res* 2011; **22**: 340–345.
14 Hernan I, Borras E, de Sousa Dias M *et al*: Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. *J Mol Diagn* 2012; **14**: 286–293.
15 De Schrijver JM, De Leeneer K, Lefever S *et al*: Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. *BMC Bioinformatics* 2010; **11**: 269.
16 Kent WJ: BLAT–the BLAST-like alignment tool. *Genome Res* 2002; **12**: 656–664.
17 De Leeneer K, De Schrijver J, Clement L *et al*: Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. *PLoS One* 2011; **6**: e25531.
18 Mattocks CJ, Morris MA, Matthijs G *et al*: A standardized framework for the validation and verification of clinical molecular genetic tests. *Eur J Hum Genet* 2010; **18**: 1276–1288.
19 del Valle J, Feliubadalo L, Nadal M *et al*: Identification and comprehensive characterization of large genomic rearrangements in the BRCA1 and BRCA2 genes. *Breast Cancer Res Treat* 2009; **122**: 733–743.
20 del Valle J, Campos O, Velasco A *et al*: Identification of a new complex rearrangement affecting exon 20 of BRCA1. *Breast Cancer Res Treat* 2011; **130**: 341–344.
21 Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**: R143.
22 Loman NJ, Misra RV, Dallman TJ *et al*: Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 562.
23 Quinlan AR, Stewart DA, Stromberg MP, Marth GT: Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 2008; **5**: 179–181.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Supplemental Data Table 1. List of variants analyzed

Gene	Mutation	Pathog.	RUN	MID	DetectionStep	RUN	MID	DetectionStep
BRCA1	c.68_69delAG	PAT	1	1	NGS	5	1	NGS
BRCA1	c.2082C>T	POL	1	1	NGS	5	1	NGS
BRCA1	c.2311T>C	POL	1	1	NGS	5	1	NGS
BRCA1	c.2612C>T	POL	1	1	NGS	5	1	NGS
BRCA1	c.3113A>G	POL	1	1	NGS	5	1	NGS
BRCA1	c.3548A>G	POL	1	1	NGS	5	1	NGS
BRCA1	c.4308T>C	POL	1	1	NGS	5	1	NGS
BRCA1	c.4837A>G	POL	1	1	NGS	5	1	NGS
BRCA2	c.1114A>C	POL	1	1	Sanger	5	1	NGS
BRCA2	c.3807T>C	POL	1	1	NGS	5	1	NGS
BRCA1	c.212+1G>A	PAT	1	2	NGS			
BRCA2	c.-26G>A	POL	1	2	NGS			
BRCA2	c.3396A>G	POL	1	2	NGS			
BRCA2	c.3807T>C	POL	1	2	NGS			
BRCA2	c.7242A>G	POL	1	2	NGS			
BRCA2	c.7806-14T>C	POL	1	2	NGS			
BRCA1	c.431dupA	PAT	1	3	NGS-Visual	5	3	Sanger
BRCA1	c.442-34T>C	POL	1	3	NGS	5	3	NGS
BRCA1	c.2082C>T	POL	1	3	NGS	5	3	NGS
BRCA1	c.2311T>C	POL	1	3	NGS	5	3	NGS
BRCA1	c.2612C>T	POL	1	3	NGS	5	3	NGS
BRCA1	c.3113A>G	POL	1	3	NGS	5	3	NGS
BRCA1	c.3548A>G	POL	1	3	NGS	5	3	NGS
BRCA1	c.4308T>C	POL	1	3	NGS	5	3	NGS
BRCA1	c.4837A>G	POL	1	3	NGS	5	3	NGS
BRCA1	c.5124G>A	USV	1	3	Sanger	5	3	NGS
BRCA2	c.1114A>C	POL	1	3	Sanger	5	3	NGS
BRCA1	c.2077G>A	POL	1	4	NGS	5	4	NGS
BRCA2	c.956dupA	PAT	1	4	HP Assay	5	4	HP Assay
BRCA1	c.2077G>A	POL	1	5	NGS	5	5	NGS
BRCA1	c.2082C>T	POL	1	5	NGS	5	5	NGS
BRCA1	c.2311T>C	POL	1	5	NGS	5	5	NGS
BRCA1	c.2612C>T	POL	1	5	NGS	5	5	NGS
BRCA1	c.3113A>G	POL	1	5	NGS	5	5	NGS
BRCA1	c.3548A>G	POL	1	5	NGS	5	5	NGS
BRCA1	c.4308T>C	POL	1	5	NGS	5	5	NGS
BRCA1	c.4837A>G	POL	1	5	NGS	5	5	NGS
BRCA2	c.2402_2420del	PAT	1	5	NGS	5	5	NGS
BRCA1	c.2082C>T	POL	1	6	NGS	5	6	NGS
BRCA1	c.2311T>C	POL	1	6	NGS	5	6	NGS
BRCA1	c.2612C>T	POL	1	6	NGS	5	6	NGS
BRCA1	c.3113A>G	POL	1	6	NGS	5	6	NGS
BRCA1	c.3548A>G	POL	1	6	NGS	5	6	NGS
BRCA1	c.4308T>C	POL	1	6	NGS	5	6	NGS
BRCA1	c.4837A>G	POL	1	6	NGS	5	6	NGS
BRCA2	c.6405_6409delCTTAA	PAT	1	6	HP Assay	5	6	HP Assay
BRCA1	c.442-34T>C	POL	1	7	Sanger			
BRCA1	c.2082C>T	POL	1	7	NGS			
BRCA1	c.2311T>C	POL	1	7	NGS			
BRCA1	c.2612C>T	POL	1	7	NGS			
BRCA1	c.3113A>G	POL	1	7	NGS			
BRCA1	c.3119G>A	POL	1	7	NGS			
BRCA1	c.3548A>G	POL	1	7	NGS			
BRCA1	c.4308T>C	POL	1	7	NGS			
BRCA1	c.4837A>G	POL	1	7	NGS			
BRCA2	c.-26G>A	POL	1	7	NGS			
BRCA2	c.1114A>C	POL	1	7	Sanger			
BRCA2	c.3396A>G	POL	1	7	NGS			
BRCA2	c.5744C>T	POL	1	7	NGS			
BRCA2	c.7242A>G	POL	1	7	NGS			
BRCA2	c.7806-14T>C	POL	1	7	NGS			
BRCA2	c.9026_9030delATCAT	PAT	1	7	HP Assay			
BRCA2	c.-26G>A	POL	2	1	NGS	6	9	NGS
BRCA2	c.1114A>C	POL	2	1	Sanger	6	9	NGS
BRCA2	c.3396A>G	POL	2	1	NGS-Visual	6	9	NGS
BRCA2	c.6275_6276delTT	PAT	2	1	HP Assay	6	9	HP Assay
BRCA2	c.7242A>G	POL	2	1	NGS	6	9	NGS
BRCA2	c.7806-14T>C	POL	2	1	NGS	6	9	NGS
BRCA2	c.9257-16T>C	POL	2	1	NGS	6	9	NGS
BRCA2	c.9976A>T	POL	2	1	Sanger	6	9	NGS
BRCA1	c.591C>T	POL	2	2	NGS	5	2	NGS
BRCA1	c.2082C>T	POL	2	2	NGS	5	2	NGS
BRCA1	c.2311T>C	POL	2	2	NGS	5	2	NGS
BRCA1	c.2612C>T	POL	2	2	NGS	5	2	NGS
BRCA1	c.3113A>G	POL	2	2	NGS	5	2	NGS
BRCA1	c.3548A>G	POL	2	2	NGS	5	2	NGS
BRCA1	c.4308T>C	POL	2	2	NGS	5	2	NGS
BRCA1	c.4837A>G	POL	2	2	NGS	5	2	NGS
BRCA1	c.5123C>A	PAT	2	2	Sanger	5	2	NGS

BRCA2	c.1114A>C	POL	2	2	Sanger	5	2	NGS
BRCA1	c.3770_3771delAG	PAT	2	4	Sanger			
BRCA1	c.5467+9C>A	USV	2	4	NGS			
BRCA2	c.1114A>C	POL	2	4	Sanger			
BRCA2	c.3807T>C	POL	2	4	NGS			
BRCA1	c.4107_4110dupATCT	PAT	2	5	NGS			
BRCA1	c.1961delA	PAT	2	6	HP Assay			
BRCA1	c.2077G>A	POL	2	6	NGS			
BRCA1	c.2612C>T	POL	2	7	NGS	5	7	NGS
BRCA2	c.-26G>A	POL	2	7	NGS	5	7	NGS
BRCA2	c.1114A>C	POL	2	7	Sanger	5	7	NGS
BRCA2	c.2803G>A	POL	2	7	NGS	5	7	NGS
BRCA2	c.3396A>G	POL	2	7	Sanger	5	7	NGS
BRCA2	c.5350_5351delAAinsT	PAT	2	7	HP Assay	5	7	HP Assay
BRCA2	c.7242A>G	POL	2	7	NGS	5	7	NGS
BRCA2	c.7806-14T>C	POL	2	7	NGS	5	7	NGS
BRCA1	c.442-34T>C	POL	2	8	Sanger			
BRCA1	c.1121_1123delCACinsT	PAT	2	8	HP Assay			
BRCA1	c.2077G>A	POL	2	8	NGS			
BRCA1	c.2082C>T	POL	2	8	NGS			
BRCA1	c.2311T>C	POL	2	8	NGS			
BRCA1	c.2612C>T	POL	2	8	NGS			
BRCA1	c.3113A>G	POL	2	8	NGS			
BRCA1	c.3548A>G	POL	2	8	NGS			
BRCA1	c.4308T>C	POL	2	8	NGS			
BRCA1	c.4837A>G	POL	2	8	NGS			
BRCA2	c.1114A>C	POL	2	8	Sanger			
BRCA2	c.1128delT	PAT	3	1	Sanger	6	10	NGS
BRCA2	c.3807T>C	POL	3	1	NGS	6	10	NGS
BRCA2	c.7806-14T>C	POL	3	1	NGS	6	10	NGS
BRCA2	c.10131_10133delAGA	USV	3	1	HP Assay	6	10	HP Assay
BRCA1	c.2082C>T	POL	3	2	NGS			
BRCA1	c.2311T>C	POL	3	2	NGS			
BRCA1	c.2612C>T	POL	3	2	NGS			
BRCA1	c.3113A>G	POL	3	2	NGS			
BRCA1	c.3548A>G	POL	3	2	Sanger			
BRCA1	c.4308T>C	POL	3	2	NGS			
BRCA1	c.4837A>G	POL	3	2	NGS			
BRCA2	c.2808_2811delACAA	PAT	3	2	HP Assay			
BRCA1	c.442-34T>C	POL	3	3	NGS			
BRCA1	c.1067A>G	POL	3	3	NGS			
BRCA2	c.-26G>A	POL	3	3	NGS			
BRCA2	c.3264dupT	PAT	3	3	NGS			
BRCA2	c.3396A>G	POL	3	3	NGS			
BRCA2	c.7242A>G	POL	3	3	NGS			
BRCA1	c.2082C>T	POL	3	4	NGS			
BRCA1	c.2311T>C	POL	3	4	NGS			
BRCA1	c.2397T>A	USV	3	4	NGS-Visual			
BRCA1	c.2612C>T	POL	3	4	NGS			
BRCA1	c.3113A>G	POL	3	4	NGS			
BRCA1	c.3548A>G	POL	3	4	NGS			
BRCA1	c.4308T>C	POL	3	4	NGS			
BRCA1	c.4837A>G	POL	3	4	NGS			
BRCA1	c.5144G>A	PAT	3	4	Sanger			
BRCA2	c.865A>C	POL	3	4	NGS			
BRCA2	c.1365A>G	POL	3	4	NGS			
BRCA2	c.2229T>C	POL	3	4	NGS-Visual			
BRCA2	c.2971A>G	POL	3	4	NGS			
BRCA1	c.442-34T>C	POL	3	5	NGS			
BRCA1	c.2082C>T	POL	3	5	NGS			
BRCA1	c.2311T>C	POL	3	5	NGS			
BRCA1	c.2612C>T	POL	3	5	NGS			
BRCA1	c.3113A>G	POL	3	5	NGS			
BRCA1	c.3548A>G	POL	3	5	NGS			
BRCA1	c.4226A>T	USV	3	5	NGS			
BRCA1	c.4308T>C	POL	3	5	NGS			
BRCA1	c.4837A>G	POL	3	5	NGS			
BRCA2	c.1114A>C	POL	3	5	Sanger			
BRCA2	c.5720_5723delCTCT	PAT	3	5	HP Assay			
BRCA2	c.7806-14T>C	POL	3	5	NGS			
BRCA1	c.1953_1956delGAAA	PAT	3	6	HP Assay	6	11	HP Assay
BRCA2	c.-26G>A	POL	3	6	NGS	6	11	NGS
BRCA2	c.865A>C	POL	3	6	NGS	6	11	NGS
BRCA2	c.1365A>G	POL	3	6	NGS	6	11	NGS
BRCA2	c.2229T>C	POL	3	6	Sanger	6	11	NGS
BRCA2	c.2971A>G	POL	3	6	NGS	6	11	NGS
BRCA2	c.3396A>G	POL	3	6	Sanger	6	11	NGS
BRCA2	c.7806-14T>C	POL	3	6	NGS	6	11	NGS
BRCA1	c.2082C>T	POL	3	7	NGS			
BRCA1	c.2311T>C	POL	3	7	NGS			
BRCA1	c.3113A>G	POL	3	7	NGS			
BRCA1	c.3418A>G	POL	3	7	NGS			

BRCA1	c.3548A>G	POL	3	7	NGS			
BRCA1	c.4308T>C	POL	3	7	NGS			
BRCA1	c.4837A>G	POL	3	7	NGS			
BRCA2	c.-26G>A	POL	3	7	NGS			
BRCA2	c.1114A>C	POL	3	7	Sanger			
BRCA2	c.3396A>G	POL	3	7	Sanger			
BRCA2	c.7242A>G	POL	3	7	NGS			
BRCA2	c.8946delA	PAT	3	7	HP Assay			
BRCA1	c.1571C>T	USV	4	1	NGS			
BRCA1	c.2082C>T	POL	4	1	NGS			
BRCA1	c.2311T>C	POL	4	1	NGS			
BRCA1	c.2612C>T	POL	4	1	NGS			
BRCA1	c.3113A>G	POL	4	1	NGS			
BRCA1	c.3257T>G	PAT	4	1	NGS			
BRCA1	c.3548A>G	POL	4	1	NGS			
BRCA1	c.4308T>C	POL	4	1	NGS			
BRCA1	c.4837A>G	POL	4	1	NGS			
BRCA2	c.3807T>C	POL	4	1	NGS			
BRCA2	c.7806-14T>C	POL	4	1	NGS			
BRCA2	c.3807T>C	POL	4	2	NGS			
BRCA2	c.7806-14T>C	POL	4	2	NGS			
BRCA2	c.7977-1G>A	PAT	4	2	NGS			
BRCA2	c.1114A>C	POL	4	3	Sanger			
BRCA2	c.3396A>G	POL	4	3	NGS			
BRCA2	c.7242A>G	POL	4	3	NGS			
BRCA2	c.7806-14T>C	POL	4	3	NGS			
BRCA1	c.2082C>T	POL	4	4	NGS	6	12	NGS
BRCA1	c.2311T>C	POL	4	4	NGS	6	12	NGS
BRCA1	c.2612C>T	POL	4	4	NGS	6	12	NGS
BRCA1	c.3113A>G	POL	4	4	NGS	6	12	NGS
BRCA1	c.3548A>G	POL	4	4	NGS	6	12	NGS
BRCA1	c.4308T>C	POL	4	4	NGS	6	12	NGS
BRCA1	c.4837A>G	POL	4	4	NGS	6	12	NGS
BRCA2	c.-26G>A	POL	4	4	NGS	6	12	NGS
BRCA2	c.7242A>G	POL	4	4	NGS	6	12	NGS
BRCA2	c.7806-14T>C	POL	4	4	NGS	6	12	NGS
BRCA1	c.442-34T>C	POL	4	5	NGS	6	13	NGS
BRCA1	c.1067A>G	POL	4	5	NGS	6	13	NGS
BRCA1	c.2077G>A	POL	4	5	NGS	6	13	NGS
BRCA1	c.2082C>T	POL	4	5	NGS	6	13	NGS
BRCA1	c.2311T>C	POL	4	5	NGS	6	13	NGS
BRCA1	c.2612C>T	POL	4	5	NGS	6	13	NGS
BRCA1	c.3113A>G	POL	4	5	NGS	6	13	NGS
BRCA1	c.3548A>G	POL	4	5	NGS	6	13	NGS
BRCA1	c.4308T>C	POL	4	5	NGS	6	13	NGS
BRCA1	c.4837A>G	POL	4	5	NGS	6	13	NGS
BRCA2	c.1114A>C	POL	4	5	NGS	6	13	NGS
BRCA2	c.3807T>C	POL	4	5	NGS	6	13	NGS
BRCA1	c.442-34T>C	POL	4	6	NGS	6	14	NGS
BRCA1	c.3119G>A	POL	4	6	NGS	6	14	NGS
BRCA2	c.3807T>C	POL	4	6	NGS	6	14	NGS
BRCA2	c.7806-14T>C	POL	4	6	NGS	6	14	NGS
BRCA1	c.2077G>A	POL	4	7	NGS	6	15	NGS
BRCA1	c.2082C>T	POL	4	7	NGS	6	15	NGS
BRCA1	c.2311T>C	POL	4	7	NGS	6	15	NGS
BRCA1	c.2612C>T	POL	4	7	NGS	6	15	NGS
BRCA1	c.3113A>G	POL	4	7	NGS	6	15	NGS
BRCA1	c.3548A>G	POL	4	7	NGS	6	15	NGS
BRCA1	c.4308T>C	POL	4	7	NGS	6	15	NGS
BRCA1	c.4837A>G	POL	4	7	NGS	6	15	NGS
BRCA2	c.-26G>A	POL	4	7	NGS	6	15	NGS
BRCA2	c.3396A>G	POL	4	7	NGS	6	15	NGS
BRCA2	c.3807T>C	POL	4	7	NGS	6	15	NGS
BRCA2	c.5744C>T	POL	4	7	NGS	6	15	NGS
BRCA2	c.7242A>G	POL	4	7	NGS	6	15	NGS
BRCA2	c.7806-14T>C	POL	4	7	NGS	6	15	NGS
BRCA1	c.442-34C>T	POL	7	1	NGS			
BRCA1	c.2082C>T	POL	7	1	NGS			
BRCA1	c.2311T>C	POL	7	1	NGS			
BRCA1	c.2612C>T	POL	7	1	NGS			
BRCA1	c.3113A>G	POL	7	1	NGS			
BRCA1	c.3548A>G	POL	7	1	NGS			
BRCA1	c.4308T>C	POL	7	1	NGS			
BRCA1	c.4837A>G	POL	7	1	NGS			
BRCA2	c.-26G>A	POL	7	1	NGS			
BRCA2	c.7242A>G	POL	7	1	NGS			
BRCA2	c.7806-14T>C	POL	7	1	NGS			
BRCA1	c.442-34C>T	POL	7	2	NGS			
BRCA1	c.1067A>G	POL	7	2	NGS			
BRCA2	c.1114A>C	POL	7	2	NGS			
BRCA2	c.3807T>C	POL	7	2	NGS			
BRCA2	c.9257-16T>C	POL	7	2	NGS			

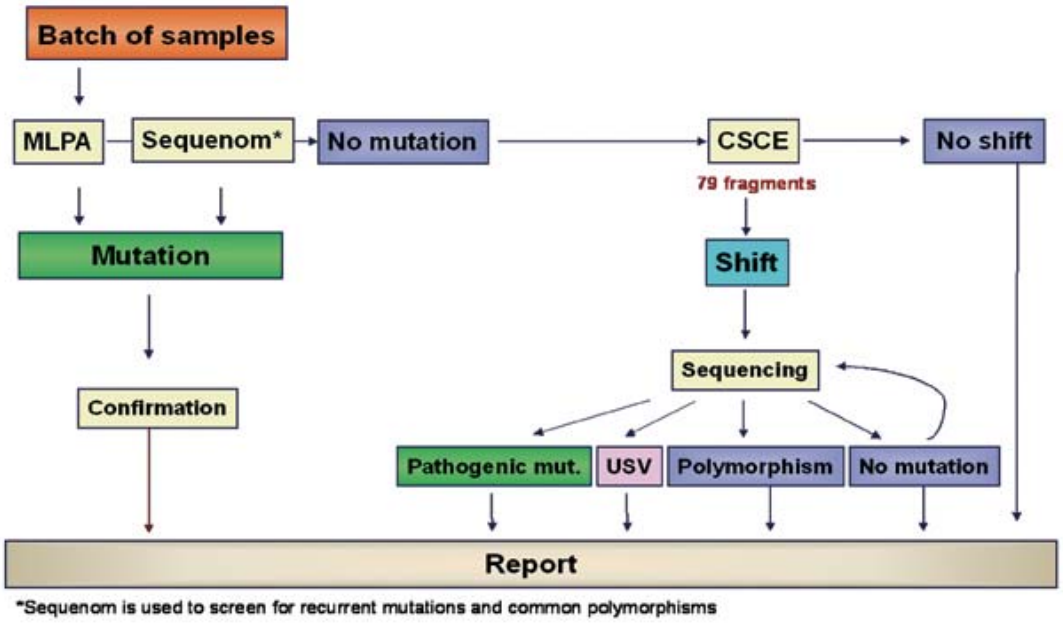
BRCA2	c.9976A>T	POL	7	2	NGS
BRCA1	c.2082C>T	POL	7	3	NGS
BRCA1	c.2311T>C	POL	7	3	NGS
BRCA1	c.2612C>T	POL	7	3	NGS
BRCA1	c.3113A>G	POL	7	3	NGS
BRCA1	c.3548A>G	POL	7	3	NGS
BRCA1	c.4308T>C	POL	7	3	NGS
BRCA1	c.4837A>G	POL	7	3	NGS
BRCA2	c.1114A>C	POL	7	3	NGS
BRCA2	c.2803G>A	POL	7	3	NGS
BRCA2	c.3807T>C	POL	7	3	NGS
BRCA1	c.2077G>A	POL	7	4	NGS
BRCA2	c.-26G>A	POL	7	4	NGS
BRCA2	c.68-7T>A	USV	7	4	NGS-Visual
BRCA1	c.2082C>T	POL	7	5	NGS
BRCA1	c.2311T>C	POL	7	5	NGS
BRCA1	c.2612C>T	POL	7	5	NGS
BRCA1	c.3113A>G	POL	7	5	NGS
BRCA1	c.3548A>G	POL	7	5	NGS
BRCA1	c.4308T>C	POL	7	5	NGS
BRCA1	c.4837A>G	POL	7	5	NGS
BRCA2	c.-26G>A	POL	7	5	NGS
BRCA2	c.1938C>T	POL	7	5	NGS
BRCA2	c.3396A>G	POL	7	5	NGS
BRCA2	c.7242A>G	POL	7	5	NGS
BRCA1	c.2082C>T	POL	8	1	NGS
BRCA1	c.2311T>C	POL	8	1	NGS
BRCA1	c.2612C>T	POL	8	1	NGS
BRCA1	c.3113A>G	POL	8	1	NGS
BRCA1	c.3548A>G	POL	8	1	NGS
BRCA1	c.4308T>C	POL	8	1	NGS
BRCA1	c.4837A>G	POL	8	1	NGS
BRCA2	c.865A>C	POL	8	1	NGS
BRCA2	c.1114A>C	POL	8	1	NGS
BRCA2	c.1365A>G	POL	8	1	NGS
BRCA2	c.2229T>C	POL	8	1	NGS
BRCA2	c.2971A>G	POL	8	1	NGS
BRCA2	c.3847_3848delGT	PAT	8	1	HP Assay
BRCA1	c.442-34C>T	POL	8	2	NGS
BRCA2	c.-26G>A	POL	8	2	NGS
BRCA2	c.3396A>G	POL	8	2	NGS
BRCA2	c.3807T>C	POL	8	2	NGS
BRCA2	c.5744C>T	POL	8	2	NGS
BRCA2	c.7242A>G	POL	8	2	NGS
BRCA2	c.7806-14T>C	POL	8	2	NGS
BRCA1	c.442-34C>T	POL	8	6	NGS
BRCA1	c.1067A>G	POL	8	6	NGS
BRCA1	c.2082C>T	POL	8	6	NGS
BRCA1	c.2311T>C	POL	8	6	NGS
BRCA1	c.2612C>T	POL	8	6	NGS
BRCA1	c.3113A>G	POL	8	6	NGS
BRCA1	c.3548A>G	POL	8	6	NGS
BRCA1	c.4308T>C	POL	8	6	NGS
BRCA1	c.4837A>G	POL	8	6	NGS
BRCA2	c.1114A>C	POL	8	6	NGS
BRCA2	c.7806-14T>C	POL	8	6	NGS
BRCA1	c.442-34C>T	POL	8	7	NGS
BRCA1	c.1067A>G	POL	8	7	NGS
BRCA1	c.2082C>T	POL	8	7	NGS
BRCA1	c.2311T>C	POL	8	7	NGS
BRCA1	c.2612C>T	POL	8	7	NGS
BRCA1	c.3113A>G	POL	8	7	NGS
BRCA1	c.3548A>G	POL	8	7	NGS
BRCA1	c.4308T>C	POL	8	7	NGS
BRCA1	c.4837A>G	POL	8	7	NGS
BRCA2	c.865A>C	POL	8	7	NGS
BRCA2	c.1114A>C	POL	8	7	NGS
BRCA2	c.1365A>G	POL	8	7	NGS
BRCA2	c.2229T>C	POL	8	7	NGS
BRCA2	c.2971A>G	POL	8	7	NGS
BRCA2	c.7806-14T>C	POL	8	7	NGS
BRCA2	c.8851G>A	POL	8	7	NGS
BRCA2	c.7806-14T>C	POL	8	8	NGS
BRCA1	c.2082C>T	POL	9	3	NGS
BRCA1	c.2311T>C	POL	9	3	NGS
BRCA1	c.2612C>T	POL	9	3	NGS
BRCA1	c.3113A>G	POL	9	3	NGS
BRCA1	c.3548A>G	POL	9	3	NGS
BRCA1	c.4308T>C	POL	9	3	NGS
BRCA1	c.4837A>G	POL	9	3	NGS
BRCA2	c.865A>C	POL	9	3	NGS
BRCA2	c.1365A>G	POL	9	3	NGS

BRCA2	c.2229T>C	POL	9	3	NGS
BRCA2	c.2971A>G	POL	9	3	NGS
BRCA2	c.-26G>A	POL	9	4	NGS
BRCA2	c.3396A>G	POL	9	4	NGS
BRCA2	c.3807T>C	POL	9	4	NGS
BRCA2	c.-26G>A	POL	9	5	NGS
BRCA2	c.3396A>G	POL	9	5	NGS
BRCA2	c.7242A>G	POL	9	5	NGS
BRCA1	c.2082C>T	POL	9	6	NGS
BRCA1	c.2311T>C	POL	9	6	NGS
BRCA1	c.2612C>T	POL	9	6	NGS
BRCA1	c.3113A>G	POL	9	6	NGS
BRCA1	c.3548A>G	POL	9	6	NGS
BRCA1	c.4308T>C	POL	9	6	NGS
BRCA1	c.4837A>G	POL	9	6	NGS
BRCA2	c.1114A>C	POL	9	6	NGS
BRCA2	c.3396A>G	POL	9	6	NGS
BRCA2	c.7242A>G	POL	9	6	NGS
BRCA2	c.7806-14T>C	POL	9	6	NGS
BRCA2	c.9257-16T>C	POL	9	6	NGS
BRCA2	c.9976A>T	POL	9	6	NGS
BRCA1	c.68_69delAG	PAT	9	7	NGS
BRCA2	c.-26G>A	POL	9	7	NGS
BRCA2	c.1114A>C	POL	9	7	NGS
BRCA2	c.3396A>G	POL	9	7	NGS

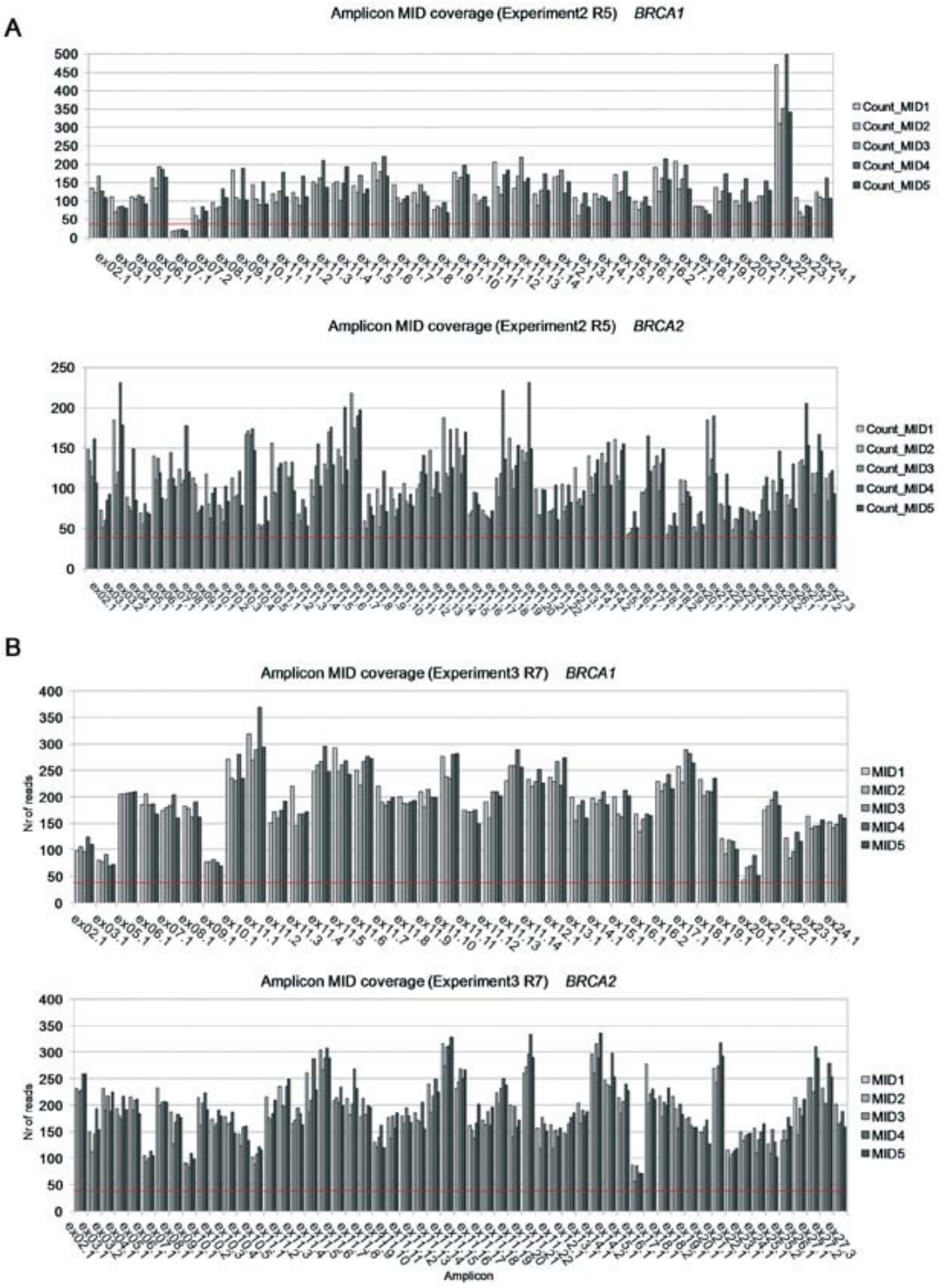
Comprehensive list of all variants analyzed in the present study. Variant names are based on HGVS nomenclature v2.0. Pathogenicity of variants is described according to our mutation database (PAT: pathogenic; USV: unknown significance variant; POL: polymorphism). Run number refers to the run in which the sample was first run, the MID assigned to the sample in this run, the step at which the variant was detected in this run (NGS: after analyzing the NGS run and applying the 4 filters of our pipeline; NGS-Visual: after analyzing the NGS run, applying the 4 filters of our pipeline and passing the visual inspection and Sanger sequencing; HP Assay: in the BRCA HP v2.0 assay (note that although 13 variants appear as found by the HP assay, all but one, have been found by VIP and filtered out by filter 1); Sanger: by Sanger Sequencing of low coverage regions (most of them were identified by VIP but discarded by filter 2). When a sample was run twice, the last 3 columns were repeated, with the information corresponding to the second run, MID and step at which the variant was found for the second time.

Supplementary Figure 1. Workflow used to screen for *BRCA1* and *BRCA2* mutations in our Molecular Diagnostics Unit

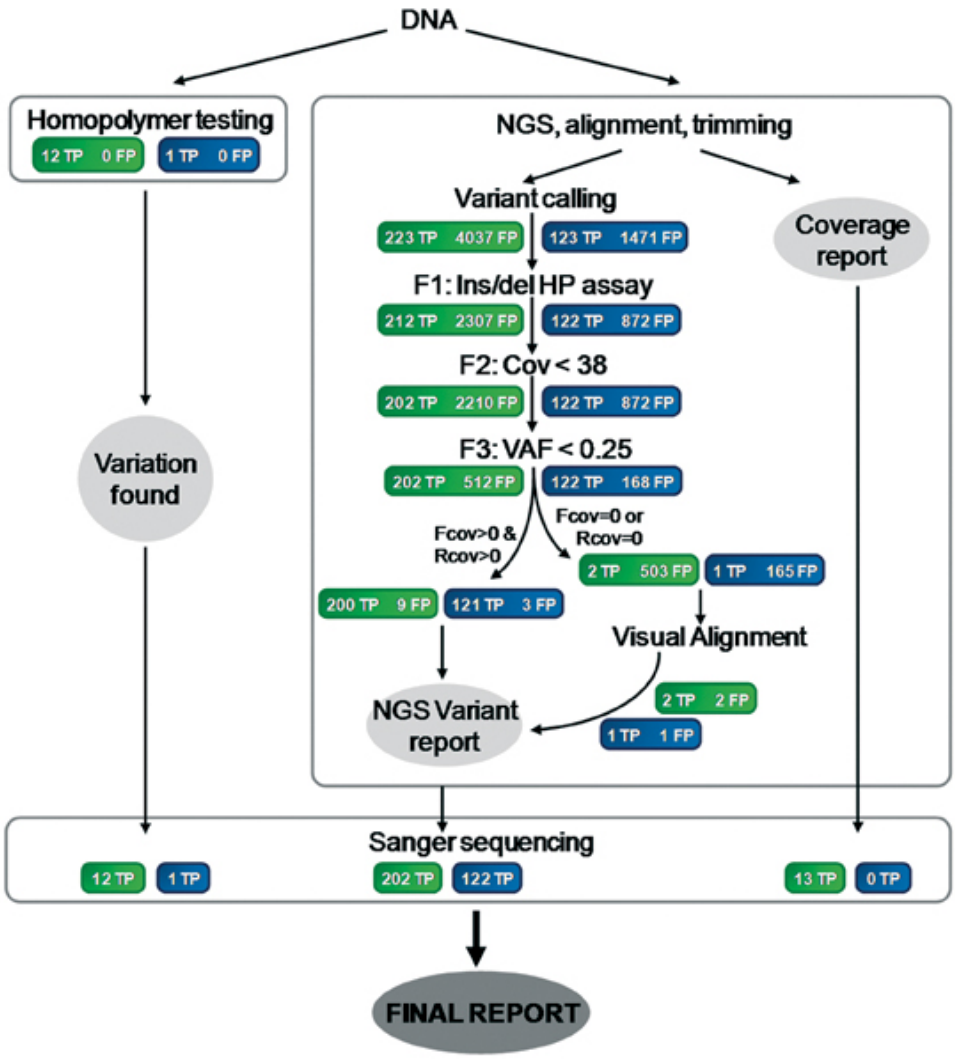
A cascade workflow for mutational analysis of *BRCA1* and *BRCA2* is used. Briefly, MLPA (Multiplex ligation-dependent probe amplification, MRC-Holland) is performed to detect large rearrangements followed by analysis of recurrent mutations using an in-house designed Sequenom assay (data not shown). If negative, Conformation Sensitive Capillary Electrophoresis (CSCE) analysis and sequencing of aberrant patterns is performed. CSCE is a method based on heteroduplex analysis and has shown sensitivity for heterozygous variants comparable to that of Sanger sequencing, with a lower cost. The limitation of this technique is that only heterozygous DNA changes are found, although this is not a drawback when searching for DNA mutations responsible for autosomal dominant syndromes. Hence, for the present study only variants in heterozygosity covered by both CSCE and Multiplicom kits were considered. However, Sequenom genotyping of the common polymorphisms included in our previous pipeline detected 99 homozygous polymorphisms, all of which were confirmed in the new NGS workflow.



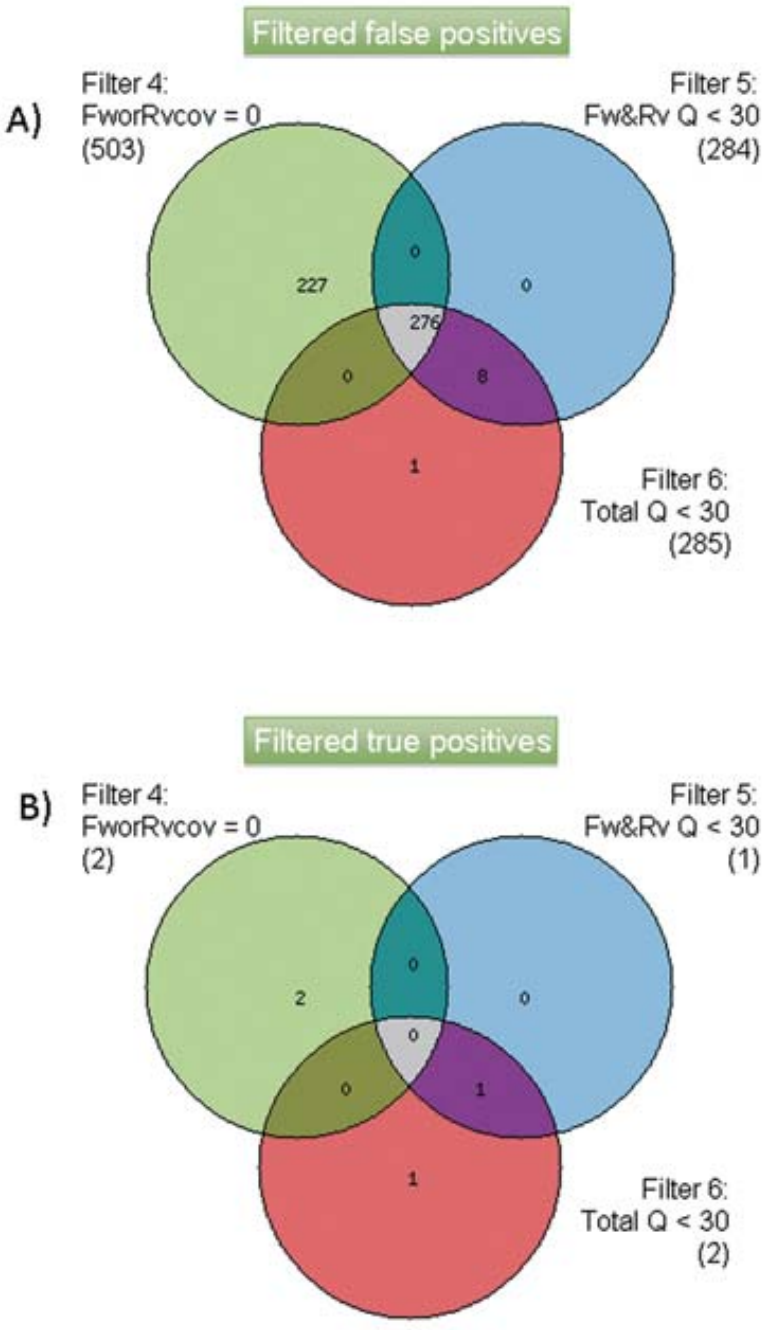
Supplementary Figure 2. Coverage distribution
Coverage distribution for each amplicon of *BRCA1* and *BRCA2* in 5 MIDs of Runs R5 and R7, from *Experiments 2* (A) and *3* (B), respectively. The red line indicates the minimum coverage threshold of 38x.



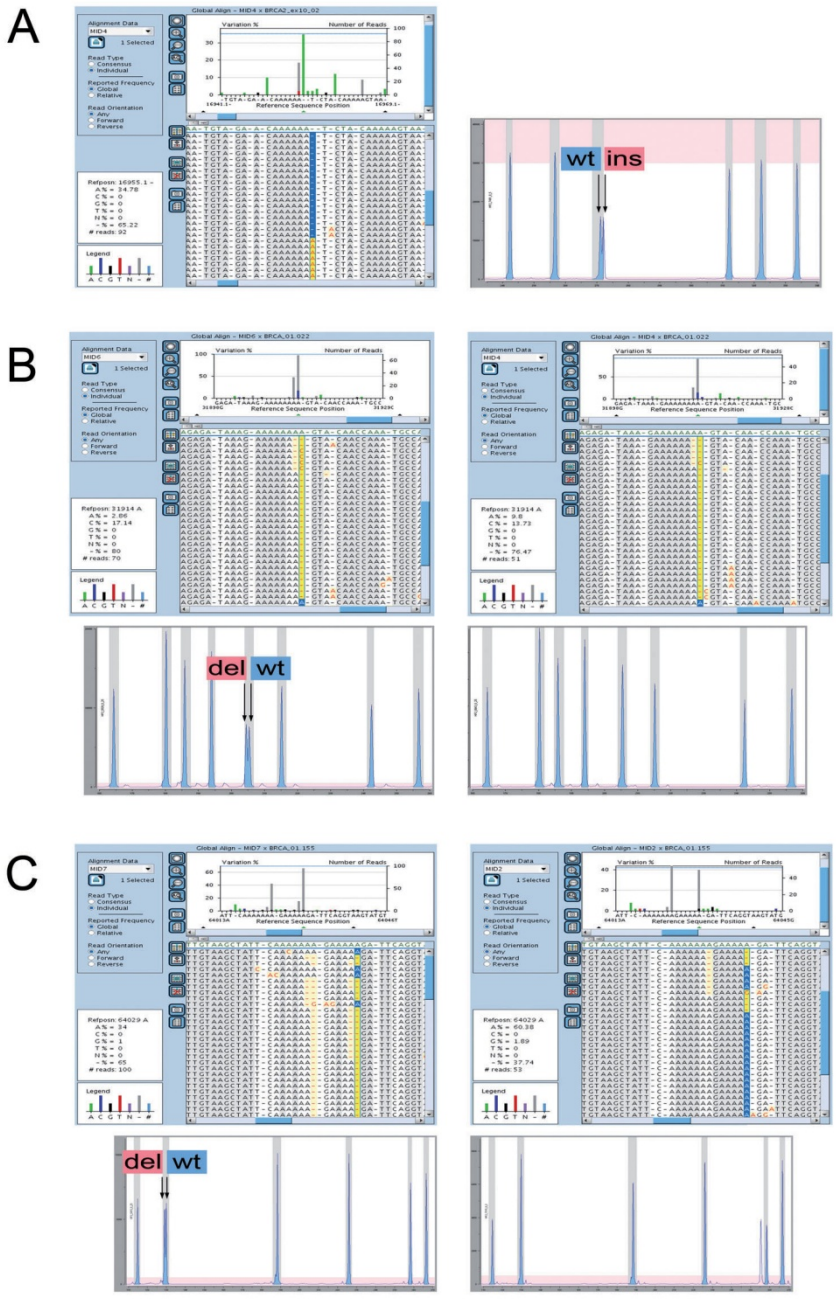
Supplementary Figure 3. NGS workflow showing true and false positives resulting from each step
Proposed workflow for *BRCA1* and *BRCA2*analysis, indicating the number of true positives (TP) and false positives (FP) resulting from each technique, filtering or inspection step. Results from the *Training Set* are marked with green labels, results from the *Validation Set* are marked with blue labels. True positives filtered out by filters 1 and 2 are recovered by the homopolymer testing and Sanger sequencing, respectively, of regions indicated by the coverage report. The Sanger sequencing load is decreased thanks to the bypass of visual classification of variants detected in only one strand, indicated by VIP as having forward coverage or reverse coverage of zero (Fcov=0 or Rcov=0). Thanks to this workflow, all variants detected in our previous workflow (227 in the *Training Set* and 123 *Validation Set*) were identified by the NGS workflow and only 11 FP in the *Training Set* and four FP in the *Validation Set* needed Sanger sequencing to be discarded.



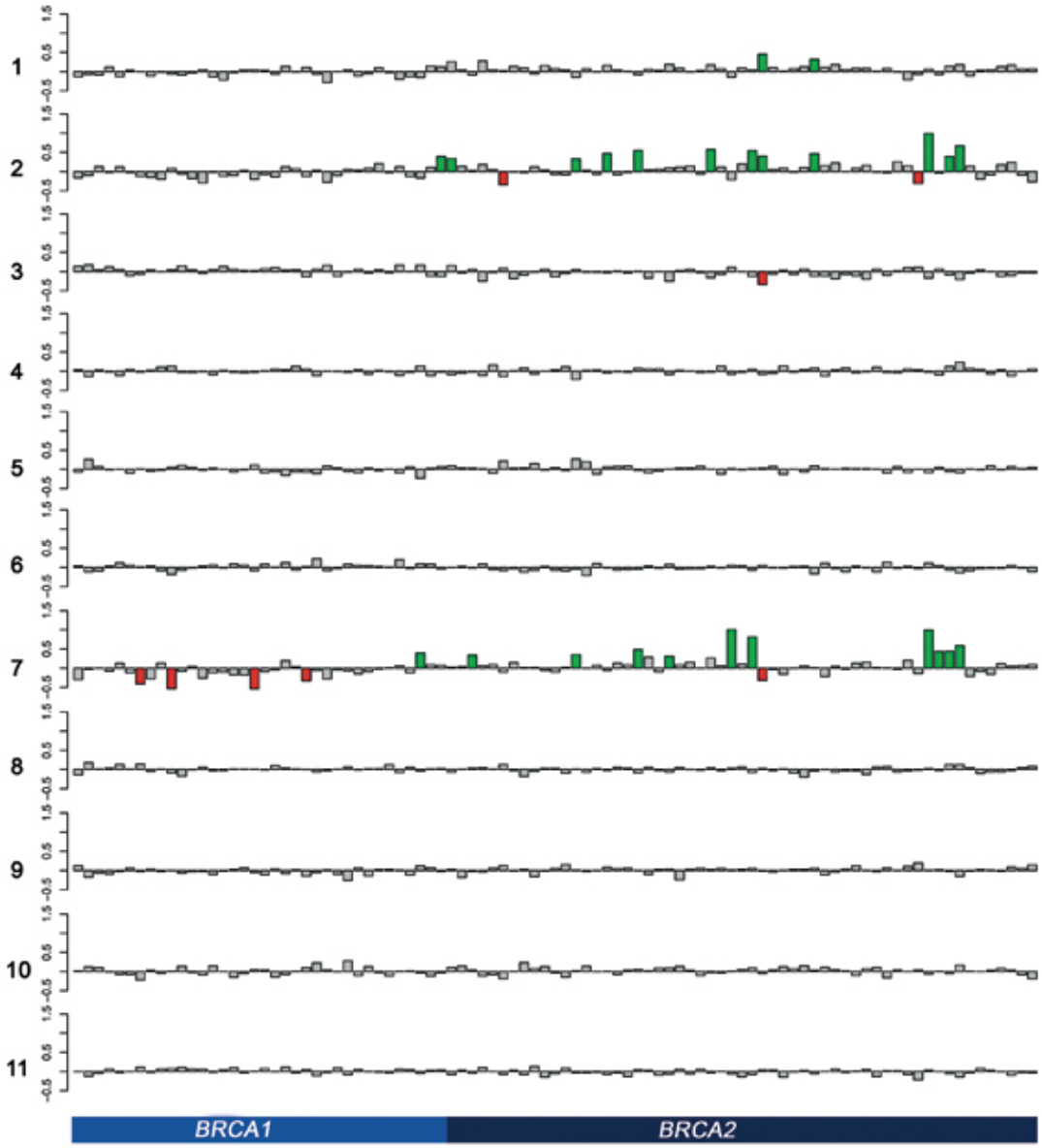
Supplementary Figure 4. Venn diagrams showing similarities between filters 4, 5 and 6
Venn diagrams showing the common and different false positives (A) and true positives (B) that filters 4 (green circle), 5 (blue circle) and 6 (red circle) would discard. Drawn with the Venn diagram generator available at the Chris Seidel web page: http://www.pangloss.com/seidel/Protocols/venn.cgi.



Supplementary Figure 5. Examples of variants in homopolymers and usefulness of the HP kit
Three scenarios in which homopolymers cause confusion and the HP assay or the visual inspection of alignment are needed to correctly classify the variant. For each of them, a portion (the window cannot accommodate all the reads) of the AVA alignment is shown, followed by a relevant portion of the MAQ-S graph for the analysis of the corresponding HP assay. A) BRCA2 c.956dupA, a duplication in a homopolymer of 6 nucleotides, found by VIP and confirmed by the HP assay. B) BRCA1 c.1961delA, a deletion in a homopolymer of 8 nucleotides, not found by VIP in the correct MID (MID6, to the left) but found in some other MIDs (one of them shown to the right) and correctly detected by the HP assay. C) BRCA2 c.8946delA, a deletion in a homopolymer of five nucleotides, covered by the HP assay due to its proximity to a homopolymer of 7 nucleotides, compared to the same region in another MID, shown to the right. Note that the sample to the right without the BRCA2 c.8946delA presents a double peak in a different amplicon of the HP assay, due to the true positive BRCA2 c.2802_2811delACAA, not in a homopolymer but also covered by the kit.

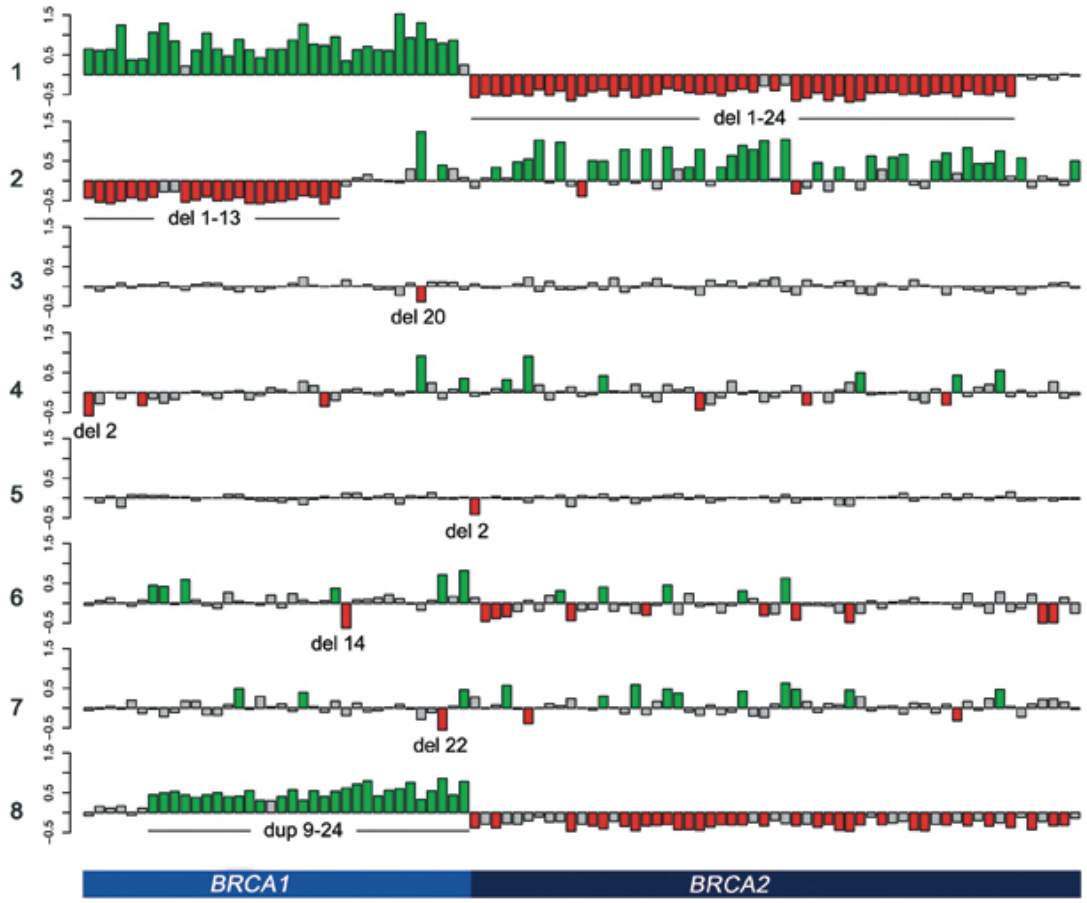


Supplementary Figure 6A. Detection of large genomic rearrangements (LGRs) using NGS results
Bar plots of the dose of NGS amplicons after normalization. X-axis: NGS amplicons. Y-axis: Count ratio minus 1. Fragments with normalized ratios over 1.3 are highlighted in green, indicating putative duplications. Fragments with ratios below 0.7 are shown in red, indicating putative deletions. A) 11 plots from control samples with no alterations, plots 1, 2, 3 and 7 show false positive alterations.



Supplementary Figure 6B. Detection of large genomic rearrangements (LGRs) using NGS results

Bar plots of the dose of NGS amplicons after normalization. X-axis: NGS amplicons. Y-axis: Count ratio minus 1. Fragments with normalized ratios over 1.3 are highlighted in green, indicating putative duplications. Fragments with ratios below 0.7 are shown in red, indicating putative deletions. B) 8 plots representing 8 samples with LGRs: (1) Sample with a deletion of the region comprising *BRCA2* exons 1-24 ; (2) Sample with a deletion comprising *BRCA1* exons 1-13 ; (3) Sample with a deletion comprising exon 20 of *BRCA1;* (4) Sample with a deletion comprising *BRCA1* exon 2*;* (5) Sample with a deletion comprising *BRCA2* exon 2; (6) Sample with a deletion comprising *BRCA1* exon 14; (7) Sample with a deletion comprising *BRCA1* exon 22; (8) Sample with a duplication of the region comprising *BRCA1* exons 9-24.



ARTICLE 2

ICO Amplicon NGS Data Analysis: A Web Tool for Variant Detection in Common High-Risk Hereditary Cancer Genes Analyzed by Amplicon GS Junior Next-Generation Sequencing

Adriana Lopez-Doriga, Lídia Feliubadaló, Mireia Menéndez, Sergio Lopez-Doriga, Francisco D. Morón-Duran, Jesús del Valle, Eva Tornero, Eva Montes, Raquel Cuesta, Olga Campos, Carolina Gómez, Marta Pineda, Sara González, Victor Moreno, Gabriel Capellá, and Conxi Lázaro

Resum del treball: La seqüenciació de nova generació (*Next Generation Sequencing*, NGS) ha revolucionat la recerca genòmica, i va en camí de tenir un major impacte en el diagnòstic genètic gràcies a l'arribada de seqüenciadors de capacitat moderada i de desenvolpaments flexibles per analitzar les regions d'interès determinades per a cada malaltia. Entre les dificultats principals de la NGS es troba la realització de l'anàlisi bioinformàtica, pel gran volum de dades que genera, i per la dificultat d'abordar la gran quantitat de falsos positius que es poden obtenir segons la tecnologia de NGS utilitzada i el protocol d'anàlisi seguit. En aquest article es presenta el desenvolupament d'una eina oberta executable via web, per a detectar i filtrar variants, i per a proporcionar informació de la cobertura, permetent a l'usuari personalitzar alguns paràmetres bàsics. Aquesta eina s'ha desenvolupat per a realitzar les anàlisis genètiques de la reseqüenciació per amplicons d'alguns gens específics d'alt risc per a càncer hereditari, utilitzant el seqüenciador GS Junior. La web està vinculada a la base de dades de mutacions de la nostra institució per ajudar a la classificació clínica de les variants identificades. Creiem que aquesta eina pot facilitar l'ús de la NGS en la rutina d'alguns laboratoris.

A més de l'eina oberta publicada a l'article, s'ha desenvolupat una eina web més personalitzada per a l'ús rutinari de la Unitat de Diagnòstic Molecular, dins del Programa de Càncer Hereditari de l'Institut Català d'Oncologia, on s'han adaptat els formats dels resultats i s'ha afegit un mòdul que permet indicar els encebadors necessaris per a validar per seqüenciació Sanger les variants detectades, d'acord amb els oligonucleòtids disponibles al laboratori.

Contribució de la doctoranda: En aquest article la doctoranda ha realizat l'estudi del nou *software* disponible per a l'anàlisi de les dades i ha fet les proves pertinents amb cadascun d'ells. Un cop triats els *software* per a l'alineament i la detecció de variants, ha realitzat un estudi dels paràmetres més adients per a les dades del GS Junior i les llibreries de Multiplicom. A continuació ha programat les funcions en R per a la generació dels informes amb els resultats de cobertura i de variants. També ha programat el protocol unint cada procés de l'anàlisi (lectura i transformació de les dades, alineament, detecció de

variants i generació d'informes) relacionant-lo amb la pàgina web. Ha dissenyat la pàgina web per a que altres investigadors la poguessin programar. Finalment, ha realitzat la major part de l'escriptura de l'article i la preparació de les figures i de tot el material suplementari de la pàgina web.

INFORMATICS

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

ICO Amplicon NGS Data Analysis: A Web Tool for Variant Detection in Common High-Risk Hereditary Cancer Genes Analyzed by Amplicon GS Junior Next-Generation Sequencing

Adriana Lopez-Doriga,[1,2] Lídia Feliubadaló,[1]* Mireia Menéndez,[1] Sergio Lopez-Doriga,[3] Francisco D. Morón-Duran,[2] Jesús del Valle,[1] Eva Tornero,[1] Eva Montes,[1] Raquel Cuesta,[1] Olga Campos,[1] Carolina Gómez,[1] Marta Pineda,[1] Sara González,[1] Victor Moreno,[2] Gabriel Capellá,[1] and Conxi Lázaro[1]**

[1]*Hereditary Cancer Program, Catalan Institute of Oncology, L'Hospitalet de Llobregat, Barcelona, Spain;* [2]*Prevention Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain;* [3]*Facultat d'Ingenieria Industrial, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain*

Communicated by Madhuri Hegde

Received 26 August 2013; accepted revised manuscript 7 November 2013.
Published online 14 November 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22484

ABSTRACT: Next-generation sequencing (NGS) has revolutionized genomic research and is set to have a major impact on genetic diagnostics thanks to the advent of benchtop sequencers and flexible kits for targeted libraries. Among the main hurdles in NGS are the difficulty of performing bioinformatic analysis of the huge volume of data generated and the high number of false positive calls that could be obtained, depending on the NGS technology and the analysis pipeline. Here, we present the development of a free and user-friendly Web data analysis tool that detects and filters sequence variants, provides coverage information, and allows the user to customize some basic parameters. The tool has been developed to provide accurate genetic analysis of targeted sequencing of common high-risk hereditary cancer genes using amplicon libraries run in a GS Junior System. The Web resource is linked to our own mutation database, to assist in the clinical classification of identified variants. We believe that this tool will greatly facilitate the use of the NGS approach in routine laboratories.

Hum Mutat 0:1–10, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: next generation sequencing; mutation analysis; bioinformatic analysis; variant identification; hereditary cancer

Introduction

In the past 2 years, an enormous shift in the field of genetic testing has begun, mainly due to the implementation of next-generation sequencing (NGS) methodologies in medium scale or benchtop sequencers. The development of these instruments, together with the launch of commercial kits to enrich for a few genes, has permitted the use of NGS for mutational analysis of a small number of genes, in a small number of samples, in a short period of time and in a cost-effective manner, with the quality required for diagnosis purposes [Ku et al., 2013]. Although laboratory protocols for the different platforms are available, analytical expertise is needed to obtain results that can be used for clinical purposes. One of the critical steps when using these instruments is the interpretation of the enormous quantity of data, which underlines the need for bioinformatic tools to better interpret results and to facilitate the implementation of NGS in routine laboratories.

Mutational analysis of high-risk cancer susceptibility genes is mandatory to better determine the familial and individual risk of developing cancer and to personalize surveillance measures and improve genetic counseling [Lynch et al., 2008; Shannon and Chittenden, 2012]. The main genetic analyses performed in hereditary cancer molecular diagnostics units comprise genetic testing of *BRCA1* and *BRCA2* in patients suspected of hereditary breast and ovarian cancer (HBOC) Syndrome, analysis of *APC*, and *MUTYH* in patients suspected of classical or attenuated familial adenomatous polyposis (FAP), and sequencing of mismatch repair genes (*MLH1*, *MLH2*, *MSH6*, and *PMS2*) in patients suspected of hereditary nonpolyposis colorectal cancer (HNPCC) or Lynch syndrome. Mutation analysis of these eight genes allows the identification of germline mutations in a significant number of patients. To sequence these genes, recent medium-throughput sequencing platforms are perfect for small- or medium-sized laboratories as they provide high-resolution

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Lídia Feliubadaló, Unitat de Diagnòstic Molecular, Programa de Càncer Hereditari, Laboratori de Recerca Translacional 2, Institut Català d'Oncologia (ICO-IDIBELL), Hospital Duran i Reynals, Gran Via 199-203, L'Hospitalet de Llobregat, Barcelona 08908, Spain. E-mail: lfeliubadalo@iconcologia.net

**Correspondence to: Conxi Lázaro, Unitat de Diagnòstic Molecular, Programa de Càncer Hereditari, Laboratori de Recerca Translacional 2, Institut Català d'Oncologia (ICO-IDIBELL), Hospital Duran i Reynals, Gran Via 199-203, L'Hospitalet de Llobregat, Barcelona 08908, Spain. E-mail: clazaro@iconcologia.net

Contract grant sponsors: the Spanish Ministry of Health ISCIII FIS grants (PI10/01422, PI13/00285, CA10/01474, RD06/0020/1050, RD12/0036/008 and RD12/0036/0031); the AGAUR Catalan Government Agency grants 2009-SGR293; the Spanish Association Against Cancer (AECC 2010).

results in a short time at a moderate price [Ku et al., 2012]. Among these platforms, 454 GS Junior from Roche AG (Basel, Switzerland) provides the longest reads and allows users to combine targeted libraries of several genes from different bar-coded samples in a single run. Multiplicom (Niel, Belgium) has developed a set of different targeted library generation kits for the analysis of several hereditary cancer syndromes using multiplex amplification, which have proved to be useful for genetic testing [Feliubadaló et al., 2013]. One of the critical points in NGS protocols is the bioinformatic analysis of the huge quantity of data obtained in each run. Some commercial applications are available for pyrosequencing data analysis, such as AVA (Roche), CLC-Bio Genomic Workbench (CLC bio, Aahrus, Denmark), and SeqNext (JSI medical systems GmbH, Kippenheim, Germany). However, the code of these applications is secret and closed, which prevents users from controlling the processes being executed and restricts the parameters that can be customized. Open-source applications like VIP [De Schrijver et al., 2010], BWA alignment algorithm [Li and Durbin, 2009], and R (a language and environment for statistical computing) [R Core Team, 2012] are not user friendly and require users to have some degree of bioinformatics skills.

Here, we present a free, accurate, and user-friendly Web-based tool to perform the bioinformatic analysis of data obtained using the GS Junior sequencer. Our algorithm detects and filters sequence variants, providing coverage information, and allowing the user to customize some of the basic parameters. We hope that this tool will provide invaluable support to users of this platform for hereditary cancer gene analysis.

Overview of the ICO Amplicon NGS Data Analysis Web Tool

Our Web tool combines different open-source applications for the comprehensive analysis of NGS data. These applications are: (1) BWA–MEM (Li, 2013. Aligning sequence reads, clone sequences, and assembly contigs with BWA–MEM, submitted. URL: arXiv:1303.3997 [q-bio.GN]), a recent alignment algorithm optimized for long reads, which is faster and produces more homogeneous multialignments than BLAT, particularly in regions containing runs of identical bases (homopolymers, HPs); (2) Cutadapt [Martin, 2011], a flexible trimming tool; (3) VarScan [Koboldt et al., 2009], a variant-calling software package that produces an exhaustive text file listing all of the variants detected in the alignment (other variant callers like GATK [McKenna et al., 2010] also provide accurate callings, but results in .vcf format are more difficult to process); and (4) a series of functions and filters that we have programmed in R. These functions include: (1) the supplementary coverage descriptive ratio (CDR) variant caller, (2) the Human Genome Variation Society (HGVS) coding sequence nomenclature application, (3) filters based on coverage, allele frequency, and position relative to the coding sequence or the HP assay-covered bases, and (4) the link to our ICO mutation database, which provides a preliminary pathogenicity classification of variants. The Web-based tool presented here is currently available for the BRCA, HNPCC, and FAP MASTR assays. However, the analysis of *PMS2* should be treated with caution due to the high homology of the different pseudogenes aligned to this gene in several regions, which produces a large number of false positive and false negative results. An overview of the complete workflow is presented in Figure 1.

Bioinformatics Data Analysis

The bioinformatics pipeline includes different steps that are automatically combined as represented in Figure 2. Reads in FASTA format and their corresponding qualities in QUAL format are generated by the GS Junior platform after the sequence run. They are analyzed with the following pipeline: (1) reads from two files (.fna and .qual) are merged into a FASTQ file and demultiplexed in different FASTQ files, one for each multiplex identifier (MID); (2) MIDs, tags, and specific primers are trimmed using the Cutadapt code; (3) trimmed reads are aligned by the BWA–MEM algorithm over the human genome hg19 reference sequence; (4) VarScan is used for variant calling; (5) raw variants from VarScan are treated with R and different variables are retrieved or created to describe them (the variables are detailed in Table 1); (6) R commands and in-house functions are implemented to generate coverage information reported in two files: (a) a base coverage report, indicating the bases that do not reach the expected coverage (by default 38x), and (b) an amplicon–MID coverage report, containing the mean coverage for each amplicon within each MID. Amplicon-MID coverage information is also used to generate a bar graph showing putative Large Genomic Rearrangements, gains or losses, after normalizing by MID, gene and plex (i.e., the group of amplicons that are amplified in the same multiplex PCR). It is important to highlight that this graph is only approximate and has not been validated due to a high number of false positives. An extra R function is implemented to calculate CDRs, to detect positions where the experimental coverage is significantly different from the expected coverage. The expected coverage is calculated as the median of the base coverage in each fragment, and a fragment is defined as each group of contiguous bases targeted by the same single amplicon or by two overlapping amplicons. Positions with a CDR < 0.6 indicate putative deletions and are included in the variant report. Most of these deletions are also reported by VarScan, but some insertions or deletions at the very end of an amplicon can be missed by VarScan when the BWA–MEM fails to anchor the short arm of the gap, or when the deletion locates exactly at the end of the trimmed sequence. In those cases, the aligner does not show a gap but rather a shorter sequence, which variant-calling applications cannot identify as a deletion. This phenomenon is common to most aligners and variant callers. However, in these situations, the decrease in coverage of the deleted bases is detected by the CDR function because the limits of the reference fragments preloaded into it do not match the limits of a proportion of the detected reads, thus the ratio (CDR) is <0.6.

A training set of 28 HBOC DNA samples (223 variants) previously characterized by conformation-sensitive capillary electrophoresis (CSCE) and Sanger sequencing was used to set up our NGS workflow. The validation was performed with an independent set of 15 HBOC DNA samples (123 variants) that were analyzed blindly by NGS in parallel with CSCE plus Sanger sequencing, showing a sensitivity of 100% ($\geq$97.5% at a confidence interval of 95%) and a specificity of 100% [Feliubadaló et al., 2013]. Since its publication, this workflow has been routinely applied for the genetic testing of over 200 HBOC patients, and has recently started to be used for the routine diagnosis of FAP and HNPCC patients. This experience has allowed us to refine the bioinformatic analysis for better alignment and visualization. The new bioinformatic analysis presented here has been tested on the same 43 samples (346 variants) and, when embedded in the analysis workflow depicted in Figure 1 and explained in the Supporting Information, has also shown a sensitivity of 100% and a specificity of 100%. The Supporting Information contains an extensive laboratory protocol for library preparation and sequencing. All steps are described in detail, with annotations provided for the most critical, and a useful troubleshooting table is included.

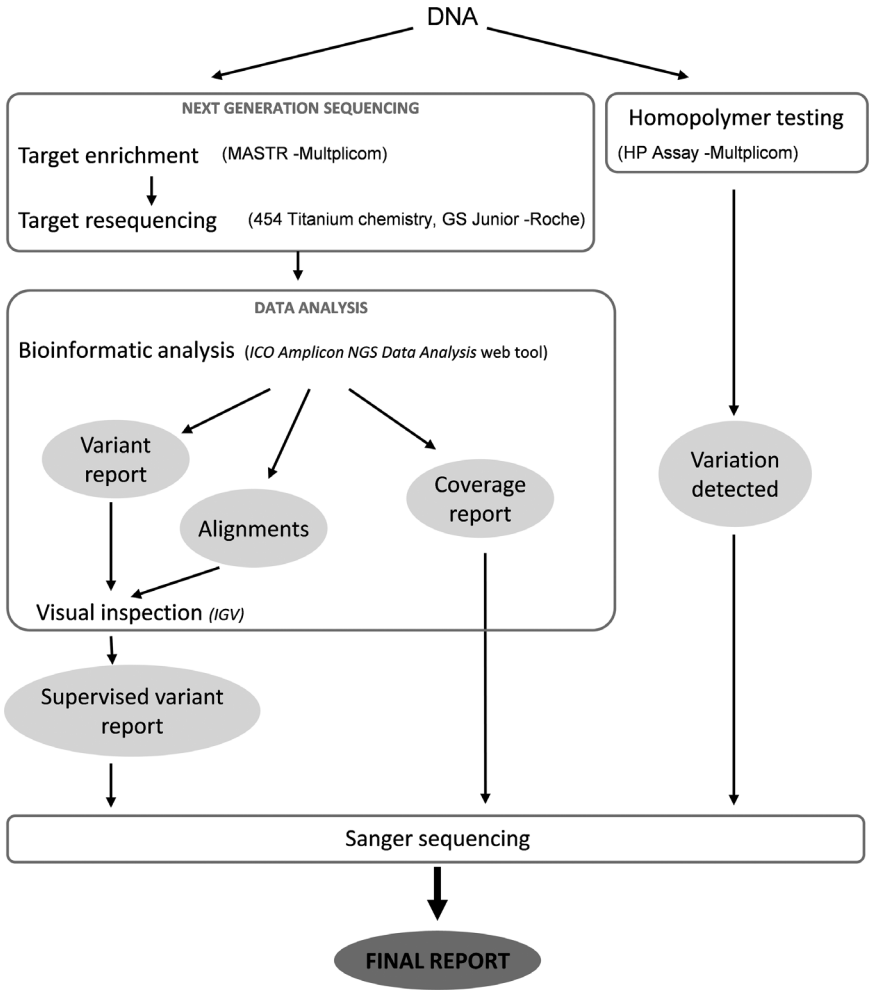


Figure 1. Workflow for analyzing hereditary cancer susceptibility genes using NGS. Starting from genomic DNA patient samples, bar-coded amplicon libraries are prepared using MASTR kits (Multiplicom). Individual libraries are pooled and sequenced using 454 Titanium chemistry in a GS Junior platform (Roche). Data generated by the sequencer are analyzed by the ICO Amplicon NGS Data Analysis Web tool, which processes reads, aligns them to the hg19 reference sequence, calls variants, and generates three kinds of output: variant reports, alignments, and coverage reports. The first two are used to discard some of the remaining false positives by visual inspection in IGV, generating the supervised variant report. Simultaneous screening using the HP kit (Multiplicom) identifies some insertions or deletions located in HPs longer than 5 bp and their surroundings. Any detected aberrant pattern is confirmed by Sanger sequencing, as are all pathogenic and unknown significance variants from the supervised variant report. Regions with low coverage (<38x), if any, are also Sanger sequenced.

Step-by-Step Guidelines for Using the Application

After performing the targeted NGS experiment (see Supporting Information for a detailed laboratory protocol), the user can connect to the Web application at the address http://bioinfo.iconcologia.net/aplicNGS/, select the gene analysis required (i.e., *BRCA1–BRCA2*, *APC–MUTYH*, or *MLH1–MSH2–MSH6–PMS2–EPCAM*) and follow the instructions to upload parameters and data (summarized in Fig. 3). The steps that the user must complete are as follows: (1) fill in the "run name," as all results will contain this name as an identifier, (2) upload the MID names and the corresponding DNA sequences used in the experiment as a tab-delimited text file (a template can be downloaded from the Web), (3) indicate the commercial kit used. The application is currently built to analyze reads from experiments using Multiplicom kits BRCA MASTRv2.1 kit, FAP MASTRv.1, and HNPCC MASTRv.1. If new kit versions are released settings will be updated. (4) Choose thresholds for the coverage report and variant filtering. By default, the application suggests thresholds of 38x for coverage to detect heterozygous alleles with a minimum 25% allele frequency and a sensitivity of 99.9% (these calculations are based on De Leeneer et al. (2011), (5) upload raw read data from GS Junior results in two files, a FASTA file (.fna), and the corresponding reads quality file (.qual), (6) enter the email address to receive the results notification or an email alert if the process has failed, (7) press the "run" button and wait to receive the email.

Data analysis will take between 20 min and 2 hr. For a GS Junior run (about 100,000 reads), the analysis will usually take less than 30 min. The processing time will only increase if many runs are submitted at the same time. If the user receives a failure notification, the process should be resubmitted or the user should contact with the Web application helpdesk.

If any parameter is missing or incorrect, an alert will appear and the user will be requested to upload the files and fill in the parameters again. If all of the resubmitted information is correct, a confirmation message will appear when the "run" button is pressed.

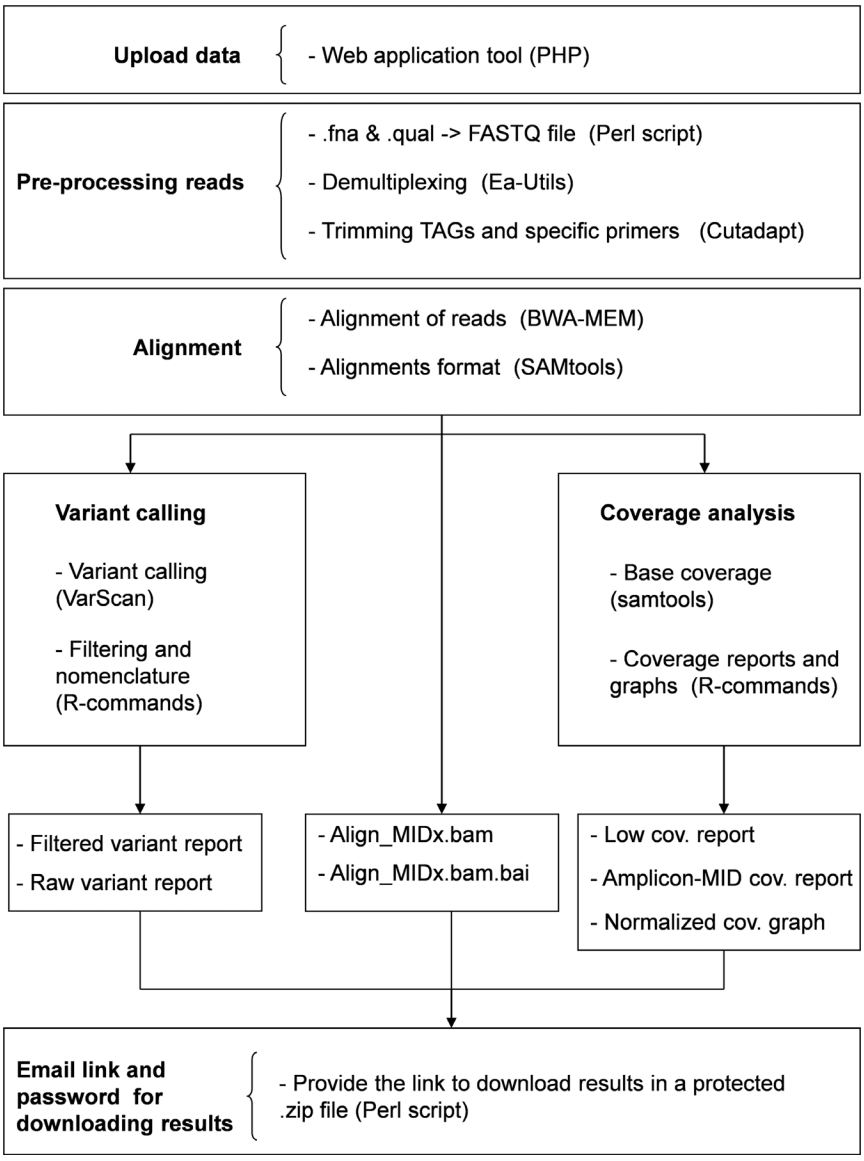


Figure 2. Bioinformatics data analysis pipeline. This bioinformatics pipeline is executed to analyze sequence data submitted to the Web tool for any of the three experiments. Once the user has uploaded data, various applications are used to process reads and generate a FASTQ file for each MID with trimmed reads. Reads are then mapped using the BWA–MEM algorithm, and SAMtools is used to modify the alignment format to allow the user to load and visualize the alignments with IGV, for example. Alignments are also used by VarScan software to call variants and by SAMtools to extract base coverage information. Finally, R-commands are used for variant filtering and nomenclature and for coverage data manipulation, to extract variant and coverage reports. All generated files are zipped and uploaded to a server, and the user is sent an e-mail containing the link for downloading the results.

Results Download and Interpretation

Once the data have been processed, the user will receive an email that contains a link for downloading the results in a .zip file (~12 MB). It is important to check the spam and unwanted email folders as the email system may classify the message wrongly.

In the variant report file, the user will find a filtered list of the detected variants and several associated variables (Table 1). Briefly, variants are defined by technical variables including chromosome, position, variant type, reference and variant alleles, coverage and frequency in each strand, proximity to HPs, and intronic localization. Finally, variants are classified according to our ICO Mutation Database, which is updated every 3 months by adding new variants or reclassifying them at the pace that they are found in our routine diagnostics pipeline. The ICO Mutation Database contains most of the variants detected in the analysis of more than 3,000 probands of HBOC, HNPCC, and FAP over a period of 15 years. The analytical tool provides a putative pathogenicity classification of all variants previously found in our cohort of patients, so it is not necessary to further confirm polymorphisms and neutral variants. In addition, in our diagnostic setting unknown significance variants and putative pathogenic variants, together with undercovered fragments (listed in file "Run_Name_fragments_low_coverage.csv"), if any, are Sanger sequenced before the final report is generated.

In the raw variants report file, the user will find a list of all detected variants without the application of any filter after VarScan calling. This report can be useful for the user willing to customize filtering.

Table 1. Description of Variables in the Variant Report

Variable	Values	Description
Run	Text	Run name uploaded by the user.
MID	Text (MID1,MID2, . . .)	MID name for the samples included in the experiment run.
Chrom	Text (chr1,chr2, . . .)	Chromosome.
Position	Number	Chromosome position, according to hg19.
Specific_gene_ref_pos	Number	Position in a genomic sequence of the corresponding gene. This genomic sequence is the same as used by AVA (Roche) when the corresponding Multiplicom script is uploaded. This position allows the mutations to be checked in AVA.
CDSpos	Text	cDNA position. Variant position according to the HGVS nomenclature guidelines http://www.hgvs.org/mutnomen/.
Variant type	Factor (SNV, INDEL, CDR)	Single-nucleotide substitution variant, insertion or deletion, and CDR.
Reference	Text (A,T,C,G,-)	Nucleotide in the reference sequence.
Alternative	Text (A,T,C,G,N +/- A,T,C,G,N)	Alternative nucleotide found on the reads.
F_Coverage	Number	Number of forward reads covering the variant position.
R_Coverage	Number	Number of reverse reads covering the variant position.
T_Coverage	Number	Number of total reads covering the variant position.
F_Variant_A_Freq	Number	Number of forward reads containing the alternative allele.
R_Variant_A_Freq	Number	Number of reverse reads containing the alternative allele.
T_Variant_A_Freq	Number	Number of total reads (forward + reverse) containing the alternative allele.
T_Variant_R_Freq	Number	Total variant allele relative frequency. Fraction of total (forward + reverse) reads containing the alternative allele versus total reads covering the variant position.
ForwRev0	Factor (no, yes)	Indicates whether the alternative allele is only found in one strand.
HP_situation	Factor (no HP, HP4, HP5, HP ≥ 6, Next_HP ≥ 6)	Indicates whether the variant belongs to a HP of 4, 5, 6, or more nucleotides, if it is located in a position next to HP ≥ 6, or if it is not related to a HP sequence.
Pos_covered_HP_assay	Factor (no, yes)	Indicates whether the variant position is covered by the HP assay. This includes the targeted HPs and also all the bases amplified by the HP assay PCRs, as an insertion or deletion in any of these will cause a mobility shift in the assay.
Coding_and_intron_boundaries	Factor (no, yes)	Yes indicates that the variant position is in or close to the coding region (50 bp before and 20 bp after it, to cover donor, acceptor, and branch splice sites). No indicates deep intronic. In our lab, deep intronic variants are filtered, as they contain lots of long HPs and unknown significance variants.
HGVS_ICO_DB	Text	Most detected variants (mainly polymorphisms but also a significant proportion of pathogenic, neutral, and unknown significance variants) have been previously found in our laboratory, and are assembled and classified in the ICO Mutation Database (unpublished results). A link between the variant found and a variant (if existing) described in the ICO Mutation Database for the same position allows the user to confirm (when the nucleotide change is the same, or the reverse complement change for genes located in the minus strand) if the corresponding mutation has already been detected.
Pathogen_ICO_DB	Factor (POL,PAT,USV,NA)	Pathogenicity classification (POLymorphism/neutral, PAThogenic, unknown significance variant, not present) of the variant using the information from the ICO Mutation Database. If the nucleotide change matches a variant classified in our database, the tool provides a suggested clinical classification, based on the reported frequency of the variant, cosegregation, colocalization in trans with pathogenic variants and clinical data, in silico predictions and/or functional (RNA and protein) internal or published data.
Variant_predictor_filters	Factor (TP,Not_clear,FP)	Indicates whether the variant passes all filters and is a putative true positive, if it is not clear because the variant is present only in the forward or reverse strand (Not_clear), or if the variant does not pass standard filters and is a putative false positive, only present in the raw variant report.

To visualize the reported variants, the user can open IGV [Robinson et al., 2011] and load the obtained alignments. By loading the .bam file from the selected MID, the corresponding .bam.bai index file (which should be in the same folder as the .bam file) will be loaded automatically. More than one MID can be loaded at the same time (for more options, the user can consult the IGV help). The user can visualize the desired variant by introducing its chromosome position; IGV will zoom in on the position and display the reads for the variant. The visualization step is highly recommendable as it allows the user to discard most of the remaining false positive variants, the vast majority of which are sequencing artifacts in or next to HPs that are common to all MIDs. The user will quickly learn to identify and discard these artifacts by comparing variant frequencies from different MIDs or by checking the proportion of variants in the forward or the reverse strand. Other variant callers ignore most variants in HPs to increase the variant-calling specificity at the expense of sensitivity. Indeed, long HPs are fairly common in some of these genes, but reaching or improving on the sensitivity of Sanger sequencing screening requires taking into account the not uncommon deleterious mutations in such regions. For diagnostic purposes, we strongly recommend a visual inspection of these more difficult variants by two independent and highly trained technicians.

The coverage reports contain information about the depth, completeness, and uniformity of the sequencing. The amplicon–MID coverage report allows the user to verify that all MIDs and amplicons have been sequenced evenly. The low-coverage file indicates which bases, if any, have been covered below the selected threshold. In addition, in analyses of *BRCA* genes, the user can display a normalized amplicon–MID coverage bar graph to identify any suspected large genomic copy-number alteration. This option is not yet available for colorectal cancer genes.

Our tool has been designed to analyze data from the three hereditary cancer MASTR libraries described above. The analysis alignment and variant-calling applications are compatible with any kind of amplicon data, but exon positions, primer sequences, and variant pathogenicity information are preloaded to allow trimming,

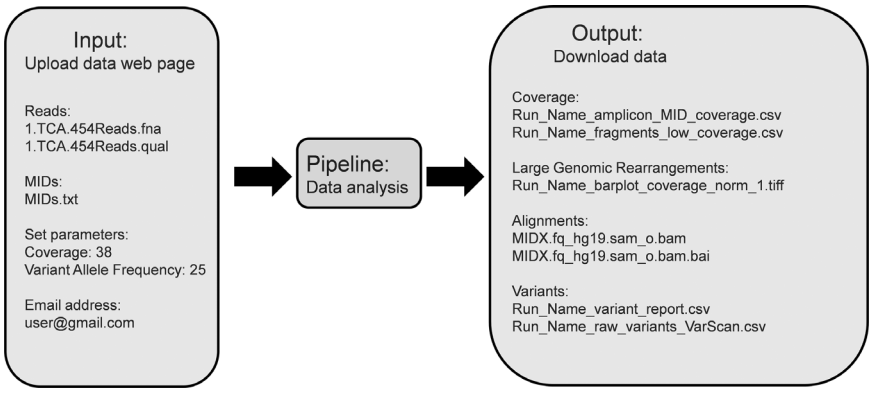


Figure 3. Input and output files for the ICO Amplicon NGS Data Analysis Web tool. Flowgram representing the input files that the user should upload to the Web application and the output files containing the analysis results that user will download. The user must upload reads in FASTA format and the corresponding qualities, a text file with the DNA sequences of the MIDs used for each sample. Coverage and variant allele frequency thresholds can be set for variant filtering. Finally, the user must provide an e-mail address to receive the notification e-mail containing the link for downloading the results. When the user unzips the compressed file with the results, several reports will appear: two .csv files for the coverage report and a plot with amplicon–MID normalized coverage, two more .csv files for the variant report, and two alignment files (.bam, containing the alignment, and .bam.bai, containing the index) for each MID.

A A comma separated file with all reported variants after both alignments and callers, and described by different variables.

MID	Chrom	Position	Gene_pos	CDSpos	Variant_type	Ref	Alternative	F_coverage	R_coverage	T_coverage	F_VA_Freq	R_VA_Freq	T_V_A_freq	ForwRev0	HP_situation	Pos_cov_HP_assay	Cod_intron_bound	Patogen_ICO_DB	HGVS_ICO_DB	Variant_predictor_filters
MID1	Chr17	41246481	31020	BRCA1 exon11 c.1067	SNV	T	C	87	76	163	34	77	47	No	No_HP	Yes	Yes	POL	c.1067A>G	TP
MID1	Chr17	41245471	31030	BRCA1 exon11 c.2077	SNV	C	T	67	52	119	29	61	57	No	No_HP	No	Yes	POL	c.2077G>A	TP
MID3	Chr17	41245466	31035	BRCA1 exon11 c.2082	SNV	G	A	56	28	84	10	41	48	No	No_HP	No	Yes	POL	c.2082C>T	TP
MID4	Chr17	41245354	32147	BRCA1 exon11 c.2194	INDEL	C	-T	85	14	99	33	0	33	Yes	No_HP	No	Yes	NA	NA	Not_Clear
MID5	chr17	32907421	17805	BRCA2 exon10 c.1806	CDR	N	N	NA	NA	61	NA	NA	0.59	No	HP>=6	Yes	Yes	NA	NA	TP

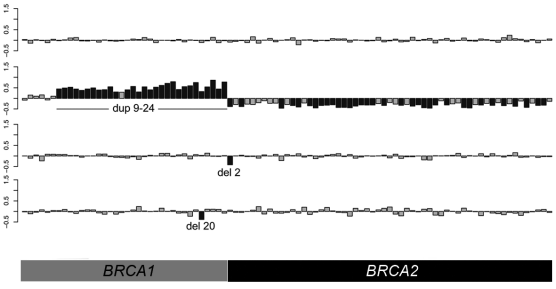
B Total coverage for each amplicon within each MID

Amplicon	MID1	MID2	MID3	MIDx…
BRCA1_ex02_01	241.1	271.5	129.6	…
BRCA1_ex03_01	131.3	233.3	133.5	…
BRCA1_ex05_01	178.3	199.8	163.5	…
Amplicon_x …..	…	…	…	…

C Fragments not reaching the expected coverage

MID	Chrom	Position	CDSpos	Coverage	HP_situation	Pos_cov_HP_assay
MID2	Chr13	32907536	BRCA2 intron10 c.1909+12	29	HP >=6	No
MID2	Chr17	41197676	BRCA1 intron24 c.5706+19	22	No_HP	No
MID2	Chr17	41197677	BRCA1 intron24 c.5706+18	22	No_HP	No
MIDX	…	…	…	…	…	…

D Bar graph showing normalized amplicon-MID coverage. In the original plot, putative duplications are displayed in green, while putative deletions are in red



E BWA-MEM alignments visualized in IGV visualizer, showing a 4 bp deletion in the left image and a substitution in the right image.

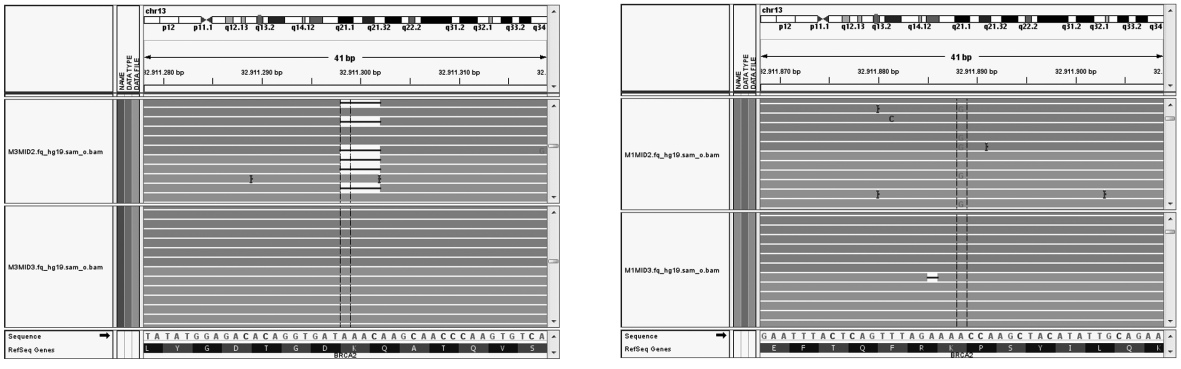


Figure 4. Anticipated results. An example of the main results files is shown. **A:** Extract of a filtered variant report, with variants in rows and descriptive variables in columns. **B:** Amplicon–MID coverage report, with amplicons in rows and coverage for the different MIDs in columns. **C:** Low coverage base report, with each low coverage base in a row, described by MID, position and coverage in columns. **D:** Example of a bar graph representing normalized amplicon–MID coverage, adapted from Feliubadaló et al. (2013). **E:** Screen shot representing the visualization in IGV of a specific position that corresponds to a deletion to the left and a substitution to the right for the alignments of two different MIDs.

HGVS-compliant variant nomenclature, amplicon coverage calculations, deep-intronic variant filtering, and HP variant management through additional user-friendly input steps. The tool has been implemented to analyze GS Junior runs.

Example of Anticipated Results

As an example of the processed bioinformatic results that the application returns, a model BRCA experiment with eight samples is depicted. Detailed information for the different result files is provided, and a snapshot of each output is shown in Figure 4.

- Raw variants from VarScan file: a comma-separated file (Run_Name_raw_variants_VarScan.csv) listing all of the variants detected by VarScan and the informative variables described in Table 1, before any filter is applied. This file allows new users to check which variants are filtered out, tune the thresholds, and even apply their own favorite filters to the detected variants.
- Variant report file: a comma-separated file (Run_Name_variant_report.csv) listing the variants that remain after filtering (Fig. 4A). Variants discarded by filters are: (1) variants with total coverage lower than the coverage threshold (38 by default), (2) variants with total variant allele frequency lower than the frequency threshold (25% by default), (3) insertions and deletions located in regions covered by the corresponding HP assay (Multiplicom; see Supporting Information for details), and (4) variants located deep in intronic regions (after position +20 bp from the donor site and before position −50 bp from the acceptor site) [Feliubadaló et al., 2013]. Variants are ordered by MID, gene, and position.
- Amplicon–MID coverage file: a comma-separated file (Run_Name_amplicon_MID_coverage.csv) listing the mean coverage for each amplicon and each MID, that is, the number of reads mapping over each amplicon for each different MID, shown in a table format with amplicons in rows and MIDs in columns (Fig. 4B).
- Fragment coverage file: a comma-separated file (Run_Name_fragments_low_coverage.csv) listing the bases that do not reach the coverage threshold. This file is useful in diagnostic settings where the few undercovered regions are usually Sanger sequenced, to ensure maximal sensitivity. If no low coverage bases are present, the file will only contain the message: "no fragment under the required coverage" (Fig. 4C).
- Bar graph indicating putative large rearrangements: a visual representation of each amplicon of the analyzed genes is presented. Amplicon coverage is normalized by the plex mean, the mean of other gene/s in the kit and the MID mean, to detect dose changes suggestive of structural variations. Green bars indicate putative duplications (normalized coverage > 1.3) and red bars indicate putative deletions (normalized coverage < 0.7) (Fig. 4B). At the time of writing this manuscript, the bar graph is only available for BRCA experiments (Fig. 4D).
- Alignments over hg19 reference sequence: a binary alignment file (MIDX.fq_hg19.sam_o.bam) and the corresponding index (MIDX.fq_hg19.sam_o.bam.bai) for each MID. These files can be easily uploaded to the IGV visualizer [Robinson et al., 2011] to check the alignments for variants of interest (Fig. 4E).

Conclusions

We present the development of ICO Amplicon NGS Data Analysis, a user-friendly Web tool for the bioinformatic analysis of sequencing data from GS Junior technology that works in any browser and provides tabulated results in comma-delimited files. This tool combines BWA–MEM alignment and VarScan variant calling, complemented by a series of functions and filters programmed in R that provide reports for coverage and for variant identification.

In the near future, our intention is to add the possibility of running the analysis for the same genes but with custom primers. This will make the application useful for more users. Moreover, Roche FLX will be run upon request (due to the larger amount of memory needed). Lastly, we are also studying the possibility of adapting the pipeline to analyze data from PGM Ion Torrent platforms. In summary, we anticipate continuous updates of our ICO Amplicon NGS Data Analysis tool as new developments emerge or upon requests from our users. All updates will be documented on the Web page.

Acknowledgments

We wish to thank all members of the ICO Hereditary Cancer Program team for their constant support and willingness to apply new technological developments within the Molecular Diagnostics Unit. Thanks to Antoni Berenguer and Eduard Serra for their help and support during the development of the tool and the preparation of the manuscript.

The project was conceived and the experiments and data analyses coordinated by A.L.D., M.M., L.F., G.C., and C.L. Samples were genetically characterized by J.D.V., M.M., E.T., E.M., R.C., C.G., O.C., M.P., and S.G. Bioinformatic analysis was performed by A.L.D., S.L.D., F.D.M., and V.M. The manuscript was written by L.F., A.L.D., M.M., and C.L. and was discussed and improved by all authors.

Disclosure statement: The authors declare no conflict of interest.

References

De Leeneer K, De Schrijver J, Clement L, Baetens M, Lefever S, De Keulenaer S, Van Criekinge W, Deforce D, Van Nieuwerburgh F, Bekaert S, Pattyn F, De Wilde B, et al. 2011. Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. PLoS One 6:e25531.

De Schrijver JM, De Leeneer K, Lefever S, Sabbe N, Pattyn F, Van Nieuwerburgh F, Coucke P, Deforce D, Vandesompele J, Bekaert S, Hellemans J, Van Criekinge W. 2010. Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. BMC Bioinformatics 11:269.

Feliubadaló L, Lopez-Doriga A, Castellsague E, Del Valle J, Menendez M, Tornero E, Montes E, Cuesta R, Gomez C, Campos O, Pineda M, Gonzalez S, et al. 2013. Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. Eur J Hum Genet 21:864–870.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285.

Ku CS, Cooper DN, Iacopetta B, Roukos DH. 2013. Integrating next-generation sequencing into the diagnostic testing of inherited cancer predisposition. Clin Genet 83:2–6.

Ku CS, Wu M, Cooper DN, Naidoo N, Pawitan Y, Pang B, Iacopetta B, Soong R. 2012. Technological advances in DNA sequence enrichment and sequencing for germline genetic diagnosis. Expert Rev Mol Diagn 12:159–173.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Lynch HT, Lynch JF, Lynch PM, Attard T. 2008. Hereditary colorectal cancer syndromes: molecular genetics, genetic counseling, diagnosis and management. Fam Cancer 7:27–39.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303.

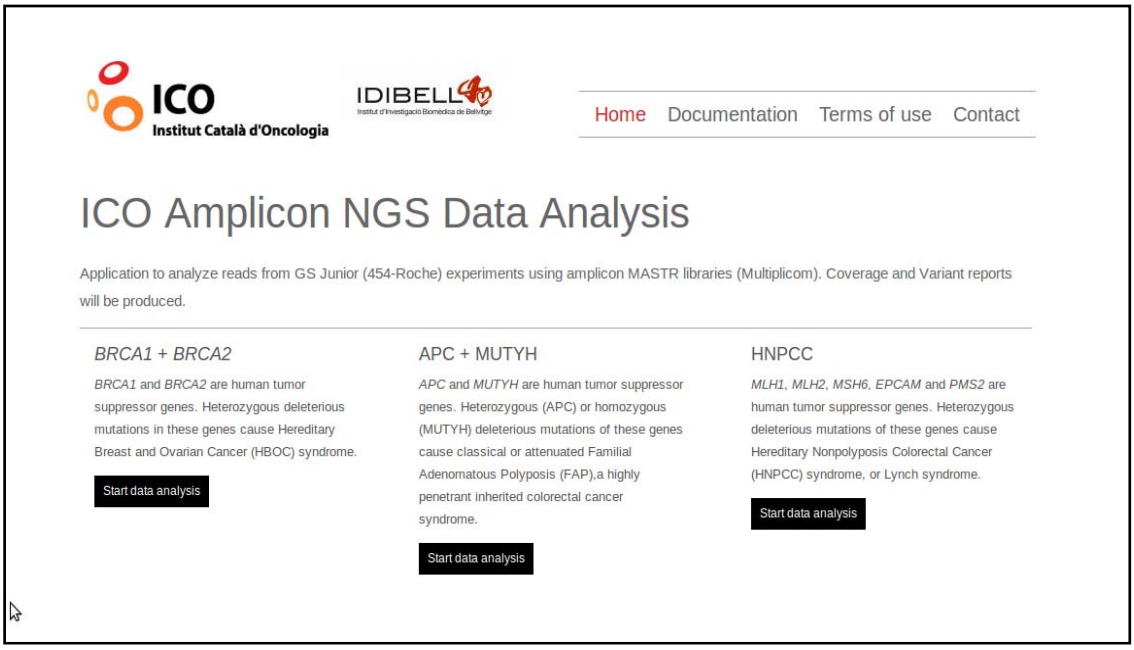
R Core Team. 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing V, Austria. ISBN: 3-900051-07-0.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

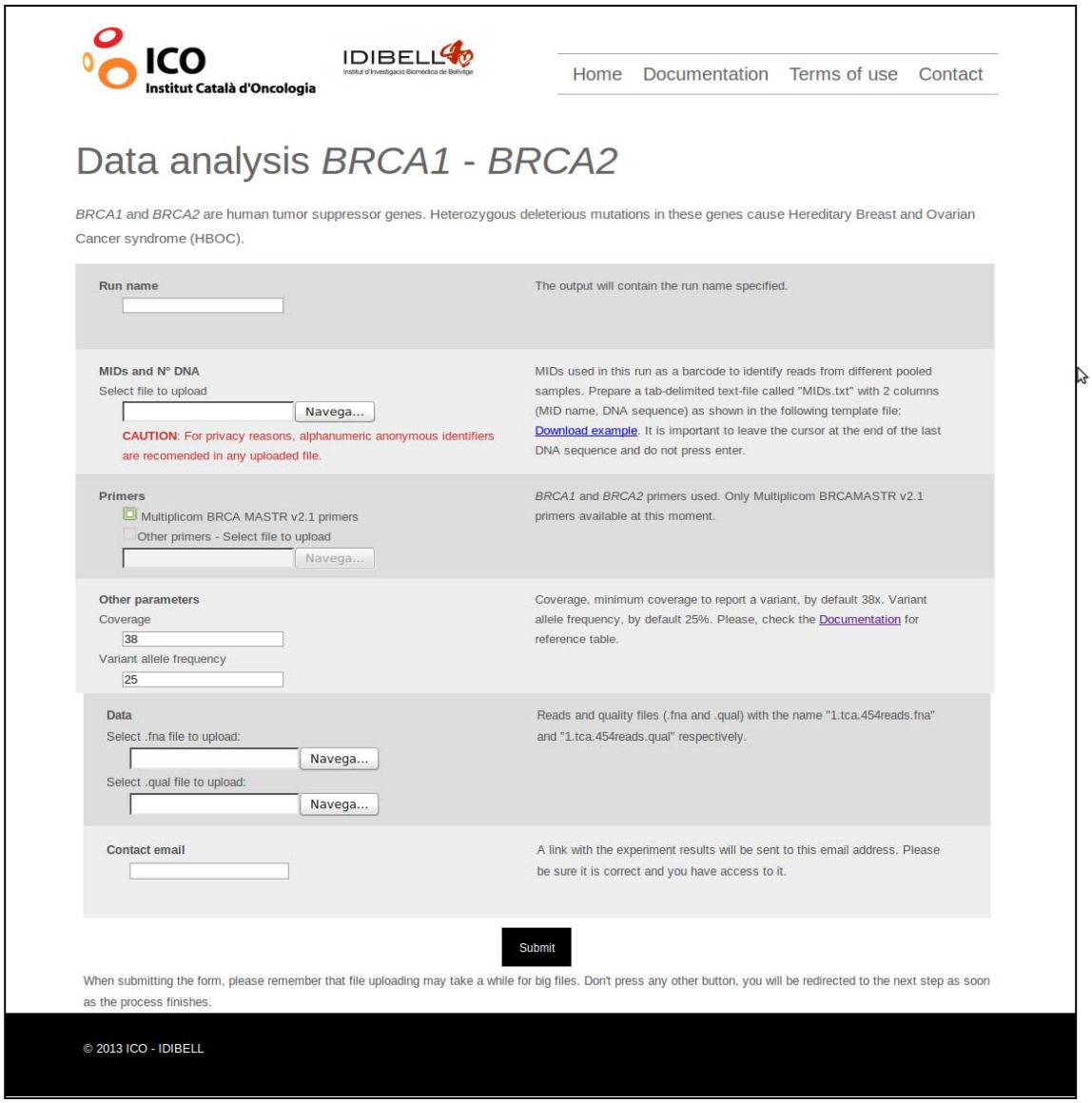
Shannon KM, Chittenden A. 2012. Genetic testing by cancer site: breast. Cancer J 18:310–319.

Material suplementari disponible a l'aplicació web

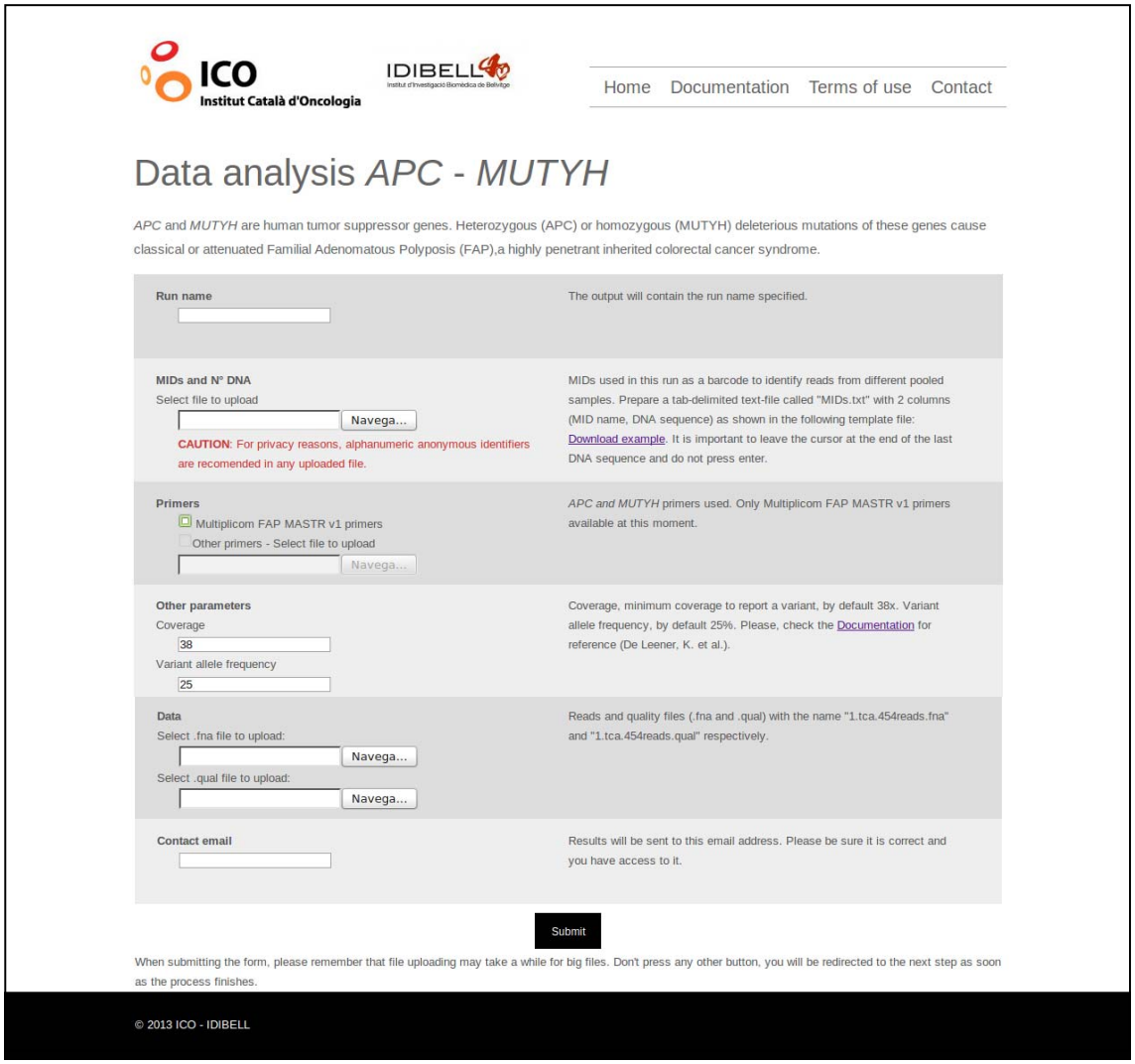
Material suplementari 1: Pàgina principal de l'aplicació web



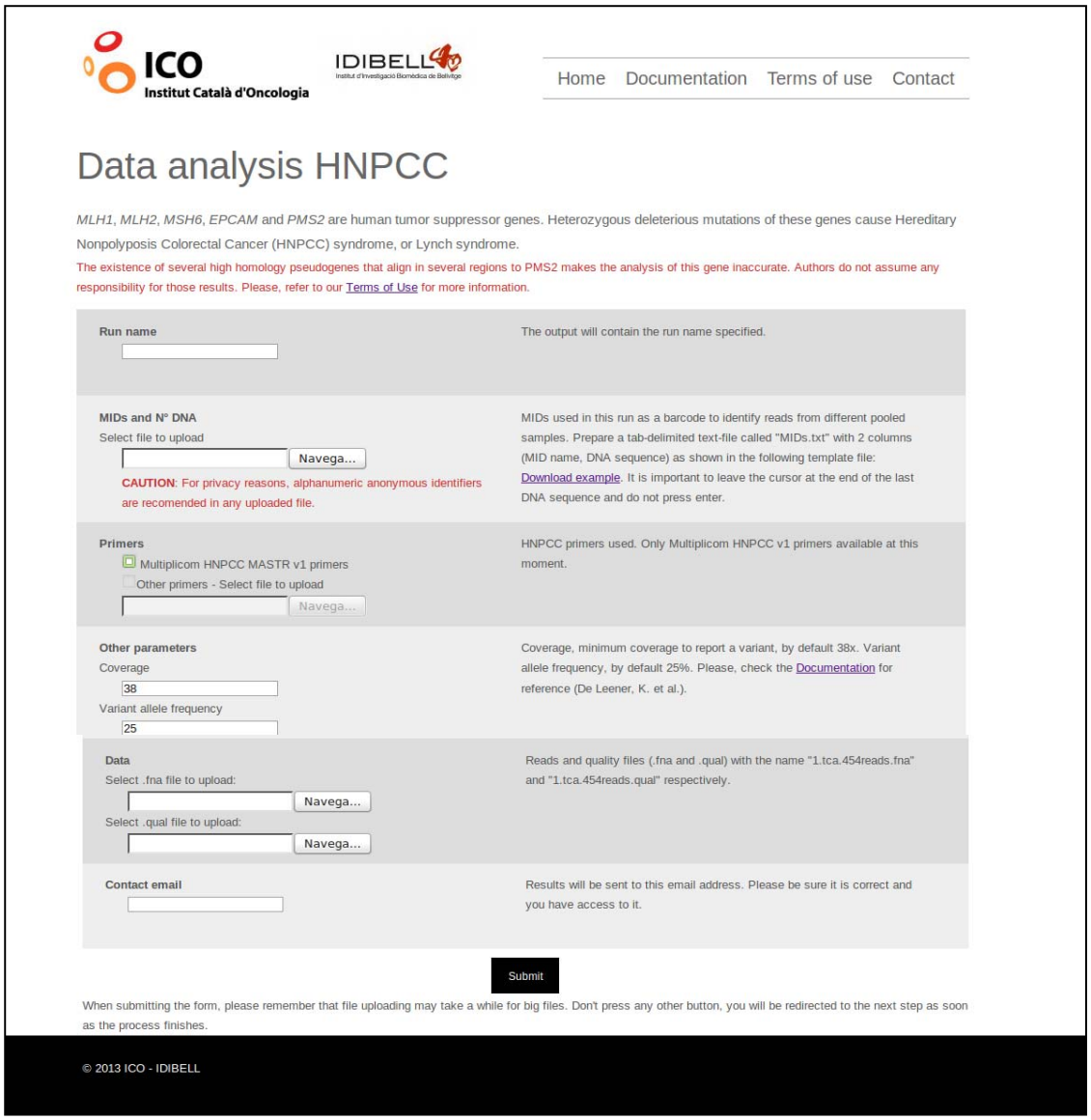
Material suplementari 2: Pàgina de l'aplicació web per a entrar les dades en l'anàlisi dels gens *BRCA1* i *BRCA2*.



Material suplementari 3: Pàgina de l'aplicació web per a entrar les dades en l'anàlisi dels gens *APC* i *MUTYH*.



Material suplementari 4: Pàgina de l'aplicació web per a entrar les dades en l'anàlisi dels gens reparadors (d'alt risc per HNPCC) *MLH1, MLH2, MSH6, EPCAM i PMS2*.



Material suplementari 5: Apèndix de la documentació de l'aplicació web amb el codi font del protocol d'anàlisi, exemple per a un anàlisi dels gens *BRCA1* i *BRCA2*.

```
#####################################################################################################################
#####################################################################################################################
#####################################################################################################################

#!/bin/bash

#Input Variables
run_name="$1"
coverage_in=$2
var_freq_in=$3
email_adress="$4"

path_data="/home/junior454/Aplic_J454_DA/DATA_454"
path_results="/home/junior454/Aplic_J454_DA/RESULTS_454"
path_Aplic="/home/junior454/Aplic_J454_DA"

r_program_1="r_program_Aplic_J454_coverage_BRCAs.R"
r_program_2="r_program_Aplic_J454_variants_BRCAs.R"



# STEP 1: Load data
#---------------------------------------------------------------------------------------------------------------------
# copy files reads.fna, reads.qual and MIDs.txt to "/home/junior454/Aplic_J454_DA/DATA_454/$run_name " folder.


# STEP 2: Create directories for intermediate and final results
#---------------------------------------------------------------------------------------------------------------------
mkdir ${path_results}/$run_name
mkdir ${path_results}/$run_name/Demultiplexed
mkdir ${path_results}/$run_name/temp
mkdir ${path_results}/$run_name/Trimmed
mkdir ${path_results}/$run_name/Alignments
mkdir ${path_results}/$run_name/Base_coverage
mkdir ${path_results}/$run_name/Variants
mkdir ${path_results}/$run_name/Variants/snps
mkdir ${path_results}/$run_name/Variants/indels
mkdir ${path_results}/$run_name/Report_files


# STEP 3: Make fastq file format
#---------------------------------------------------------------------------------------------------------------------
code_fastq=${path_Aplic}/soft
raw_data=${path_data}/$run_name

cd ${raw_data}
/share/apps/Perl/bin/perl ${code_fastq}/perl_fna_qual_to_fastq.pl 1.TCA.454Reads.fna


# STEP 4: Demultiplexing
#---------------------------------------------------------------------------------------------------------------------
eautils=${path_Aplic}/soft/ea_utils/bin
raw_data=${path_data}/$run_name
demultiplexed=${path_results}/$run_name/Demultiplexed

${eautils}/fastq-multx -bx ${raw_data}/MIDs.txt ${raw_data}/1.TCA.454Reads.fastq -o ${demultiplexed}/%.fq

#Remove unmatched.fq
rm ${demultiplexed}/unmatched.fq


# STEP 5: Trimming
#---------------------------------------------------------------------------------------------------------------------
cutadapt=${path_Aplic}/soft/cutadapt-1.2.1/bin
demultiplexed=${path_results}/$run_name/Demultiplexed
temp=${path_results}/$run_name/temp
trimmed=${path_results}/$run_name/Trimmed
```

```
# First trimming forward MID+TAG+specific_primer and then trimming reverse MID+TAG+specific_primer
cd ${demultiplexed}
for fl in *.fq; do
${cutadapt}/cutadapt $(<${path_Aplic}/soft/BRCAsTAGsPrimers_Forw_cutadaptconfig.conf) ${demultiplexed}/${fl} >
$(International Cancer Genome, Hudson et al.)/${fl}
done

cd $(International Cancer Genome, Hudson et al.)
for fl in *.fq; do
${cutadapt}/cutadapt $(<${path_Aplic}/soft/BRCAsTAGsPrimers_Rev_cutadaptconfig.conf) $(International Cancer Genome,
Hudson et al.)/${fl} > ${trimmed}/${fl}
done


# STEP 6: Alignment over hg19 using BWA-MEM
#-------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------
bwa=/share/apps/bwa
ref=${path_Aplic}/soft/ref_bwa/hg19ref
trimmed=${path_results}/$run_name/Trimmed
alignments=${path_results}/$run_name/Alignments

cd ${trimmed}
for fl in *.fq; do
cd $bwa
./bwa mem ${ref} ${trimmed}/${fl} > ${alignments}/${fl}_hg19.sam
done


# STEP 7: Alignment manipulation using samtools, .sam to binary .bam, order and create index for IGV visualization
#-------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------
samtools=/share/apps/samtools/samtools
alignments=${path_results}/$run_name/Alignments

cd ${alignments}
for fl in *.sam; do
$samtools view -bS ${alignments}/${fl} -o ${alignments}/${fl}.bam
$samtools sort ${alignments}/${fl}.bam ${alignments}/${fl}_o
$samtools index ${alignments}/${fl}_o.bam
done

for fl in *hg19.sam; do
rm ${fl}
done

for fl in *hg19.sam.bam; do
rm ${fl}
done


# STEP 8: Extract base coverage from alignments
#-------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------
samtools=/share/apps/samtools
alignments=${path_results}/$run_name/Alignments
base_cov=${path_results}/$run_name/Base_coverage
samtools=/share/apps/samtools

cd ${alignments}
for fl in *.bam; do
$samtools/samtools depth ${alignments}/${fl} > ${base_cov}/${fl}.txt
done


# STEP 9: Variant calling using VarScan
#-------------------------------------------------------------------------------------------------------------------------------------------------------
---------------------
varscan=${path_Aplic}/soft/VarScan
samtools=/share/apps/samtools/samtools
ref=${path_Aplic}/soft/ref_hg19_fasta
alignments=${path_results}/$run_name/Alignments
variants=${path_results}/$run_name/Variants

cd ${alignments}
```

```
for fl in *_o.bam; do
#snps
$samtools mpileup -f ${ref}/ucsc.hg19.fasta ${alignments}/${fl} | /usr/java/latest/bin/java -jar ${varscan}/VarScan.v2.3.3.jar
pileup2snp --min-avg-qual 0 >${variants}/snps/snps_varscan_${fl}.txt
#indels
$samtools mpileup -f ${ref}/ucsc.hg19.fasta ${alignments}/${fl} | /usr/java/latest/bin/java -jar ${varscan}/VarScan.v2.3.3.jar
pileup2indel --min-avg-qual 0 >${variants}/indels/indels_varscan_${fl}.txt
done


# STEP 10: Read base coverage and report low coverage bases or fragments, amplicon-MID distribution, and CDRs as a putative
indels
#-----------------------------------------------------------------------------------------------------------------------------
---------------------
cd $path_Aplic
/share/apps/R --no-save --args "${run_name} ${coverage_in} ${var_freq_in}" < $r_program_1


# STEP 11: Read variants from VarScan, filtering, add putative CDR indels ,add nomenclature in CDS, and add ICO-Database
information
#-----------------------------------------------------------------------------------------------------------------------------
---------------------
cd $path_Aplic
/share/apps/R --no-save --args "${run_name} ${coverage_in} ${var_freq_in}" < $r_program_2


# STEP 12: Compress results in a zip file, make a code and a link for download results
#-----------------------------------------------------------------------------------------------------------------------------
---------------------
cd ${path_results}/$run_name
zip -r results_$run_name Report_files Alignments

codimd5=`md5sum results_$run_name.zip|cut -d' ' -f1`

cp results_$run_name.zip ../${codimd5}.zip
codibase64=$(echo -n $codimd5 | base64)
link="http://bioinfo.iconcologia.net/aplicNGS/downloads.php?task=$codibase64"

# STEP 13: Send the email with the link to download results
#-----------------------------------------------------------------------------------------------------------------------------
---------------------
code_mail=${path_Aplic}/soft
/share/apps/Perl/bin/perl ${code_mail}/mail.pl $email_adress ${link}

###############################################################################################################################
###############################################################################################################################
###############################################################################################################################
```

Note: code for R-functions are not publicly available as it contains implicit information about primers positions which is property of Multiplicom. However, it is not difficult for an R user to process raw results and get the optimal information. We provide steps to process the information in appendix 2 and 3. If user has troubles reproducing those steps, do not hesitate to contact us by the web contact desk and we will try to give specific help.

```
###############################################################################################################################
###############################################################################################################################
###############################################################################################################################
```

Steps of r_program_1="r_program_Aplic_J454_coverage_BRCAs.R"

##
##
##

--
STEP 1: Pass parameters and load homemade functions
#--

1.1: Pass parameters (run name, coverage and variant allele frequency)

1.2: Load base coverage reference sequence positions and functions: a function indicating all positions covered by the MASTR assay, a function indicating fragment positions (a fragment is defined as each group of contiguous bases targeted by the same single amplicon or by two overlapping amplicons), amplicon names and chromosome positions where the are mapped.

1.3: Load functions for CDS nomenclature, HP positions, amplicon positions and HP assay: a function that makes the conversion from chromosome position to CDS nomenclature based on ensemble exon positions from the specific transcript (attached there is an excel sheet with these positions correspondences), chromosome positions covered by the MASTR assay, a function that indicates the homopolymer situation of a specific position and also if it is covered by the corresponding HP assay.

--
STEP 2: Read base coverage from all MIDs
#--

2.1: Read base coverage from all MIDs in folder "/home/junior454/Aplic_J454_DA/RESULTS_454/",run_name,"/Base_coverage" and save it in a matrix. Matrix has 4 variables ("Chrom","position","base_cov","MID")

--
STEP 3: Define low base coverage fragments
#--

3.1: From the matrix with all MIDs, select only bases that are suppose to be covered, some reads could be mapped out of the target genes and could produce wrong results. Also check that all target region is covered, if not, insert positions with coverage 0 to report them later as uncovered fragment.

3.2: Filter out bases with coverage < defined coverage

3.3: Add CDS nomenclature, HP situation and HP assay information, and intronic indication if it is (< c.-50 or > c.+20)

3.4: Filter out intronic regions (< c.-50 or > c.+20)

3.5: Select variables, and order. At the end we obtain a data frame with all positions with low coverage in rows and variables ("MID", "Chrom", "position", "CDSpos", "base coverage", "Homopolymer situation", "pos covered by HP assay") in columns, ordered by MID, chromosome and position.

3.6: Print results if the number of rows is >1, or print a message ("No fragments under coverage defined") if there is no bases with low coverage.

--
STEP 4: Indels detector based on base coverage (Coverage Descriptive Ratio, CDR)
#--

4.1: Recover the matrix with coverage from all target bases and all MIDs from the STEP 3.

4.2: Make a function to detect all positions where the coverage is different from coverage of the surrounding bases, but it is not due to amplicon overlapping.

4.3: Estimate the median for each fragment, fragments are defined in loaded functions. Then calculate the ratio between the coverage from each position and the corresponding fragment coverage median. (We used the fuction "tapply" for these calculations)

4.4: Select positions where ratio is <=0.6 as a putative indel, we call them CDRs.

4.5: In order to avoid repetitive False Positives due to indels next to HP, we discard CDRs that are present in 3 or more MIDs in the analyzed run.

4.6: Print results in a temporally file to later include CDRs in the variant report.

--

STEP 5: Amplicon-MID coverage report
#----------

5.1: Recover the matrix with coverage from all target bases and all MIDs from the STEP 3.

5.2: Select only positions that are not in overlapping amplicons.

5.3: Calculate the median for each amplicon (only bases previously selected)

5.4: Print Amplicon-MID coverage results in a data frame where amplicons from the MASTR assay are in rows and different MIDs in columns, the matrix contains the median coverage for each amplicon and MID.

STEP 6: Coverage Graph normalized by plex, MID and gene
#----------

6.1: Recover matrix from Amplicon-MID coverage in step 5

6.2: Add plex and gene information to each amplicon

6.3: Normalize coverages by plex and by gene

6.4: Divide normalized coverages by the mean of all samples except the one analyzed to normalize by sample

6.5: Order amplicons

6.6: Substract 1 from the normalized ratios

6.7: Barplot showing differences to 1 (red <-0.3, grey -0.3to0.3, green>0.3).

##########
##########
##########

Steps of r_program_2="r_program_Aplic_J454_variants_BRCAs.R"

```
#########################################################################################################################
#########################################################################################################################
#########################################################################################################################

# ---------------------------------------------------------------------
# STEP 1: Pass parameters and load homemade functions
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# 1.1: Pass parameters (run name, coverage and variant allele frequency)

# 1.2: Load functions for CDS nomenclature, HP positions, amplicon positions and HP assay: a function that makes the conversion from chromosome position to CDS nomenclature based on ensemble exon positions from the specific transcript (in appendix 4 there is a table with these positions correspondences), chromosome positions covered by the MASTR assay, a function that indicates the homopolymer situation of a specific position and also if it is covered by the corresponding HP assay.

# 1.3: Load variants from ICO Mutation Database

# ---------------------------------------------------------------------
# STEP 2: Read variants from VarScan calling for all MIDs
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# 2.1: Read separately snps and indels from VarScan results from all MIDs reported in folder "/RESULTS_454/",run_name,"/Variants/" and save them in a matrix.

# 2.2: Join SNPs and INDELs in the same matrix with a veriable indicating if it is a snp or an indel, and create variables of *Total_coverage* and *Total_variant_allele_frequency.*

# ---------------------------------------------------------------------
# STEP 3: Add coverage descriptive ratios (CDR) as a putative indels from the file created in step 4 in coverage code.
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# ---------------------------------------------------------------------
# STEP 4: Create variables that will be used for the analysis
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# 4.1: Add gene name depending on chromosome and positions

# 4.2: Add a variable indicating if variant coverage in forward or reverse is 0.

# 4.3: Add CDS nomenclature, information about HP situation and HPassay, and intronic position indication if it is (< c.-50 or > c.+20)

# 4.4: Add information about known variants from ICO Mutation Database

# ---------------------------------------------------------------------
# STEP 5: Print raw variants report
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# 5.1: Select variables and print variants in a data frame with no filters applied.

# ---------------------------------------------------------------------
# STEP 6: Filtering
#----------------------------------------------------------------------------------------------------------------------------------------------------------------

# 6.1: Filter out variants with T_Coveargе<defined coverage (default=38)

# 6.2: Filter out variants with T_Variant_Rel_Freq<defined variant allele frequency (default=0.25)

# 6.3: Annotate as "Not Clear" those variants that are only in forward or revers strand.

# 6.4: Filter out indels covered by HP assay but not CDRs

# 6.5: Filter out intronic variants further than c.-50 or c.+20 from the exon.


# ---------------------------------------------------------------------
```

```
# STEP 7: Order, name variables and print
#-------------------------------------------------------------------------------------------------------------------------------------------------------------------

# 7.1: Order variants by CDS position and MID

# 7.2: Name variables for a correct interpretation

# 7.3: Print variant report with remaining variants and selected variables for visualization and interpretation.


#####################################################################################################################
#####################################################################################################################
#####################################################################################################################
```

ALTRES CONTRIBUCIONS

AVALUACIÓ DE L'EFICIÈNCIA DEL PANELL *TRUSIGHT CANCER* D'ILLUMINA EN EL DIAGNÒSTIC GENÈTIC DEL CÀNCER HEREDITARI

Introducció

Amb els avenços tecnològics de la seqüenciació massiva es fa més factible seqüenciar més gens de manera més cost-efectiva. Això està donant peu a l'aparició de panells de múltiples gens per al diagnòstic genètic. Els panells permeten analitzar múltiples mutacions en múltiples gens a la vegada.

A dia d'avui ja s'han utilitzat panells de gens per al diagnòstic d'un gran nombre de condicions genètiques com cardiomiopaties, epilèpsia o distròfia muscular congènita, entre d'altres (Tarpey et al. 2009, Voelkerding et al. 2010, Lemke et al. 2012). Estudis com el del primer article d'aquesta tesi, confirmen les avantatges de la utilització de la NGS en el diagnòstic del càncer hereditari. Així, vàries cases comercials estan apostant per la generació de panells de múltiples gens associats al càncer (LaDuca et al. 2014).

El diagnòstic genètic a la nostra unitat està evolucionant cap a l'ús dels panells. Per aquest motiu el nostre equip va decidir avaluar el rendiment del panell comercial d'Illumina "*Trusight Cancer Panel*" (http://www.illumina.com/products/trusight_cancer.html) per tal de valorar la seva efectivitat en la rutina diagnòstica del Càncer Hereditari. Aquest panell inclou 94 gens associats a càncer hereditari més un grup de 284 SNPs correlacionats amb càncer en diferents estudis d'associació.

Mètodes

S'han seqüenciat els 94 gens i 284 SNPs que formen el Trusight Cancer Panel, en 24 pacients de famílies procedents de la Unitat de Consell Genètic. Deu mostres tenien una mutació coneguda (sèrie *Training)*, i catorze mostres eren de persones amb famílies amb una elevada incidència de càncer però sense mutació patogènica coneguda (sèrie *Discovery*). La generació de la llibreria es va fer amb la tecnologia Nextera que es basa en una captura amb transposomes (Nextera Rapid Capture) i un enriquiment amb sondes. La seqüenciació es va fer en un MiSeq (Illumina). La regió d'interès està formada per tots els exons codificants descrits, les zones intròniques adjacents (+/- 20 pb) en general estan cobertes però Illumina no dóna garanties de que arribin a la cobertura suficient per a una confiança acceptable per al diagnòstic (30x). A les taules 1 i 2 es troben els gens i SNPs que inclou el panell. Dels 94 gens, es van seleccionar un conjunt de 20 gens principals que s'analitzarien amb més cura, gens *Core*, aquest conjunt de gens principals es caracteritza per tenir implicacions clíniques en cas d'estar alterats, ja sigui prenent mesures de prevenció, de seguiment, o de tractament, i presenten risc de càncer amb una prevalença important a la nostra població. Concretament són 20 gens rellevants per al diagnòstic del Càncer de Mama i Ovari Hereditari (CMOH), el Càncer Colorectal Hereditari No Polipòsic (CCHNP) i la Poliposis Adenomatosa Familiar (FAP).

L'anàlisi bioinformàtica es va realitzar mitjançant dos protocols per comprovar-ne la robustesa. En primer lloc es va realitzar un control de qualitat de les lectures amb el FastQC. Posteriorment, d'una banda es van analitzar les dades mitjançant l'alineament i la detecció de variants seguint el protocol proposat per la casa comercial Illumina i disponibles al BaseSpace (Illumina BaseSpace 2016) ("BWA-MEM" per a l'alineament + "GATK" per a la detecció de variants + "Variant Studio" per a l'anotació de

les variants + "IGV" per a la visualització dels alineaments), amb els paràmetres per defecte i filtrant les variants amb cobertura menor de 30x i freqüència menor del 25%. D'altra banda, es van analitzar les dades amb el *software* comercial SeqNext (JSI), el SeqNext permet realitzar tots els passos de l'anàlisi. En aquest cas es van optimitzar els paràmetres de l'alineament per a que permetés detectar insercions i delecions majors de 30 nucleòtids. Per al filtratge de les variants, el propi *software* reporta un nivell de confiança per a cada variant indicant aquelles amb alta probabilitat de ser falsos positius, tot i així es va considerar tot el llistat de variants reportades.

Després de la crida general de variants es va realitzar un filtratge més acurat on es va tenir en compte el filtratge de les variants en zones intròniques, les variants de *PMS2* descartades per ser potencials falsos positius a causa dels pseudogens d'alta homologia, les variants presents en vàries mostres, classificant aquelles variants presents en tres o més mostres com a polimorfismes si estan presents al dbSNP, o com a falsos positius deguts a errors en la seqüenciació, i per últim, descartant aquelles variants situades en regions on l'alineament no és acurat.

Paral·lelament, amb els alineaments resultants de cadascun dels protocols d'anàlisi, es va analitzar la cobertura dels gens en totes les mostres, avaluant la capacitat del panell d'arribar a una cobertura de 30x en la regió d'interès, estimant la quantitat de fragments que quedarien baixos de cobertura i en seria necessària la seqüenciació Sanger.

Resultats

La qualitat de la seqüenciació va ser homogènia en totes los mostres, amb una mitjana del 45% de lectures alineades en la regió d'interès, i amb un 95% de les bases amb una qualitat de 30 o superior. A la Taula 3 es mostren els paràmetres principals de la qualitat de la seqüenciació amb el *Trusight Cancer Panel* d'Illumina per a les 24 mostres.

En relació a la cobertura, s'han estudiat en detall la selecció dels 20 gens *Core* (Taula 1). La majoria d'exons queden ben coberts, excepte alguns com els primers exons d'*MSH6* o *EPCAM,* que fallen en quasi totes les mostres i aproximacions. A la figura 1 s'observa com la majoria d'exons més 20 pb de regions intròniques adjacents queden ben coberts.

Pel que fa a la crida de variants, després dels filtres de cobertura (>30x) i freqüència al·lèlica (>25%) els dos protocols donaven un nombre molt alt i diferent de variants. Després d'un filtratge més acurat descartant regions intrònques més profundes, variants del gen *PMS2*, variants presents en més d'una mostra o variants classificades com a polimorfismes al dbSNP, es va reduïnt el nombre de variants progressivament tal com es mostra a la figura 2, fins a arribar al mateix nombre de variants i ser coincidents pels dos protocols. A la figura 3 s'il·lustra com es presenten els resultats de les variants i dels alineaments els dos protocols d'anàlisi utilitzats.

Cal destacar que totes les mutacions patogèniques conegudes en les deu mostres de la sèrie *Training* es van detectar mitjançant els dos protocols d'anàlisi, excepte una. La mutació no identificada es troba a l'exó 1 de *MSH6 i* no va ser cridada ja que es localitzava en una de les regions amb baixa cobertura citades anteriorment. Així, aquesta mutació s'hagués cridat posteriorment en la seqüenciació per Sanger

de les regions amb baixa cobertura. A la taula 3 es mostren de forma resumida les variants conegudes i si es van trobar o no després de l'anàlisi bioinformàtica per ambdós protocols.

En l'estudi de les mostres *Discovery* es van trobar 88 variants candidates, i entre elles 6 es podien classificar com a potencialment patogèniques, a la Taula 5 es descriuen les mutacions i els gens on es troben.

Discussió

Els dos protocols d'anàlisi bioinformàtica utilitzats són eficients i robustos per donar els resultats de forma acurada. Una de les diferències entre els dos protocols és la seqüència de referència que utilitzen per alinear, l'opció que s'ha utilitzat en el SeqNext només considera la regió d'interès, de manera que totes les lectures s'alineen contra els gens seqüenciats, en canvi el BWA-MEM d'Illumina considera tot el genoma humà com a referència. Aquest fet sembla ser la causa de que el nombre de variants a *PMS2* detectades pel SeqNext sigui molt superior, ja que lectures que pertanyin a altres pseudogens s'alineen sobre la regió d'interès de *PMS2* indicant variants falses, i el mateix succeeix amb altres regions, incrementant el nombre de falsos positius. El SeqNext té una opció més avançada que permet considerar les seqüències de pseudogens però per implementar aquesta funció es requereix una major formació de la que disposàvem. Una altra de les diferències entre els dos protocols és la classificació de les variants *a priori*, el SeqNext classifica les variants amb un indicador del nivell de confiança de ser falsos positius. En aquesta anàlisi es van considerar totes les variants reportades tot i que el classificador indiqués alta probabilitat de fals positiu, la majoria d'aquestes variants són posteriorment descartades per estar en 3 o més mostres i no ser polimorfismes. La petita variació en el nombre de polimorfismes és deguda a les diferents bases de dades que utilitzen el SeqNext o el VariantStudio per a la classificació de les variants.

En definitiva, el panell Trusight Cancer d'Illumina mostra un bon rendiment en la rutina del diagnòstic de càncer hereditari. S'està realitzant un estudi sobre el rendiment d'un altre panell amb disseny propi per comparar-ne els resultats. També s'estan estudiant les avantatges i els inconvenients que suposaria seqüenciar tot l'exoma en la rutina del diagnòstic genètic, considerant els costos, les connotacions ètiques i el coneixement genètic.

Taula 1. Els 94 gens inclosos en el panell Trusight Cancer destacant en taronja els 20 *Core* per al diagnòstic del Càncer de Mama i Ovari Hereditari (CMOH), el Càncer Colorectal Hereditari No Polipòsic (CCHNP) i la Poliposis Adenomatosa Familiar (FAP).

BRCA1	*AIP*	*ERCC4*	*HRAS*	*SDHB*
BRCA2	*ALK*	*ERCC5*	*KIT*	*SDHC*
MLH1	*ATM*	*EXT1*	*MAX*	*SDHD*
MSH2	*BAP1*	*EXT2*	*MEN1*	*SLX4*
MSH6	*BLM*	*EZH2*	*MET*	*SMARCB1*
PMS2	*BRIP1*	*FANCA*	*NBN*	*SUFU*
EPCAM	*BUB1B*	*FANCB*	*NF1*	*TMEM127*
APC	*CDC73*	*FANCC*	*NF2*	*TSC1*
MUTYH	*CDK4*	*FANCD2*	*NSD1*	*TSC2*
TP53	*CDKN1C*	*FANCE*	*PHOX2B*	*VHL*
PMS1	*CDKN2A*	*FANCF*	*PRF1*	*WRN*
PTEN	*CEBPA*	*FANCG*	*PRKAR1A*	*WT1*
STK11	*CEP57*	*FANCI*	*PTCH1*	*XPA*
CDH1	*CYLD*	*FANCL*	*RB1*	*XPC*
BMPR1A	*DDB2*	*FANCM*	*RECQL4*	
SMAD4	*DICER1*	*FH*	*RET*	
PALB2	*DIS3L2*	*FLCN*	*RHBDF2*	
CHEK2	*EGFR*	*GATA2*	*RUNX1*	
RAD51C	*ERCC2*	*GPC3*	*SBDS*	
RAD51D	*ERCC3*	*HNF1A*	*SDHAF2*	

Taula 2. Llistat dels 284 SNPs associats amb risc de càncer inclosos en el panell TrusightCancer.

284 SNPs					
rs17401966	rs7584330	rs971074	rs4324798	rs2180341	rs7014346
rs9430161	rs2292884	rs7679673	rs29232	rs9485372	rs1447295
rs7538876	rs757978	rs10069690	rs3129055	rs2046210	rs4242382
rs11249433	rs4973768	rs2242652	rs2860580	rs651164	rs4242384
rs7412746	rs1052501	rs2736100	rs2517713	rs9364554	rs7837688
rs3790844	rs2660753	rs2853676	rs6457327	rs7758229	rs9642880
rs6691170	rs9284813	rs4635969	rs130067	rs4487645	rs2019960
rs6687758	rs17181170	rs4975616	rs2894207	rs11978267	rs10088218
rs801114	rs9841504	rs401681	rs2596542	rs4132601	rs891835
rs1465618	rs10934853	rs31489	rs2248462	rs6465657	rs4295627
rs7579899	rs6763931	rs12653946	rs3117582	rs1495741	rs2294008
rs1432295	rs6774494	rs2255280	rs204999	rs1512268	rs7040024
rs721048	rs10936599	rs13361707	rs9268542	rs2439302	rs755383
rs10187424	rs10936632	rs2121875	rs6903608	rs16892766	rs3814113
rs17483466	rs4488809	rs4415084	rs2395185	rs1016343	rs7023329
rs12621278	rs10937405	rs889312	rs2858870	rs1456315	rs2157719
rs2072590	rs17505102	rs10052657	rs674313	rs16901979	rs1412829
rs13016963	rs710521	rs20541	rs28421666	rs2456449	rs1011970
rs13393577	rs2131877	rs4624820	rs2647012	rs16902094	rs4977756
rs3768716	rs798766	rs10058728	rs10484561	rs445114	rs965513
rs6435862	rs1494961	rs872071	rs9275572	rs13281615	rs865686
rs13387042	rs12500426	rs12210050	rs210138	rs1562430	rs505922
rs966423	rs17021918	rs4712653	rs10484761	rs10505477	rs10795668
rs13397985	rs1229984	rs6939340	rs339331	rs6983267	rs11012732
rs3123078	rs10896449	rs9510787	rs4785763	rs738722	rs7584993
rs10993994	rs7130881	rs753955	rs4795519	rs36600	rs17272796
rs10821936	rs7105934	rs9600079	rs4430796	rs2284063	rs1155741
rs7089424	rs614367	rs9573163	rs7501939	rs1014971	rs161792
rs10822013	rs1393350	rs9543325	rs7210100	rs5759167	rs11940551
rs10995190	rs1801516	rs7335046	rs1859962	rs5768709	rs9293511
rs224278	rs3802842	rs944289	rs17674580	rs1327301	rs9352613
rs704010	rs498872	rs116909374	rs7238033	rs5945572	rs685449
rs3765524	rs735665	rs4444235	rs4939827	rs5945619	rs7808249
rs2274223	rs2900333	rs4779584	rs8170	rs5919432	rs1106334
rs3781264	rs718314	rs4924410	rs8102137	rs1321311	rs11017876
rs17119461	rs10875943	rs4775302	rs10411210	rs3824999	rs9572094
rs12413624	rs11169552	rs8030672	rs8102476	rs5934683	rs4905366
rs11199874	rs902774	rs7176508	rs11083846	rs2283873	rs4775699
rs2981579	rs995030	rs8034191	rs2735839	rs807624	rs1528601
rs2981575	rs3782181	rs1051730	rs961253	rs1027643	rs11655512
rs1219648	rs4474514	rs8042374	rs910873	rs3755132	rs4793172
rs2981582	rs11066015	rs3803662	rs4925386	rs790356	rs242076
rs3817198	rs671	rs4784227	rs6010620	rs5955543	rs6603251
rs7127900	rs4767364	rs3112612	rs4809324	rs10974944	AMG_mid100
rs110419	rs2074356	rs9929218	rs372883	rs1210110	rs149617956
rs1945213	rs11066280	rs391525	rs2014300	rs7555566	ATM_SNP
rs11228565	rs4765623	rs258322	rs45430	rs1364054	rs138213197
rs7931342	rs1572072	rs1805007	rs1547374	rs6734275	

Taula 3. Paràmetres de qualitat de la seqüenciació

Qualitat de la seqüenciació			
Mostra	Número de lectures	Percentatge de lectures alineades	Percentatge de bases amb qualitat >30
9508	1078238	45,31	95,16
9249	567086	46,93	94,38
10737	95954	45,35	94,82
7359	1021616	46,30	94,80
7272	114395	46,59	95,50
14010	1146242	45,16	93,32
12111	1057048	45,49	94,60
12856	1080798	44,48	94,89
11497	102991	46,02	93,30
11147	106374	44,67	94,84
5097	1252286	45,75	94,91
5029	2087782	44,85	95,52
15302	165769	44,23	94,51
12051	709374	46,88	94,54
13177	1117538	43,19	95,00
11978	73146	46,81	94,07
3202	1256498	43,30	94,56
14290	1038974	38,91	94,74
7508	730242	46,50	94,62
14560	717384	45,40	94,61
9569	1395808	45,49	93,82
15263	1107138	45,09	95,19
8449	105105	45,23	95,01
14290	1038974	38,91	94,74
Mitjana	**798615**	**44,87**	**94,64**

Taula 4. Variants conegudes trobades segons els dos protocols d'anàlisi.

Mostra	Mutació coneguda	SeqNext	BWA-GATK-*V.Studio*
9249	*APC* c.2344A>T	Sí	Sí
9249	*APC* c.730-22G>C	Sí	Sí
10737	*MSH6* c.255dupC	No	No
11147	*APC* c.1548+1G>C	Sí	Sí
11147	*APC* c.1744-37A>G	Sí	Sí
11497	*MLH1* c.1590_1598dupCGTGGGCTG	Sí	Sí
12111	*MUTYH* c.1187G>A	Sí	Sí
5029	*BRCA1* c.1961delA	Sí	Sí
5097	*BRCA1* c.1953_1956delGAAA	Sí	Sí
7359	*BRCA2* c.8946delA	Sí	Sí
9508	*BRCA1* c.68_69delAG	Sí	Sí
14010	*BRCA1* c.3869_3870delAA	Sí	Sí
14010	*BRCA1* c.2584A>G	Sí	Sí

Taula 5. Variants potencialment patogèniques trobades a les mostres de la sèrie *Discovery*.

Gene	Mutation
ATM	c.4776+2_4776+13delTAATAAAAATTT
CHEK2	c.792+2T>C
ERCC3	c.325C>T
FANCL	c.1111_1114dupATTA
FANCM	c.5791C>T
MSH2	c.2785C>T

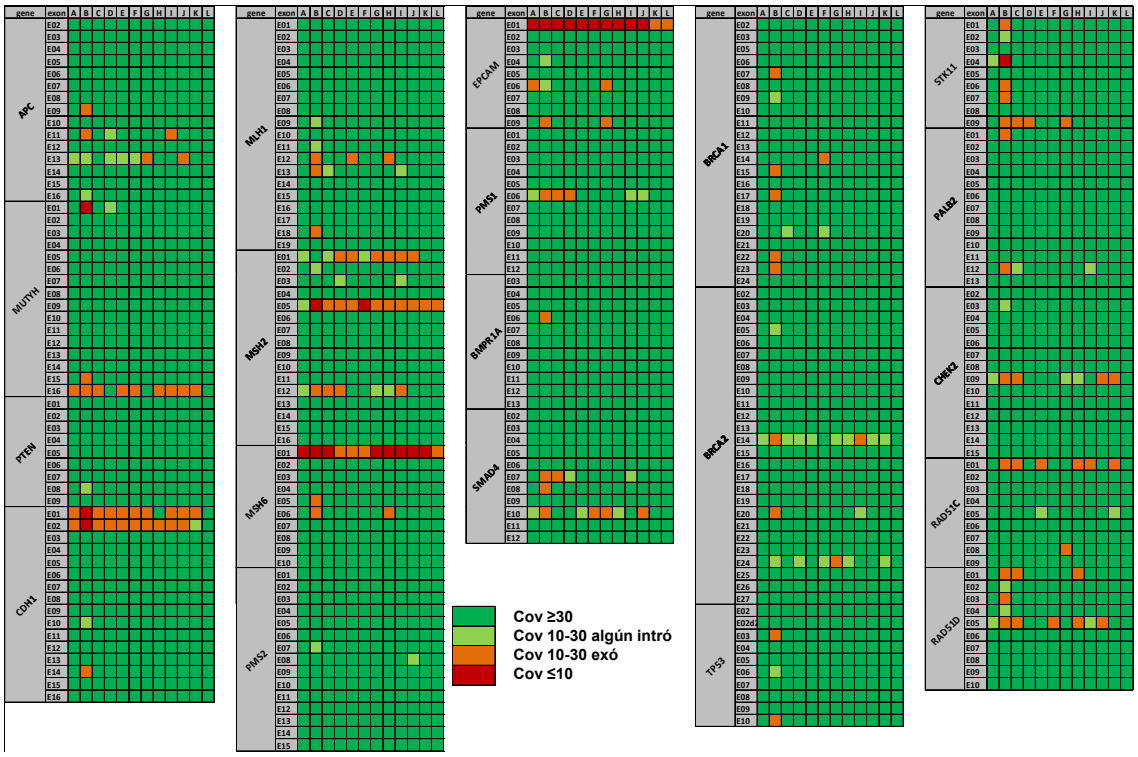


Figura 1. Cobertura dels exons més les 20 bases intròniques flanquejants per als 20 gens *Core* en el diagnòstic de càncer hereditari en 12 mostres, 10 mostres de la sèrie *Training* i dues de la sèrie *Discovery*.

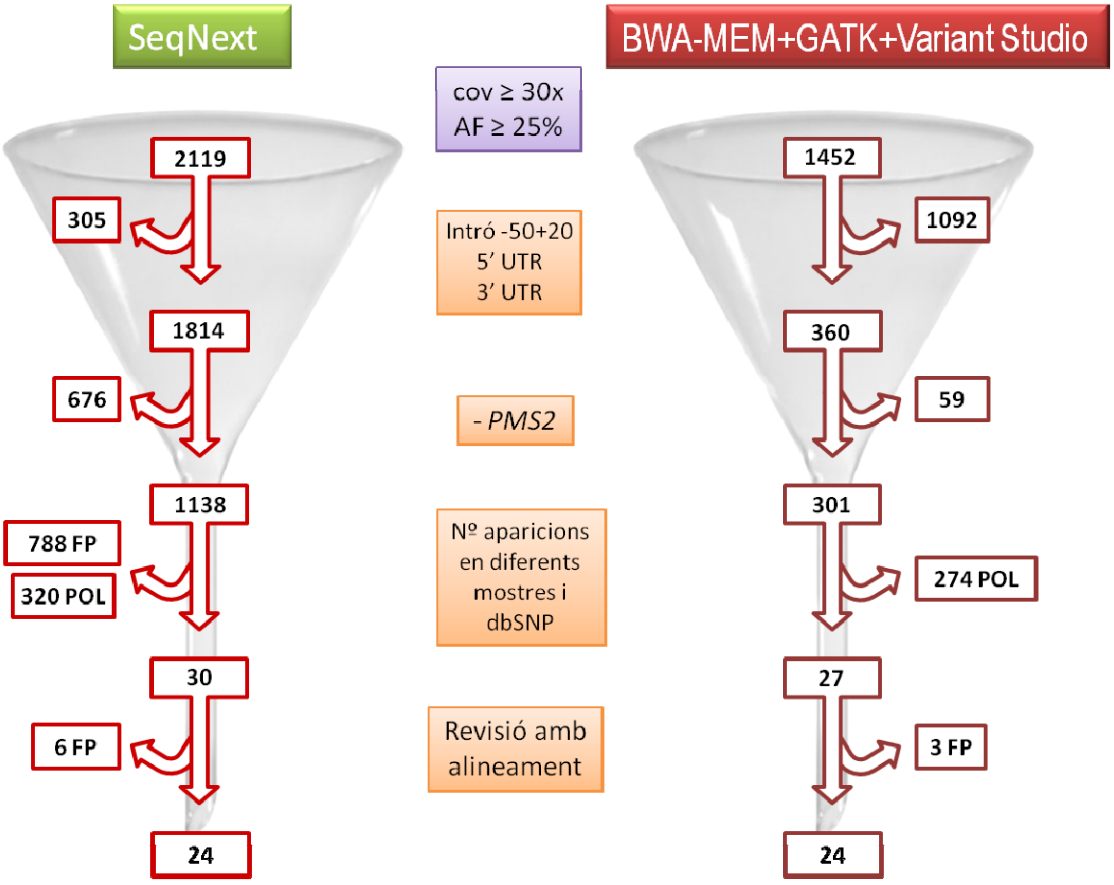
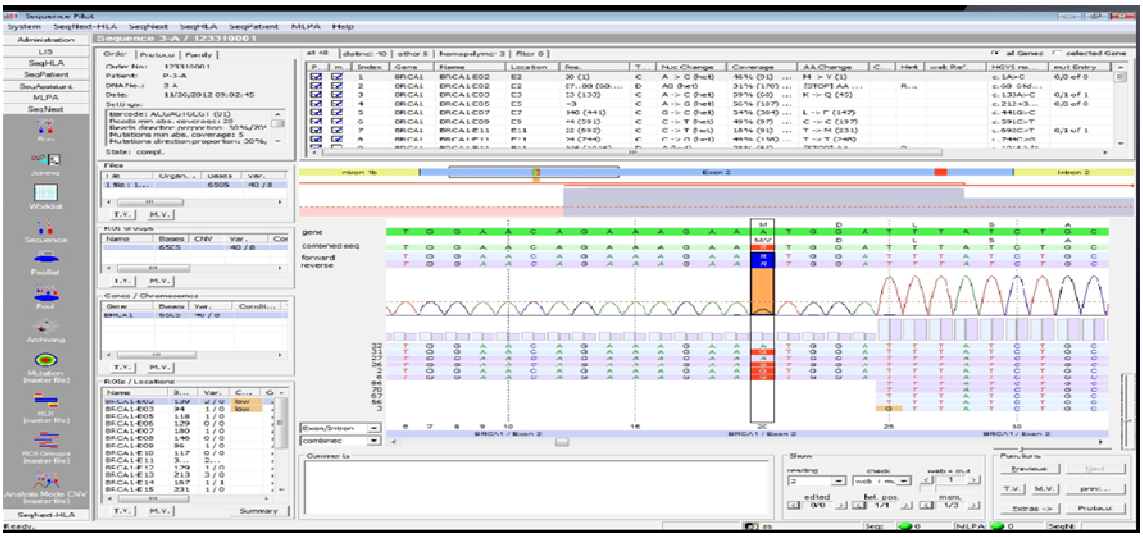


Figura 2. Filtratge de variants en els 20 gens *Core* de la sèrie *Training* mitjançant els 2 protocols d'anàlisi

SeqNext



BWA-MEM + GATK + VS + IGV

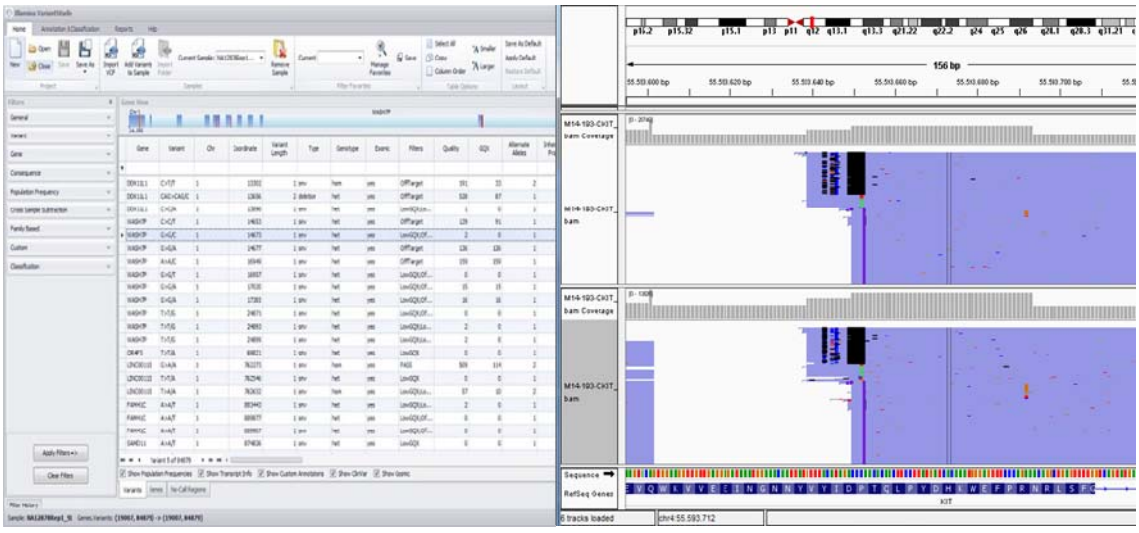


Figura 3. Presentació dels resultats mitjançant els protocols d'anàlisi basats en SeqNext i en el conjunt d'aplicacions del BaseSpace (BWA-MEM+GATK+VS+IGV).

NGS per a la recerca del CCR esporàdic

ARTICLE 3

Exome sequencing reveals *AMER1* as a frequently mutated gene in colorectal cancer

Rebeca Sanz-Pamplona, **Adriana Lopez-Doriga**, Laia Paré-Brunet, Kira Lázaro, Fernando Bellido, M. Henar Alonso, Susanna Aussó, Elisabet Guinó, Sergi Beltrán, Francesc Castro-Giner, Marta Gut, Xavier Sanjuan, Adria Closa, David Cordero, Francisco D. Morón-Duran, Antonio Soriano, Ramón Salazar, Laura Valle, Victor Moreno

Resum del treball: En la progressió del Càncer Colorectal (CCR), les mutacions somàtiques apareixen en els estadiatges inicials dels adenomes i es van acumulant a mesura que progressa cap a CCR. L'objectiu d'aquest treball és caracteritzar l'estat mutacional dels tumors estadiatge II i trobar noves mutacions recurrents que puguin estar implicades en la tumorigenesis del CCR.

Es va seqüenciar l'exoma de 42 mostres de tumors de CCR estadi II i amb estabilitat de microsatèl·lits, així com de les mucoses normals aparellades. A més del DNA obtingut, altres dades moleculars disponibles en aquest grup de mostres (expressió gènica, metilació i CNV) també es van utilitzar per tal de caracteritzar aquests tumors. Per a la validació de les troballes mutacionals es van utilitzar les dades d'un conjunt de 553 mostres adicionals de CCR.

Com a resultat, es van trobar 4886 SNV somàtiques. La gran majoria de les variants eren úniques, només unes poques eren compartides per més d'un tumor, el que reflecteix un panorama mutacional específic de cada tumor. Tot i això, aquesta diversitat de mutacions convergia en vies metabòliques comunes com el cicle cel·lular o l'apoptosi. Enmig d'aquesta heterogeneïtat mutacional, ressaltaven les variants truncants al gen *AMER1* (també anomenat *FAM123* o *WTX*) que apareixien de forma recurrent en el casos de CCR. A més, també es van trobar pèrdues del gen *AMER1* per altres mecanismes com són la metilació i les CNV. Els tumors amb deficiència d'aquest gen supressor mostraren un fenotip mesenquimal caracteritzat per la inhibició de la via metabòlica de Wnt.

Validacions *in silico* i experimentals en grups de dades independents confirmaven l'existència de mutacions amb alta probabilitat d'afectar la funció en *AMER1* en aproximadament el 10% dels tumors de CCR analitzats. A més, aquests tumors mostraven un fenotip característic.

Contribució de la doctoranda: En aquest treball la doctoranda ha realitzat les anàlisis bioinformàtiques de les dades de seqüenciació per a totes les mostres de l'estudi, els 84 exomes de la sèrie principal, i els 239 de la sèrie de validació del TCGA. Ha realitzat un estudi minuciós dels filtres de les variants per a aconseguir màxima sensibilitat i especificitat. Ha realitzat les anàlisis estadístiques amb R. Ha participat en la redacció de l'article i en la preparació de taules i figures.

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

Clinical Cancer Research

Exome Sequencing Reveals *AMER1* as a Frequently Mutated Gene in Colorectal Cancer

Rebeca Sanz-Pamplona[1], Adriana Lopez-Doriga[1], Laia Paré-Brunet[1], Kira Lázaro[1], Fernando Bellido[2], M. Henar Alonso[1], Susanna Aussó[1], Elisabet Guinó[1], Sergi Beltrán[3], Francesc Castro-Giner[3], Marta Gut[3], Xavier Sanjuan[4], Adria Closa[1], David Cordero[1], Francisco D. Morón-Duran[1], Antonio Soriano[5], Ramón Salazar[6,7], Laura Valle[2], and Victor Moreno[1,8]

Abstract

Purpose: Somatic mutations occur at early stages of adenoma and accumulate throughout colorectal cancer progression. The aim of this study was to characterize the mutational landscape of stage II tumors and to search for novel recurrent mutations likely implicated in colorectal cancer tumorigenesis.

Experimental Design: The exomic DNA of 42 stage II, microsatellite-stable colon tumors and their paired mucosae were sequenced. Other molecular data available in the discovery dataset [gene expression, methylation, and copy number variations (CNV)] were used to further characterize these tumors. Additional datasets comprising 553 colorectal cancer samples were used to validate the discovered mutations.

Results: As a result, 4,886 somatic single-nucleotide variants (SNV) were found. Almost all SNVs were private changes, with few mutations shared by more than one tumor, thus revealing tumor-specific mutational landscapes. Nevertheless, these diverse mutations converged into common cellular pathways, such as cell cycle or apoptosis. Among this mutational heterogeneity, variants resulting in early stop codons in the *AMER1* (also known as *FAM123B* or *WTX*) gene emerged as recurrent mutations in colorectal cancer. Losses of *AMER1* by other mechanisms apart from mutations such as methylation and copy number aberrations were also found. Tumors lacking this tumor suppressor gene exhibited a mesenchymal phenotype characterized by inhibition of the canonical Wnt pathway.

Conclusion: *In silico* and experimental validation in independent datasets confirmed the existence of functional mutations in *AMER1* in approximately 10% of analyzed colorectal cancer tumors. Moreover, these tumors exhibited a characteristic phenotype. *Clin Cancer Res; 1–10. ©2015 AACR.*

Introduction

Colorectal cancer is the third most common cancer and the second leading cause of cancer death in the world (1). The classic adenoma-to-carcinoma model postulates that colorectal cancer tumorigenesis proceeds through a progressive accumulation of genetic alterations in oncogenes and tumor suppressors genes (2). However, colorectal cancer is currently considered a heterogeneous disease. While tumors fitting into the classic progression model (or chromosomal instability model, CIN) are the most frequent, other tumor phenotypes have been described, such as microsatellite instability (MSI) and CpG island methylator phenotypes (CIMP; ref. 3). Recent studies based on high-throughput technologies have addressed the issue of colorectal cancer molecular complexity, revealing high level of heterogeneity among tumors (4).

Among other biologic mechanisms, it is widely accepted that somatic mutations lead to tumor development in colorectal cancer. It is postulated that most mutations within a tumor are undamaging byproducts of tumorigenesis (passenger mutations) whereas only a few are responsible for driving the initiation and progression of the tumor (driver mutations; ref. 5). In colorectal cancer, a number of mutations have been proposed as drivers, such as those in the *KRAS* and *BRAF* oncogenes, or in the tumor suppressor genes *APC* and *TP53* (6). However, the seminal study by Wood and colleagues revealed that the mutational landscapes of colorectal cancer genomes are composed of a few frequently mutated genes across patients, "mountains," but are dominated by a much larger number of infrequently mutated genes, "hills" (7). Although still controversial, these rarely mutated genes may also contribute to tumor development, thus accounting for intertumor variability (8).

[1]Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain. [2]Hereditary Cancer Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [3]Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain. [4]Pathology Service, University Hospital Bellvitge (HUB–IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [5]Gastroenterology Service, University Hospital Bellvitge (HUB–IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [6]Department of Medical Oncology, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [7]Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [8]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Victor Moreno, Catalan Institute of Oncology, Avenue Gran Vía 199-203. 08908, L'Hospitalet de Llobregat, Barcelona 08908, Spain. Phone: 34-932607186; Fax: 34932607188; E-mail: v.moreno@iconcologia.net

doi: 10.1158/1078-0432.CCR-15-0159

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

Translational Relevance

Exome sequencing analysis in colorectal cancer samples reveals that variants resulting in stop codons in *AMER1* (also known as *FAM123B* or *WTX*) gene appeared in approximately 10% of the analyzed tumors. Moreover, although less commonly, *AMER1* function may also be lost by other mechanisms different from mutations such as promoter hypermethylation and chromosome deletions. The subset of tumors lacking *AMER1* expression showed Wnt pathway inhibition and, regarding molecular subtyping, could belong to type-C tumors. These results may enlighten about the mechanisms of carcinogenesis and biomarker discovery in those patients lacking *AMER1*.

Next-generation sequencing technologies have revolutionized cancer genomics research by providing fast and accurate information about individual tumors, bringing us closer to personalized medicine (9). It has been reported that approximately 85% of cancer-associated mutations are located in protein-coding regions (10). In consequence, exome sequencing has been revealed as a useful technique for mutation discovery in cancer tissues. Indeed, several studies have successfully described the mutational background of different types of tumors by using this approach (11, 12). Here, we have performed an exome sequencing analysis aimed to explore the somatic genomic landscape of microsatellite-stable (MSS) stage II colorectal tumors.

Materials and Methods

Patients and samples

This study included a subset of 42 paired adjacent normal and tumor tissues (84 samples) from a previously described set of 100 patients with colon cancer diagnosed at stage II and MSS tumors (ref. 13; colonomics project, CLX-: www.colonomics.org; NCBI BioProject PRJNA188510; Supplementary Table S1). All patients were recruited at the Bellvitge University Hospital (Barcelona, Spain). Written informed consent was obtained from all patients and the Institution's Ethics Committee approved the protocol. Prior to DNA extraction, purity of the sample was assessed by a pathologist to ensure that at least 80% was tumoral. DNA was extracted using a standard phenol–chloroform protocol. To ensure that adjacent and tumor tissues were paired, dynamic arrays were used to genotype 13 SNPs in the 84 samples (see Validation of *KRAS* and *TP53* point mutations). All 42 adjacent normal tissues correctly matched with their corresponding tumor (Supplementary Fig. S1). Tumor DNA from an additional series of 227 colorectal cancer patients from the same hospital was used for validation purposes (Supplementary Table S1). This extended series was not restricted regarding site, stage, and MSI phenotype.

In addition, raw exome sequencing data from 513 samples were downloaded from The Cancer Genome Atlas (TCGA) repository. TCGA discovery dataset comprised 239 colorectal cancer tumors and 100 adjacent mucosae and was used to expand the exome sequencing analysis. These are public samples available in TCGA repository but had not been used in the published work characterizing colorectal cancer exomes (14). Moreover, 87 matched nontumoral and tumoral colorectal samples, herein named TCGA validation dataset, were used as a validation cohort for *AMER1* mutations (Supplementary Table S1). These second set of samples included 44 that already had been analyzed by the TCGA consortium (14), not all of available samples because we requested a paired germline sample to ensure that mutations were somatic. Finally, 224 tumors from the TCGA published work with suitable information about molecular characteristic of the samples were used to assess the relationship between *AMER1* mutations and colorectal cancer molecular subtypes (MSS and CIMP status; ref. 14).

Exome sequencing pipeline

Genomic DNA from the set of 42 adjacent tumor paired samples was sequenced in the National Center of Genomic Analysis (Barcelona, Spain; CNAG) using the Illumina HiSeq-2000 platform. Exome capture was performed with the commercial kit Sure Select XT Human All Exon 50MB (Agilent). Tumor exomes were sequenced at 60× coverage (2 × 75 bp reads), and exomes from adjacent tissues were sequenced at 40× (2 × 75 bp reads). FastQ software was used to assess the quality of the sequences (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Bowtie 2.0 software was used to align sequences over the human reference genome HG19 (15). To refine data, reads unmapped, reads with unmapped mate, nonprimary alignments, and reads that were PCR or optical duplicates were discarded (http://picard.sourceforge.net/). We also executed a local realignment around indels defined in dbSNP (16) and 1000G (17) and also for the indels detected in this particular study. Variant calling was executed with GATK software, and low-quality variants (mapping quality below 30, read depth below 10 or frequency < 10%) were discarded (18). GATK has been proved to achieve higher sensitivity and specificity in exome variant calling than other softwares (19). Germline variants were also removed, that is, variants that were present in normal adjacent paired sequence for each tumor and variants reported in the 1000G project. Because of the high frequency of indels that were later considered as false-positive results, only single-nucleotide variants (SNV) were taken into account in this study. Finally, variants were annotated using the SeattleSeq Variant Annotation web tool (20). The same pipeline was applied to analyze exome data downloaded from TCGA repository. No correlation was observed between the number of mutations found per sample and the quality parameters "number of reads," "number of no matched reads," "percentage of unique aligned reads," "number of duplicate reads," and "coverage," reinforcing the robustness of the analytic pipeline (Supplementary Fig. S2).

Raw exome sequencing data were also used to search for copy number variations (CNV). Coverage data were used to compare the amount of DNA in adjacent versus tumor samples. The Varscan2 *copynumber* algorithm was run on adjacent and tumor mpileup obtained from Samtools (21). Next, the *copyCaller* command was run to adjust raw copy number values for GC content. Finally, R-GADA package was used to perform the segmentation analysis (22). A region was considered lost if the $\log_2$ tumor to adjacent mucosa ratio was less than −0.5.

Functional and pathways analysis

Databases containing function and pathway information "KEGG," "Biocarta," "Reactome," and "GO" were downloaded from MSigDB in gene set enrichment analysis (GSEA; ref. 23). Only the potentially functional SNVs, that is, coding nonsynonymous, stop gain, stop lost, splice-5′, splice-3′, coding

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

synonymous near splice site, 3′-untranslated region (UTR), and 5′-UTR variants, were analyzed. For each gene set in each database, a score was calculated by dividing the number of mutations mapping into genes constituting the dataset by the number of genes in such dataset. The score was corrected by dividing the number of samples and multiplying by 100. A *P* value was calculated by randomly permuting the original matrix of SNVs. The number of permutations was calculated in each case to ensure that the minimum *P* value was at least as small as required by the Bonferroni correction at nominal 0.05 significance level.

Validation of *KRAS* and *TP53* point mutations

KASPar genotyping assays (KASP-By-Design; LGC Group) on the Fluidigm genotyping platform (48.48 Dynamic Array IFG, Fluidigm) were used to validate 6 mutations in *KRAS* and 8 mutations in *TP53*. The same methodology was applied to genotype 13 SNPs used to ensure that adjacent and tumor tissues were matched (Supplementary Table S2). Each genotyping assay was previously validated and optimized on the Light-Cycler 480 real-time PCR detection system (Roche Diagnostics GmbH).

Functional prediction of mutations

MutSig software was used to identify the more likely cancer-associated genes from other less suspicious genes. Mutated genes were ranked using 3 criteria: (i) abundance of mutations relative to the background mutation rate; (ii) clustering of mutations in hotspots within the gene; and (iii) evolutionary conservation of the mutated positions (24). In addition, protein damage predictions of missense mutations in *AMER1* were performed by using the *in silico* algorithms SIFT (25), PolyPhen-2 (26), and PMut (27). Possible alterations of the protein structure were evaluated using the Hope software (28).

Sanger sequencing

Sanger sequencing was used to perform a technical validation of exome sequencing. Mutations in *KRAS*, *APC*, and *TP53* genes were sequenced using a standard protocol. Sanger sequencing was also used to validate the recurrent mutations found in *AMER1*. Two regions in exon 1 (covering 335 and 236, respectively) were sequenced. Sequencing was performed on an ABI Sequencer 3730 and data analyzed using Mutation Surveyor v.3.10. Primer sequences are shown in Supplementary Table S2.

Expression data

Gene expression data from GSE44076 dataset (deposited in GEO repository), which includes the 42 sequenced tumors, were used to search for phenotypic similarities among tumors exhibiting loss of *AMER1* functionality. "Sub-type B score" and "Sub-type C score" were calculated for each tumor as the mean of absolute expression of those genes described in Roepman and colleagues (29) as B-type characteristic (53 genes) or C-type characteristic (102 genes). Supplementary Fig. S3 showed all molecular data used in this study.

AMER1 methylation and CIMP phenotype assessment

Tumor methylation in our discovery dataset (CLX data) was analyzed with the Illumina Infinium HumanMethylation450 BeadChip assay covering approximately 20,000 genes (99% of RefSeq genes). *AMER1* promoter methylation was extracted from this large dataset. Also, this information has been used to asses CIMP phenotype by interrogating *MLH1*, *RUNX3*, *CACNA1G*, *IGF2*, *NEUROG1*, *SOCS1*, *CRABP1*, and *CDKN2A* promoters, as previously reported (30). A tumor was considered to be CIMP-high (CIMP-H) if at least 6 of these 8 genes were hypermethylated.

Immunohistochemistry

Xylene-dewaxed paraffin tissue sections (4-μm thick) were obtained from 4 *AMER1*-mutated tumors and from 3 wild-type AMER1 tumors. For antigen retrieval, the slides were boiled after deparaffinization in a pressure cooker for 2 minutes in sodium citrate buffer (10 mmol/L sodium citrate, 2 mmol/L citric acid, pH = 6). Endogen peroxidase was blocked by sample immersion in 3% H_2O_2 during 15 minutes. Blocking was carried out by applying goat serum 1:10 diluted in PBS for 60 minutes at room temperature (RT). Subsequently, the primary antibody against AMER1 (OAAB03558, Aviva Systems Biology, diluted 1:100 in blocking solution) was added and incubated overnight at 4°C in a humidified chamber. After rinsing, EnVision system-Goat secondary antibody (Dako) was applied for 60 minutes at RT and subsequently revealed with DAB substrate (Dako) exposed for 4 minutes. Slides were counterstained with hematoxylin.

Results

Somatic mutational landscape in stage II colon tumors

Exome analysis revealed a total of 11,122 somatic SNVs within the 42 analyzed tumors (Supplementary Table S3). Many were intergenic and intronic mutations (most of them likely to be false-positives due to the lower coverage). Indeed, approximately 50% of SNVs were located in intronic regions (away from canonical splice sites). As expected, all the tumors showed enrichment in C:G > T:A nucleotide changes (Fig. 1A). There was no concordance between the number of mutations and the age of the patients, even if only C>T nucleotide changes were taken into account (Supplementary Fig. S4). The number of SNVs identified in coding regions and flanking sequences was 4,725. The average number of mutations in coding regions per sample was 117 (range, 6–185; Fig. 1B). From those, 9.6% were coding synonymous, thus *a priori* not affecting the protein structure. Considering only potentially functional mutations, 22.4% of SNVs were missense, 5.9% in UTR, 2.1% stop gain or stop lost, and 0.7% splice-site variants (Supplementary Fig. S5).

Remarkably, the vast majority of the identified somatic SNVs (10,985 of 11,122 total and 4,699 of 4,725 located in coding regions—more than 99%-) were private events, whereas only 137 SNVs were shared by two or more tumors. Of those, 112 were intronic or intergenic. As expected, the *KRAS* G12D mutation was the most recurrent change, occurring in 8 of 42 samples (19%; Supplementary Table S4). As a methodologic validation of the overall discovery pipeline of SNVs, 6 mutations in *KRAS* (Q61H, A146T, G12V, G12D, G12S, G13D) and 7 in *TP53* (G245D, R248Q, R237H, R273C, R175H, R282W, R213*, G245S) were tested using KASPar genotyping assays in the Fluidigm Biomark platform (dynamic arrays), achieving 65% concordance. Of note, 10 of 11 nonconcordant mutations were only found by exome sequencing confirming the better performance and higher sensitivity of this technique (Supplementary Table S5). To further

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

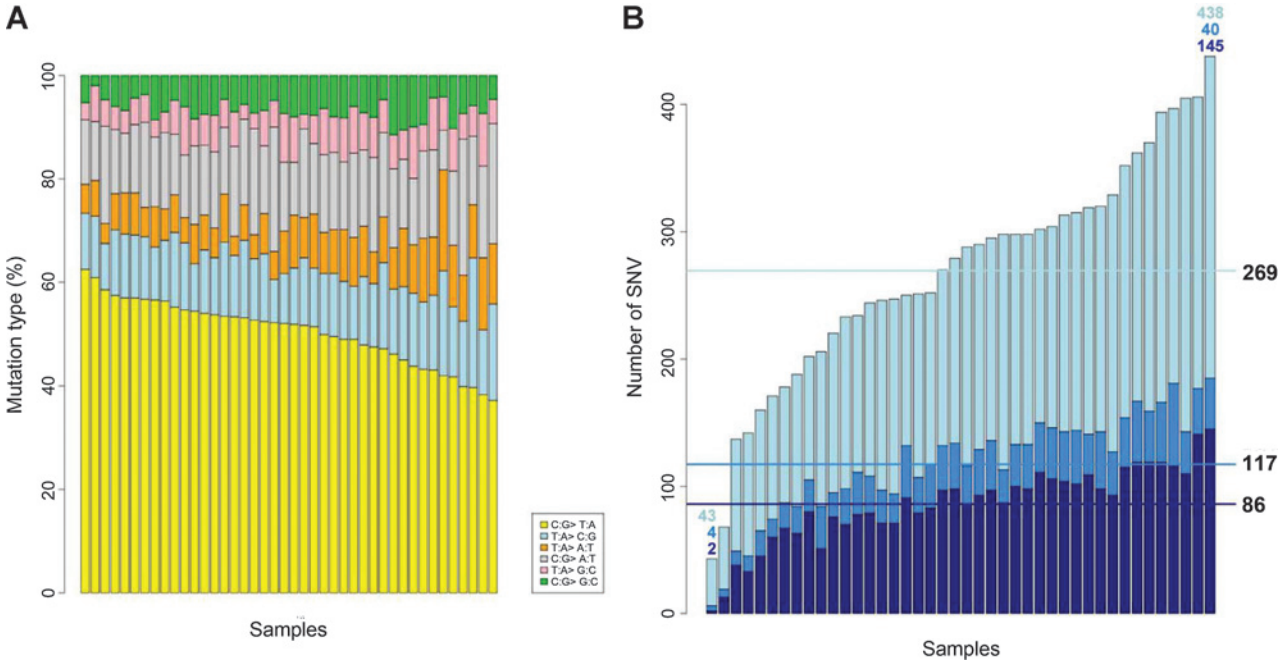


Figure 1.
Single-nucleotide somatic mutations (SNV) across samples. A, percentage of mutations by transition/transversion type. B, number of mutations per sample. Light blue represents the fraction of intronic/intergenic mutations, whereas blue represents exonic variants. Dark blue emphasizes those potentially functional exonic variants. Horizontal lines represent the mean of SNV for each category.

validate the sensitivity of our mutation calling pipeline, 9 point mutations (including 4 not previously validated by dynamic arrays) in *APC* (1), *KRAS* (4), and *TP53* (4) were validated by Sanger sequencing (the gold standard technique) in all the CLX tumor samples, achieving a 100% concordance (Supplementary Table S5).

After removal of intergenic SNVs ($n = 962$), the remaining 10,160 variants were located in 6,433 genes and 174 of them mapped to more than one gene. Across samples, a total of 723 genes were mutated in more than one tumor. *TTN* (the largest gene in the human genome) was the most frequently mutated, followed by *APC*, *KRAS*, and *TP53*. If only potentially functional mutations were taken into account, *APC* (22 tumors), *KRAS* (21 tumors), and *TP53* (20 tumors) were the most mutated genes (Fig. 2). MutSig software was also used to rank mutated genes on the basis of recurrence and functional effect of mutations. As expected, spurious genes (like the well-known *TTN*) noticeably went down in the list whereas *APC*, *KRAS*, and *TP53* continued standing out (Supplementary Table S6). These findings were in agreement with previous studies performed to discover mutated genes in colorectal cancer (Supplementary Fig. S6; refs. 14, 31, 32).

Next, a pathway analysis was performed, including 2,856 genes harboring 3,595 potentially functional SNVs. SNVs with putative functional impact accumulated in pathways and functions classically related to cancer such as cell cycle ($P < 0.001$), apoptosis ($P < 0.001$), or cell signaling ($P < 0.001$). Moreover, pathways and functions specifically related to colorectal cancer tumors, such as the Wnt pathway ($P < 0.001$), the NOTCH expression and translation ($P < 0.001$), the VEGF pathway ($P < 0.001$), or the TGFB pathway ($P < 0.001$); among others, also appeared enriched in mutated genes. The complete list of statistically significant functions is shown in Supplementary Table S7.

SNV analysis in TCGA data

Exome sequencing data from 239 tumors and 100 adjacent mucosae from TCGA (TCGA discovery) that had not been included in the TCGA Consortium analysis (14) were analyzed using the same pipeline. Only 3.6% of somatic SNVs discovered in our 42 tumors were also found in TCGA discovery data confirming the high heterogeneity of the colorectal cancer mutational landscape. On the other hand, 20 of 137 recurrent SNVs (15%) were present in the TCGA tumors analyzed (Table 1). Of these, 11 were intronic, 1 intergenic, and 1 occurred in the 5′-UTR of the *MASP1* gene. Only 6 of these 20 recurrent mutations were predicted to have a functional effect at the protein level; 5 of them being the well-known *KRAS* G12C, *KRAS* G12D, *TP53* R282W, *APC* R232*, and *APC* E1353*. A stop gain mutation in *AMER1* c.1489C>T (R497*) was identified in 4 tumors, 2 from our series and 2 from the TCGA discovery subset, thus deserving further consideration. If only stage II tumors were taken into account, barely 1.8% of somatic SNVs discovered in our 42 tumors were validated, but these included 13 of 137 recurrent ones.

Somatic mutations in *AMER1*

In addition to the recurrent mutation R497*, 2 more tumors from our series showed stop gain mutations in *AMER1* (also known as *FAM123B* or *WTX*): c.1891C>T (R631*) and c.1876C>T (R626*), with 31%, 64%, 58%, and 55% of allelic frequency, respectively. All 4 mutations were validated by Sanger sequencing (Supplementary Fig. S7A). In the TCGA discovery dataset, 25 of 239 tumors (10.5%) accumulated 26 different somatic mutations

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

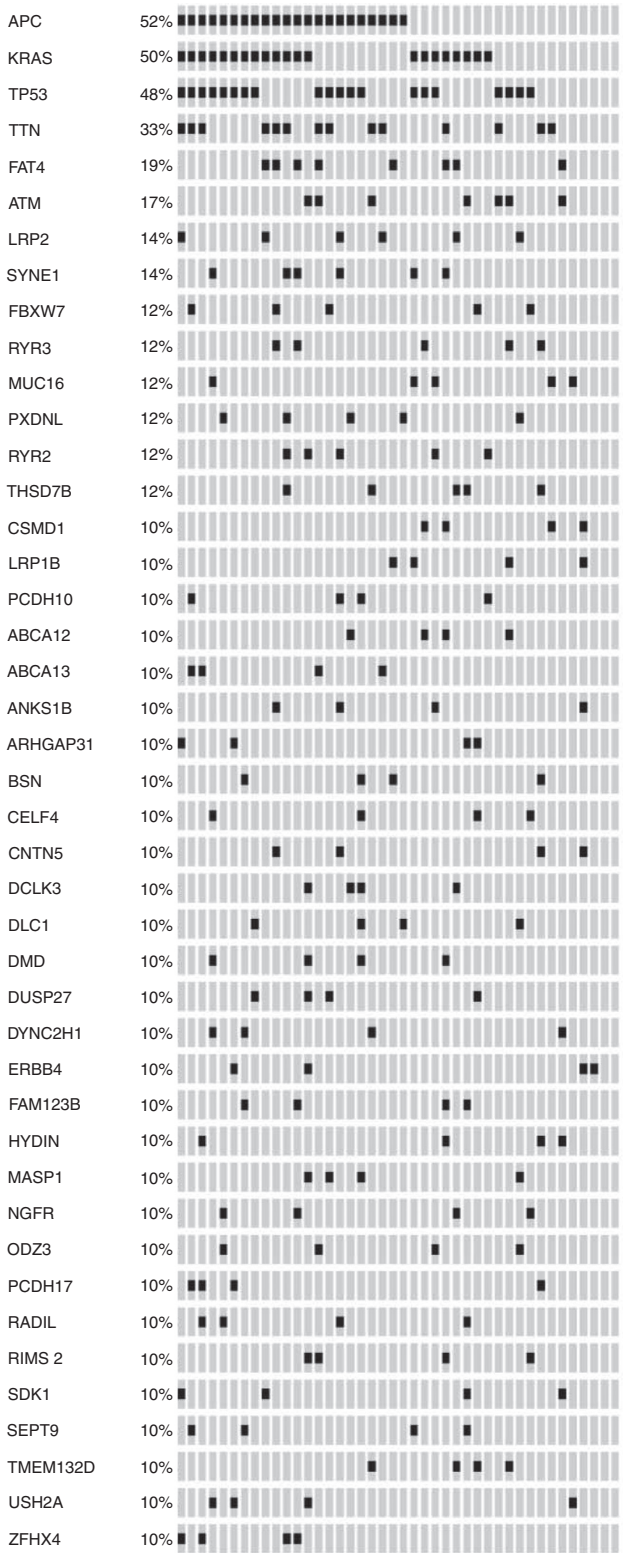


Figure 2.
Mutational map representing top mutated genes. Each column corresponds to each analyzed tumor ($n = 42$), and each line corresponds to genes recurrently mutated. Black hits indicate a functional mutation in such tumor and gene. OncoPrinter tool from cBioPortal has been used to generate this figure (33).

in *AMER1*, including the recurrent R497* (Supplementary Fig. S7B). From these, 10 were stop codon, 10 missense, 3 coding synonymous, and 3 were located in the UTR of the gene (Fig. 3; Supplementary Table S8).

Additional validation of *AMER1* mutations was performed in independent series of colorectal cancer, including 87 TCGA cases (TCGA validation subset) and 227 colorectal cancer tumors recruited at the Bellvitge University Hospital. Exome sequencing data analysis of TCGA validation revealed that 15% (13 of 87) of the tumors carried somatic variants in *AMER1*. In all, 2 variants were stop gain, 4 missense, 2 synonymous, and 5 were located in the 3′-UTR (Supplementary Table S8). The most recurrently mutated regions of *AMER1* were Sanger-sequenced in the 227 colon tumors from the Bellvitge University Hospital. In this series, 3 previously identified recurrent stop gain mutations, R497*, R631*, and R626*, were detected in 5 tumors (Supplementary Fig. S7C). In all mutated cases, adjacent mucosa exhibited a wild-type genotype supporting the somatic origin of the mutations (Supplementary Fig. S7A and S7C). Regarding the functional effect of the identified missense mutations, 10 of the 14 were predicted to be damaging by at least one of the prediction algorithms used (Supplementary Table S8).

Under the hypothesis that chromosome deletions can also cause the somatic loss of *AMER1*, exome sequencing data were used to detect CNVs in chromosome X. One of the 42 analyzed tumors exhibited the complete loss of one copy of such chromosome (Supplementary Fig. S8). Actually, the patient was a female whose tumor also carried a truncating mutation in *AMER1*, suggesting the complete inactivation of *AMER1*.

To assess whether other molecular mechanisms, apart from mutation or CNV, could inactivate *AMER1*, its methylation status was evaluated in 96 samples (including the 42 sequenced tumors, NCBI BioProject PRJNA188510). In females, several tumors were found to be hypo- and hypermethylated in the promoter region when compared with their paired adjacent samples. As expected, a negative correlation (Pearson $r = -0.22$, $P = 0.03$) between the level of methylation and *AMER1* expression was found in this subset of patients. This trend suggests that some tumors may have *AMER1*-inactive by an epigenetic regulatory mechanism (Supplementary Fig. S9). Nevertheless, nor CNV neither methylation was as frequent inactivating events as mutations in our data.

Phenotypic features associated with *AMER1* inactivation

To confirm the effect of *AMER1* truncating mutations at the protein level, immunohistochemical staining was performed in the 4 (stop gain) mutated samples of the original series. As expected, no protein expression or reduction in expression was detected in 3 of 4 analyzed tumors, whereas strong staining was detected in all adjacent samples and tumors not harboring mutations in *AMER1* (Supplementary Fig. S10). To decipher whether *AMER1* mutational inactivation confers a characteristic tumor phenotype, we used expression data from the same tumors to assess whether *AMER1*-silenced tumors showed characteristic patterns of expression of related pathways and functions (data deposited in GEO repository with access code GSE44076 and project code PRJNA188510). In addition to *AMER1*-mutated and hypermethylated tumors, one male patient from the GSE44076 set showing the complete loss of chromosome X was included in this analysis (data not shown). As expected, tumors with altered

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

Table 1. Validated recurrent mutations

Gene	Variant	rs identifier	Function/location	CLX samples, *n*	TCGA samples, *n*
TP53	chr17:7577094; G>A	rs28934574	Missense	$n = 4$ AF = 0.64; 0.6; 0.59; 0.63	$n = 16$ AF = 0.63; 0.87; 0.51; 0.84; 0.58; 0.73; 0.57; 0.48; 0.39; 0.93; 1
KRAS	chr12:25398284; C>A	rs121913529	Missense	$n = 5$ AF = 0.67; 0.46; 0.4; 0.43; 0.46	$n = 25$ AF = 0.6; 0.41; 0.45; 0.46; 0.67; 0.45; 0.29; 0.43; 0.48; 0.3; 0.49; 0.67; 0.2; 0.21; 0.43; 0.37; 0.7; 0.44; 0.45; 0.31; 0.36; 0.35; 0.48; 0.33; 0.4
KRAS	chr12:25398284; C>T	rs121913529	Missense	$n = 8$ AF = 0.19; 0.46; 0.35; 0.33; 0.31; 0.24; 0.46; 0.49	$n = 17$ AF = 0.35; 0.29; 0.6; 0.25; 0.22; 0.3; 0.29; 0.32; 0.27; 0.76; 0.26; 0.35; 0.35; 0.42; 0.48; 0.35; 0.52; 0.49; 0.31
APC	chr5:112128191; C>T	rs0	Stop gained	$n = 4$ AF = 0.71; 0.33; 0.29; 0.23	$n = 3$ AF = 0.24; 0.34; 0.18; 0.8
APC	chr5:112175348; G>T	rs0	Stop gained	$n = 2$ AF = 0.71; 0.57	$n = 4$ AF = 0.24; 0.34; 0.18; 0.81
AMER1	chrX:63411678; G>A	rs0	Stop gained	$n = 2$ AF = 0.31; 0.64	$n = 2$ AF = 0.76; 0.65
MASP1	chr3:187009800; A>T	rs0	5′-UTR	$n = 3$ AF = 0.38; 0.54; 0.41	$n = 1$ AF = 0.64
MS4A2	chr11:59861311; A>G	rs113221333	Intronic	$n = 3$ AF = 0.20; 0.25; 0.2	$n = 3$ AF = 0.31; 0.21; 0.31
KIF7	chr15:90173735; G>A	rs0	Intronic	$n = 3$ AF = 0.21; 0.25; 0.2	$n = 1$ AF = 0.25
KIF13A	chr6:17787892; C>A	rs62394104	Intronic	$n = 3$ AF = 0.21; 0.16; 0.27	$n = 1$ AF = 0.19
EMC2	chr8:109468234; T>A	rs111255731	Intronic	$n = 3$ AF = 0.17; 0.27; 0.18	$n = 1$ AF = 0.38
HOXD10	chr2:176983604; C>A	rs73974643	Intronic	$n = 2$ AF = 0.21; 0.5	$n = 3$ AF = 0.34; 0.21; 0.19
SETD2	chr3:47143125; C>T	rs200952697	Intronic	$n = 2$ AF = 0.22; 0.22	$n = 3$ AF = 0.16; 0.44; 0.29
PDS5B	chr13:33253144; A>C	rs199860513	Intronic	$n = 2$ AF = 0.18; 0.18	$n = 3$ AF = 0.23; 0.74; 0.80
AGBL1	chr15:87474796; T>G	rs0	Intronic	$n = 2$ AF = 0.23; 0.36	$n = 1$ AF = 0.36
NUP133	chr1:229623415; A>T	rs0	Intronic	$n = 2$ AF = 0.2; 0.19	$n = 1$ AF = 0.3
DPP9	chr19:4720039; A>C	rs0	Intronic	$n = 2$ AF = 0.23; 0.35	$n = 1$ AF = 0.54
PARL	chr3:183584713; C>T	rs199558489	Intronic	$n = 2$ AF = 0.25; 0.3	$n = 1$ AF = 0.23
VCL	chr10:75874194; T>C	rs0	Intronic	$n = 2$ AF = 0.23; 0.19	$n = 1$ AF = 0.35
Intergenic	chr3:195433152; G>A	rs76183393	Intergenic	$n = 2$ AF = 0.52; 0.33	$n = 3$ AF = 0.42; 0.41; 0.66

NOTE: Reference genome: hg19.
Abbreviation: AF, allelic frequency.

AMER1 tended to cluster together when analyzing the genes related to β-catenin binding (Fig. 4A) and to the Wnt pathway (Fig. 4B). Regarding the latter, overall Wnt-related genes were underexpressed in cluster 1, grouping 7 of 9 tumors with aberrant *AMER1*. In fact, overexpressed genes in this cluster included Wnt inhibitors such as *PRICKLE1*, *PRICKLE2*, and *DAAM2*, Wnt antagonists such as the SFRP family and *SOX17*, and noncanonical Wnt pathway activators such as *WNT5A*. Moreover, the potential prognostic value of *AMER1* was assessed, but no association was found with disease-free survival (Cox proportional hazards: $P = 0.58$).

We also evaluated the co-occurrence of *AMER1* mutations with mutations in genes related to the Wnt pathway as described (14). Within 266 tumors (discovery dataset and TCGA), 72% of *AMER1*-mutated samples had also *APC* mutated. However, co-occurrence of *AMER1* mutations with other Wnt genes were rare (i.e., *CTNNB1*, 1 of 12; *DKK2*, 1 of 7; *TCF7L2*, 3 of 19) or even mutually exclusive (i.e., *LRP5*; Supplementary Fig. S11).

We next assessed the relationship between *AMER1* mutational status and MSI–MSS–CIMP molecular subtypes, as well as *KRAS* and *BRAF* mutational status. Samples from CLX and TCGA were used ($n = 322$). We found that the majority of *AMER1*-lacking tumors were MSS (70%), *BRAF* wild-type (91%), and CIMP-H–negative (88%; Supplementary Fig. S12). We also wanted to assess if *AMER1*-deficient tumors belong to any of the recently described molecular subtypes of colorectal cancer. We used the gene lists reported by Roepman and colleagues (29) to construct a score able to rank tumors according to subtype B or subtype C gene

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

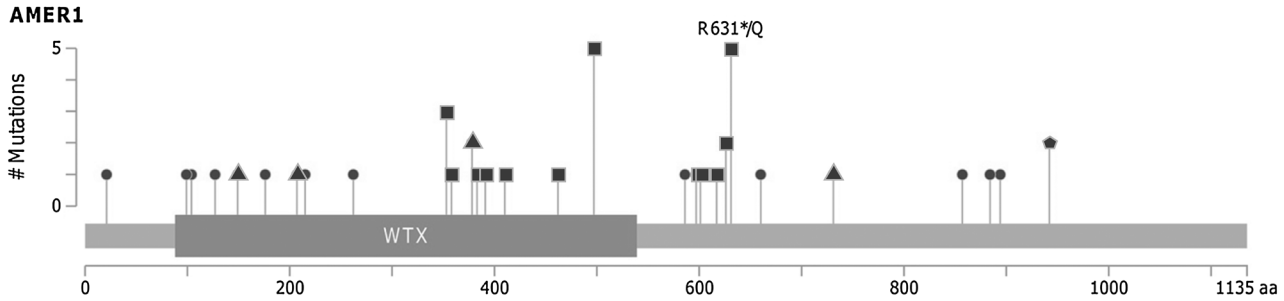


Figure 3.
AMER1 mutations. Lollipop plot showing the distribution of *AMER1* mutations across the coding protein. The *y*-axis represents the number of mutations. Circles indicate missense mutations, squares truncating mutations, and triangles synonym mutations. Pentagon indicates residues affected by different mutation types. MutationMapper tool from cBioPortal has been used to generate this figure (33).

expression. Because our study only included MSS tumors, subtype A (which mainly comprises MSI tumors) was not included in the analysis. Subtype B mainly comprises epithelial tumors with active Wnt and better prognosis whereas those subtype C were mesenchymal tumors exhibiting worse prognosis. We observed that *AMER1*-deficient tumors tend to score higher in the C than in the B subtype (also if only mutated tumors were taking into account; Supplementary Fig. S13). This trend was even more marked when tumors from the 2 main clusters defined in Fig. 4B were compared. The lower score in B subtype shown by *AMER1*-mutant tumors was also validated in 224 tumors from published TCGA data (Fig. 4C). This result was supported by the level of expression of molecular markers associated with the C subtype: a decrease in proliferation markers and an increase in EMT, NOTCH, and VEGF markers (Supplementary Fig. S14).

Public data from cBio Cancer Genomics Portal was used to compare mutations and deletions of *AMER1* gene across different cancer types (33). Colorectal was the tumor that accumulated more mutations (more than 10%) followed by lung, endometrial, and melanoma. On the contrary, other tumors such as leukemia, medulloblastoma, breast, or ovary showed none or low levels of *AMER1* mutations. One study in prostate cancer found more than 20% of tumors with amplifications in this locus. However, 6 more prostate studies showed neither mutations nor CNV in *AMER1* (Supplementary Fig. S15).

Discussion

In an attempt to better understand the colorectal cancer pathobiology at a genomic level, 42 colon tumors were profiled by means of exome sequencing. Despite the high mutational heterogeneity among tumors, mutations in *AMER1* emerged as a recurrent feature in colorectal cancer. Reinforcing its putative role as a driver gene, MutSig software scored *AMER1* between the top 10 functional genes ($P = 0.018$) along with *TP53*, *APC*, and *KRAS*. Although mutations in *AMER1* have been observed in other studies, this gene has not received proper attention as a potential driver for colorectal cancer.

AMER1 (also known as *FAM123B* or *WTX*) is a gene located in chromosome X that codifies a highly conserved membrane protein that acts as scaffold for β-catenin degradation. AMER1 is associated with the plasma membrane via 2 N-terminal domains and forms complexes with APC, β-catenin, Axin, and β-TrCP. It can recruit APC from microtubules to the plasma membrane and it is also involved in stimulating LRP6 phosphorylation (34). In tumors, AMER1 is a negative regulator of the Wnt/β-catenin pathway by promoting β-catenin ubiquitination and degradation (35). Also, it has been reported as a repressor of Wnt signaling when cells establish cell–cell contacts. AMER1 maintains the integrity of cellular junctions by mediating the membrane localization of APC (36).

Truncating mutations in the *AMER1* gene are frequent in Wilm tumors (30%), which are pediatric tumors of the kidney (37). Yoo and colleagues performed a mutational analysis of *AMER1* in gastric, colorectal, and hepatocellular carcinomas and found no mutations in colorectal cancer tumors (0 of 141; ref. 38). However, the TCGA consortium reported *AMER1* as a frequently mutated gene in colorectal cancer (14). Seshagiri and colleagues also found functional mutations in *AMER1* gene in colorectal cancer tumors: R177C, E384*, G105D, and E244* (39). Recently, mutations in *AMER1* have also been described in metastatic colorectal cancer samples and their paired primary tumors (40). Interestingly, somatic mutations affecting an X chromosome gene raise the possibility of one-hit inactivation of a tumor suppressor gene. Indeed, 6 of 7 mutations in our set occurred in male tumors, and the loss of the *AMER1* wild-type copy of chromosome X was observed in the *AMER1*-mutated tumor developed by a female. Interestingly, Han and colleagues previously reported the deletion of the *AMER1* locus (Xq11) in Wilm tumors, providing evidence of the inactivation of the gene via copy number changes (41). Regarding our analysis of the TCGA data, 10 of the 14 stop gain mutations were found in males as well as 10 of the 14 missense mutations. *AMER1* has been classically catalogued as a Wnt signaling inhibitor due to its belonging to the β-catenin complex. In this scenario, loss of *AMER1* would lead to activation of canonical Wnt pathway by β-catenin translocation into the nucleus. However, our results point to the inactivation of Wnt signaling or to the activation of a noncanonical Wnt pathway in *AMER1*-mutated tumors. These findings suggest that *AMER1* has an alternative function to its role in the canonical Wnt signaling pathway in colorectal cancer tumors, as has been proposed by other authors (42). Indeed, it has been described that *AMER1* inhibits or activates the Wnt pathway in Wilm disease, depending on the mesenchymal or epithelial origin of the tumor. *AMER1* mutations in the mesenchymal lineage lead to nuclear accumulation of β-catenin and subsequent upregulation of Wnt targets. On the other hand, *AMER1* mutations in the epithelial lineage are not associated with active Wnt (42). This observation agrees with our findings, where *AMER1*-mutated colorectal cancer (epithelial) tumors show inactive Wnt. Interestingly, more than

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

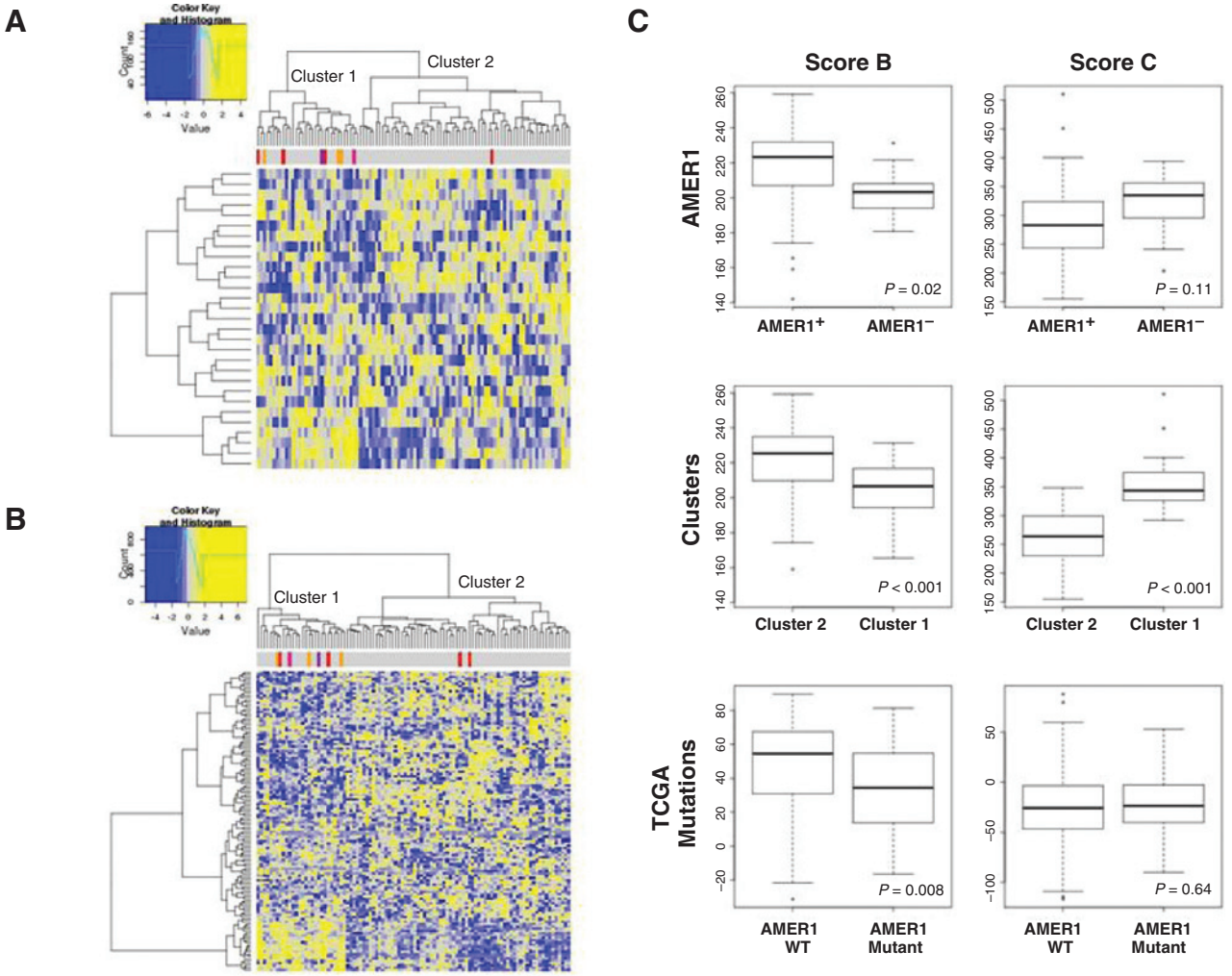


Figure 4.
Phenotypic features associated with *AMER1* inactivation. Heatmap showing gene expression profile of genes implicated in β-catenin–binding function (A) and Wnt pathway (B). Those tumors with loss of *AMER1* tend to aggregate in the same cluster. Color bars represent *AMER1*-mutated tumors (in red from exome sequencing and in orange from Sanger sequencing), a tumor from a male patient with loss of chromosome X by CNV (violet), and a tumor with hypermethylation in *AMER1* (dark pink). Underexpression is painted in yellow whereas downexpression is painted in blue. C, boxplots showing differences in score B subtype and score C subtype between tumors lacking and nonlacking *AMER1* gene (named as AMER1), between the 2 main clusters defined in B (named as Clusters) and between tumors with and without mutations in *AMER1* in TCGA data (named as TCGA mutations). *P* value is based on the nonparametric Mann–Whitney test.

50% of tumors harboring *AMER1* mutations were also *APC* mutants whereas they rarely co-occurred with *CTNNB1* and other genes in the Wnt pathway, probably indicating a characteristic and exclusive pathway activation of *AMER1* tumors. Regarding molecular classification, our results pointed to *AMER1*-mutant tumors as mainly MSS and in rare co-occurrence with the CIMP phenotype and *BRAF* mutations. Recently, molecular subtyping of colorectal cancer that takes into account different molecular features of the tumors has been proposed (4, 29). Our results suggested that the subset of tumors lacking *AMER1* expression could belong to type C tumors. This subtype is characterized by the expression of EMT markers and shows lower proliferative ratio. Clinically, type C tumors exhibit poor prognosis and are chemotherapy-resistant. However, our results showed no association between mutational status of *AMER1* and tumor relapse.

In concordance with previous observations (43, 44), the mutational patterns of colon tumors are highly heterogeneous. Our results indicate that the vast majority of SNVs were private (non recurrent). Indeed, an independent validation only found 3% of SNVs shared by more than one tumor, confirming the high heterogeneity of the colorectal cancer mutational landscape. Nevertheless, 15% of the small fraction of recurrent mutations found in our study were also observed in the TCGA tumors, the well-known driver mutation being *KRAS* G12D, the most recurrent one (14). Also in consistence with previous colorectal cancer genomic studies, mutational changes in colorectal cancer are predominated by C:G > T:A transitions (14, 32, 45). The background rate of somatic mutations in colorectal cancer has been reported to be approximately 1 mutation per megabase (46). However, mutation frequencies in colorectal cancer tumors are not homogeneous due to differences in the status of the mismatch

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

repair machinery (MSI vs. MSS) or to the presence or absence of POLE mutations (14, 47). Because our sample did not include MSI tumors, the observed mutation rate and number of SNVs are similar to those previously reported in MSS colorectal tumors.

In our series, *APC* appeared as the most mutated gene followed by *KRAS*, *TP53*, and *TTN*. *TTN* mutations have been previously identified in colorectal cancer (14) and in other tumors (48), probably due to the fact that it codes for the longest human protein, increasing the likelihood of passenger mutations, most probably unrelated to cancer (24). Other recurrently mutated genes deserve further consideration: mutations in *FBXW7*, a gene encoding a protein implicated in Notch signaling, were identified in 12% of the tumors. Similar results have been reported in a recent colorectal cancer study comparing primary and metastatic colorectal tumors (49). Also, 9.5% of tumors showed mutations in *CSMD1*, which have been associated with poor prognosis in colorectal cancer (50, 51). Interestingly, in our study, all functional *CSMD1* mutations occurred in patients who relapsed. Recently described as recurrently mutated genes in colorectal cancer, *SYNE1*, *FAT4*, *ATM*, and *USH2A*, (32), were also found in our study with more than one tumor mutated (Fig. 2).

The mutational patterns of colorectal cancer are highly heterogeneous among patients. However, mutations tend to accumulate in common pathways and functions crucial for tumorigenesis (i.e., apoptosis, cell cycle). This suggests that a broad range of equivalent genetic aberrations could deregulate key pathways in carcinogenesis, as has already been postulated (52). In other words, diverse molecular alterations could converge in similar phenotypes.

Exome sequencing is a useful technique to discover still unknown mutations that can lead us to a better understanding of the mechanisms underlying colorectal carcinogenesis. However, it also has technical limitations. Mutations are more easily detected in high-coverage regions, so regions at the extremes of the captured exons many suffer from smaller sensitivity. Also, capture kits do not cover equally all exons in the genome. Moreover, tumor heterogeneity and stromal contamination must be taken into account, as it may lead to difficulties in differentiating low-frequency mutations from technical artifacts. Our mutation-calling algorithm required a minimum number of reads with the mutation and total coverage to increase the likelihood of a correct mutation calling.

In conclusion, our exome sequencing approach has revealed that MSS stage II colon tumors exhibit a highly heterogeneous somatic mutational landscape. In concordance with previous studies, this finding clearly suggests that colorectal cancer is not a single disease and supports the necessity of pathway-directed treatments. We have also described that approximately 10% of colorectal cancer tumors often harbor *AMER1* inactivation due to somatic mutation, loss of the chromosome X, or hypermethylation (in lower fraction). This subgroup of tumors with *AMER1* deficiency also exhibited a particular gene expression pattern in Wnt signaling pathway genes and showed an overall gene expression phenotype similar to the molecular subtype C described by Roepman and colleagues(29). Although promising, further experimental work is required to clearly demonstrate the role of *AMER1* as a tumor suppressor gene in colon cancer tumors.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: R. Sanz-Pamplona, L. Valle, V. Moreno
Development of methodology: R. Sanz-Pamplona, F. Bellido, M. Gut
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): E. Guinó, X. Sanjuan, A. Soriano, R. Salazar
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R. Sanz-Pamplona, A. Lopez-Doriga, M.H. Alonso, S. Aussó, S. Beltrán, F. Castro-Giner, A. Closa, D. Cordero, F.D. Morón-Duran, V. Moreno
Writing, review, and/or revision of the manuscript: R. Sanz-Pamplona, A. Lopez-Doriga, M.H. Alonso, S. Beltrán, D. Cordero, F.D. Morón-Duran, R. Salazar, L. Valle, V. Moreno
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): L. Paré-Brunet, E. Guinó, S. Beltrán, M. Gut, D. Cordero
Study supervision: V. Moreno
Other (performed experiments): K. Lázaro

Acknowledgments

The authors thank Carmen Atencia, Pilar Medina, and Isabel Padrol for their help with the clinical annotation of the samples. They also thank Ana M[a] Corraliza for helping with bioinformatics analysis and Gemma Aiza for technical assistance.

Grant Support

This study was supported by the Instituto de Salud Carlos III grants (FIS PI09-01037, PI11-01439, and PIE13-00022), CIBERESP CB07/02/2005, Spanish Ministry of Economy and Competitiveness (SAF2012-38885), the Spanish Association Against Cancer (AECC) Scientific Foundation, the Catalan Government DURSI grant 2014SGR647, and NIH grant U19CA148107 (CORECT). Sample collection was supported by the Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d'Oncología de Catalunya (XBTC) and ICOBiobanc, sponsored by the Catalan Institute of Oncology.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 24, 2015; revised April 28, 2015; accepted May 17, 2015; published OnlineFirst June 12, 2015.

References

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer 2013;49:1374–403.
2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–67.
3. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. J Mol Diagn 2008;10:13–27.
4. Sanz-Pamplona R, Santos C, Grasselli J, Molleví DG, Dienstmann R, Paré-Brunet L, et al. Unsupervised analyses reveal molecular subtypes associated to prognosis and response to therapy in colorectal cancer. Colorectal Cancer 2014;3:277–88.
5. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24.
6. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. N Engl J Med 2009;361:2449–60.
7. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108–13.
8. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res 2012;40:e169.
9. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet 2004;5:335–44.

Published OnlineFirst June 12, 2015; DOI: 10.1158/1078-0432.CCR-15-0159

10. Izarzugaza JM, Redfern OC, Orengo CA, Valencia A. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. Proteins 2009;77:892–903.
11. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science 2011;333:1154–7.
12. Jia D, Dong R, Jing Y, Xu D, Wang Q, Chen L, et al. Exome sequencing of hepatoblastoma reveals novel mutations and cancer genes in the Wnt pathway and ubiquitin ligase complex. Hepatology 2014;60:1686–96.
13. Sanz-Pamplona R, Berenguer A, Cordero D, Molleví DG, Crous-Bou M, Sole X, et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. Mol Cancer 2014;13:46.
14. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.
15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.
16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29:308–11.
17. 1000 Genomes Project Consortium. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
18. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
19. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics 2014;15:244.
20. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272–6.
21. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76.
22. Pique-Regi R, Caceres A, Gonzalez JR. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. BMC Bioinformatics 2010;11:380.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
24. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
25. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
27. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 2005;21:3176–8.
28. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 2010; 11:548.
29. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer 2014;134:552–62.
30. Ogino S, Kawasaki T, Kirkner GJ, Kraft P, Loda M, Fuchs CS. Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. J Mol Diagn 2007;9:305–14.
31. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep 2013;3:2650.
32. Yu J, Wu WK, Li X, He J, Li XX, Ng SS, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. Gut 2015; 64:636–45.
33. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401–4.
34. Tanneberger K, Pfister AS, Kriz V, Bryja V, Schambony A, Behrens J. Structural and functional characterization of the Wnt inhibitor APC membrane recruitment 1 (Amer1). J Biol Chem 2011;286: 19204–14.
35. Major MB, Camp ND, Berndt JD, Yi X, Goldenberg SJ, Hubbert C, et al. Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. Science 2007;316:1043–6.
36. Grohmann A, Tanneberger K, Alzner A, Schneikert J, Behrens J. AMER1 regulates the distribution of the tumor suppressor APC between microtubules and the plasma membrane. J Cell Sci 2007; 120:3738–47.
37. Rivera MN, Kim WJ, Wells J, Driscoll DR, Brannigan BW, Han M, et al. An X chromosome gene, WTX, is commonly inactivated in Wilms tumor. Science 2007;315:642–5.
38. Yoo NJ, Kim S, Lee SH. Mutational analysis of WTX gene in Wnt/beta-catenin pathway in gastric, colorectal, and hepatocellular carcinomas. Dig Dis Sci 2009;54:1011–4.
39. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature 2012; 488:660–4.
40. Brannon AR, Vakiani E, Sylvester BE, Scott SN, McDermott G, Shah RH, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol 2014;15:454.
41. Han M, Rivera MN, Batten JM, Haber DA, Dal Cin P, Iafrate AJ. Wilms' tumor with an apparently balanced translocation t(X;18) resulting in deletion of the WTX gene. Genes Chromosomes Cancer 2007;46:909–13.
42. Fukuzawa R, Anaka MR, Weeks RJ, Morison IM, Reeve AE. Canonical WNT signalling determines lineage specificity in Wilms tumour. Oncogene 2009;28:1063–75.
43. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science 2013;339:1546–58.
44. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–74.
45. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502: 333–9.
46. Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, et al. Mutational analysis of the tyrosine kinome in colorectal cancers. Science 2003;300:949.
47. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 2013; 501:338–45.
48. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007; 446:153–8.
49. Xie T, Cho YB, Wang K, Huang D, Hong HK, Choi YL, et al. Patterns of somatic alterations between matched primary and metastatic colorectal tumors characterized by whole-genome sequencing. Genomics 2014;104: 234–41
50. Zhang R, Song C. Loss of CSMD1 or 2 may contribute to the poor prognosis of colorectal cancer patients. Tumour Biol 2014;35:4419–23.
51. Shull AY, Clendenning ML, Ghoshal-Gupta S, Farrell CL, Vangapandu HV, Dudas L, et al. Somatic mutations, allele loss, and DNA methylation of the Cub and Sushi Multiple Domains 1 (CSMD1) gene reveals association with early age of diagnosis in colorectal cancer patients. PLoS One 2013;8: e58731.
52. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.

Material suplementari:

Supplementary Table 1: Baseline characteristics of CRC patients

CRC patients exome sequencing (n=42)	
Gender	
Male	31 (73,8 %)
Female	11 (26,2 %)
Median age (range,years)	70 (43 - 84)
Site	
Right	12 (28,6 %)
Left	30 (71,4 %)
Stage	
II A	38 (90,5 %)
II B	4 (9,5 %)
Recurrence	
No relapse	21 (50 %)
Relapse	21 (50 %)
Recurrence-free median time (range,months)	60,9 (6,8 – 127,8)

CRC patients Sanger validation cohort (n=56)	
Gender	
Male	40 (71,4 %)
Female	16 (28,6 %)
Median age (range,years)	72 (49 - 87)
Site	
Right	26 (46,4 %)
Left	30 (53,6 %)
Stage	
II A	52 (92,8 %)
II B	4 (7,1 %)

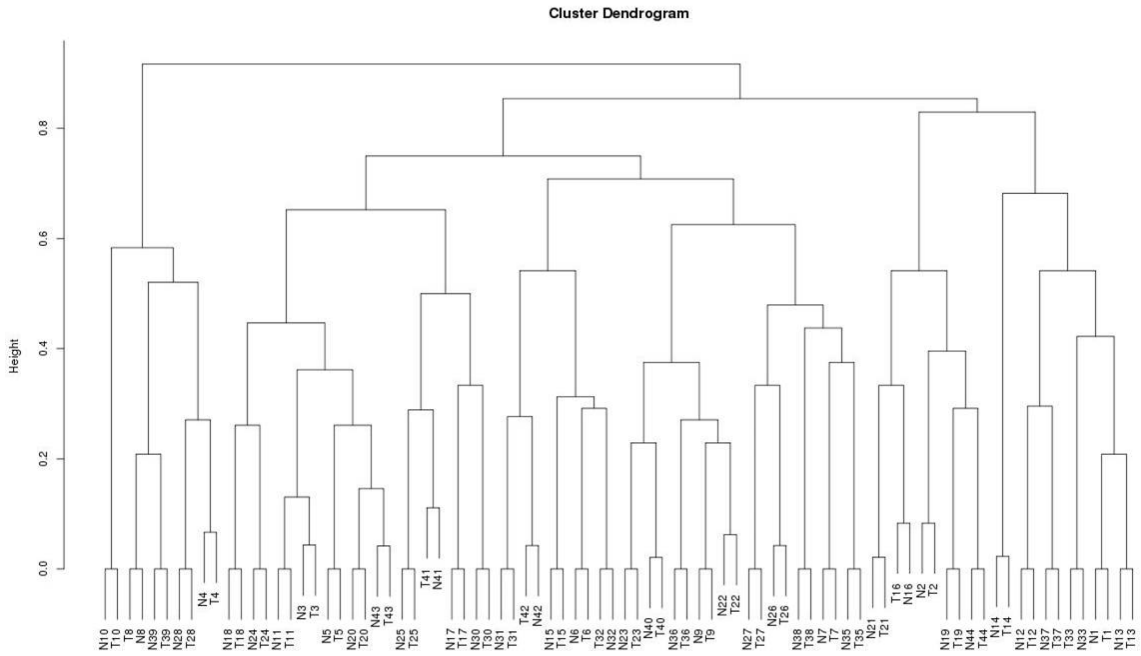
TCGA-discovery (n=239)	
Gender	
Male	138 (58%)
Female	101 (42%)
Median age (range,years)	66
Location	
Colon	173 (73%)
Rectum	64 (27%)
Site	
Right	96 (40)
Left	128 (54)
Transverse	14 (6%)
Stage	
Stage I	36 (15%)
Stage II	90 (40%)
Stage III	75 (32)
Stage IV	31 (13%)

TCGA-validation (n=87)	
Gender	
Male	46 (53%)
Female	41 (47%)
Median age (range,years)	71
Location	
Colon	87 (100%)
Rectum	*0*
Site	
Right	38 (44%)
Left	39 (45%)
Transverse	10 (11%)
Stage	
Stage I	15 (18%)
Stage II	28 (33%)
Stage III	25 (31%)
Stage IV	15 (18%)

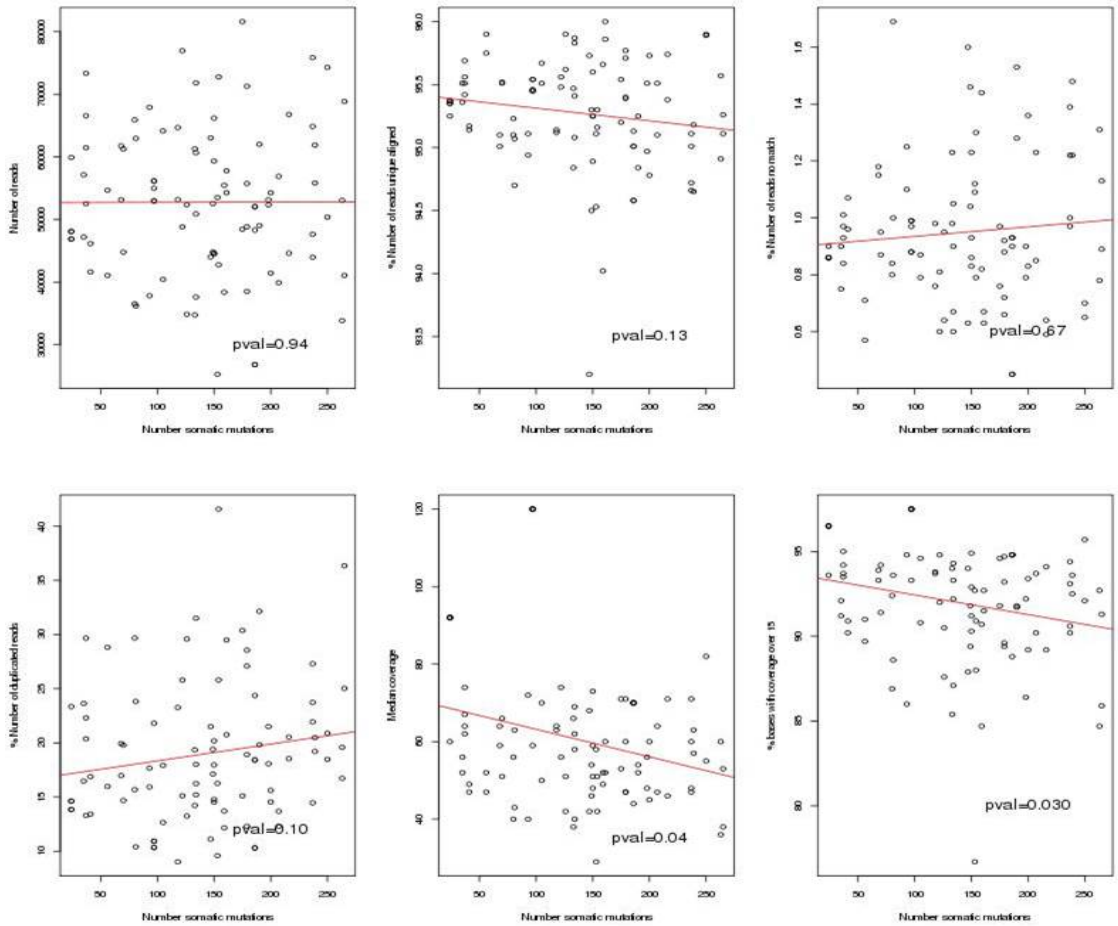
Supplementary Table 2: List of primers used in Sanger sequencing and SNPs used to genotype normal adjacent and tumor samples

PRIMERS	Sequence	Length	Tm (°C)	GC content (%)
Primer F1	CGGTGGGAAATCTGAGAGGT	20	61.4	55
Primer R1	TTGTTGTATTGGGAGCTTCG	20	58.8	45
Primer F2	AAGGCTGTCATCTGGCTCAT	20	59.8	50
Primer R2	TGCTCCTTGACCCAGTTAGG	20	60.2	55
SNPs	**rs**			
	rs6983267, rs4939827, rs16892766, rs10795668, rs9929218, rs961253, rs11169552, rs4444235, rs10411210, rs3802842, rs6691170, rs10936599, rs4925386			

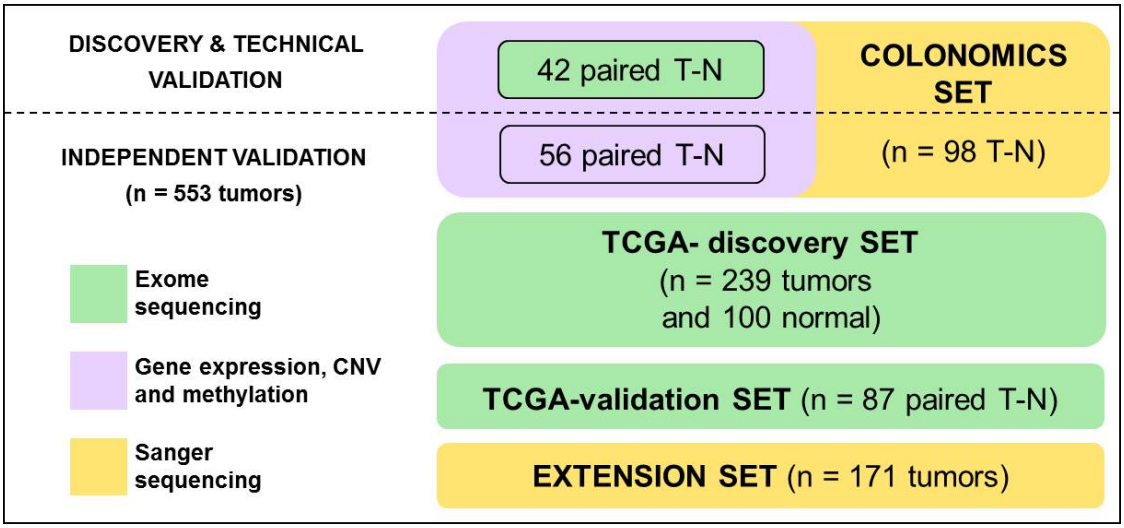
Supplementary Figure 1:



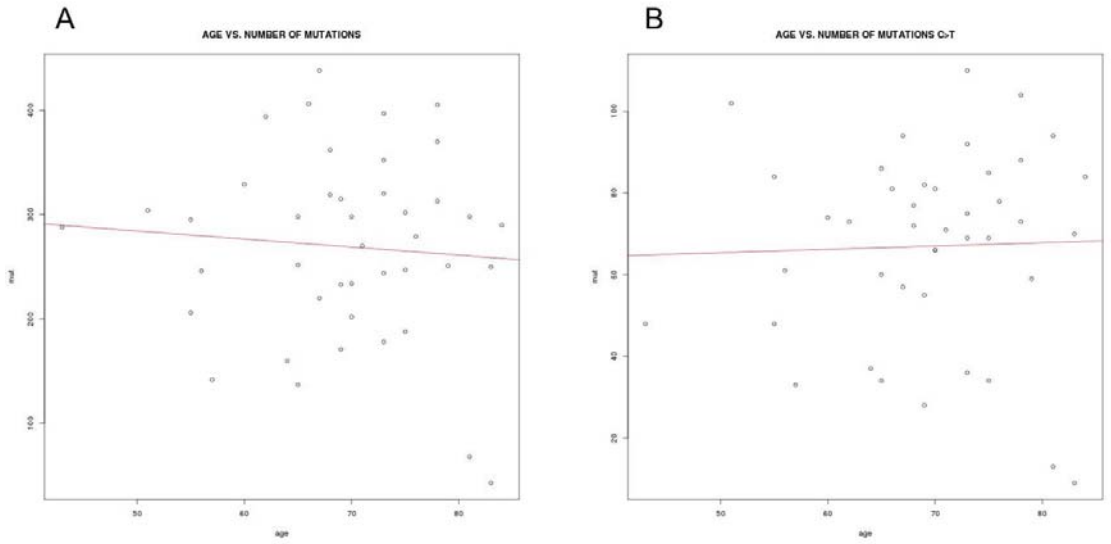
Supplementary Figure 2:



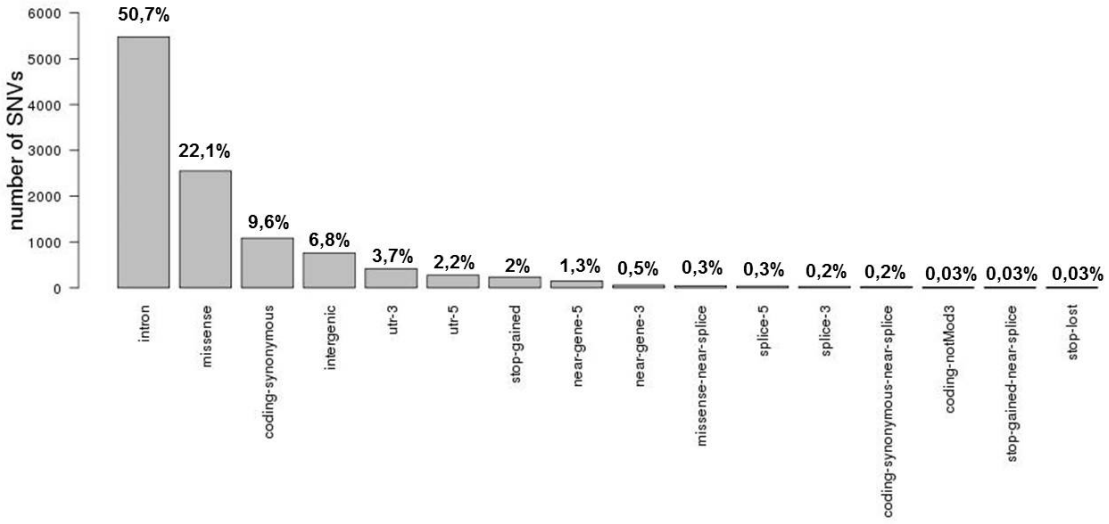
Supplementary Figure 3:



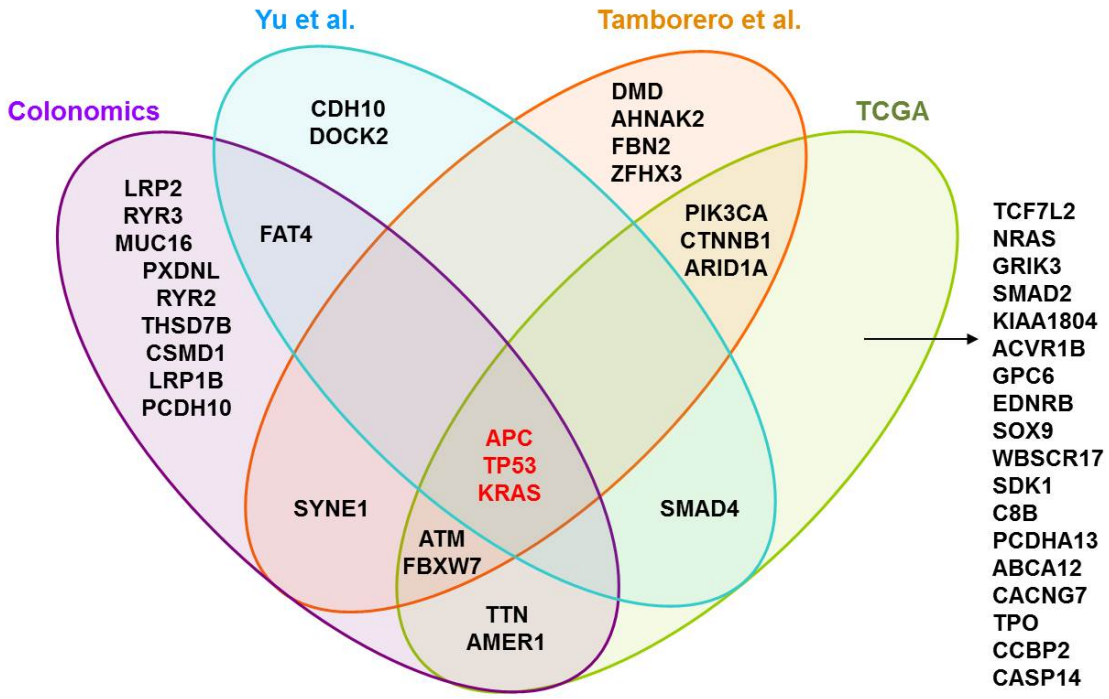
Supplementary Figure 4:



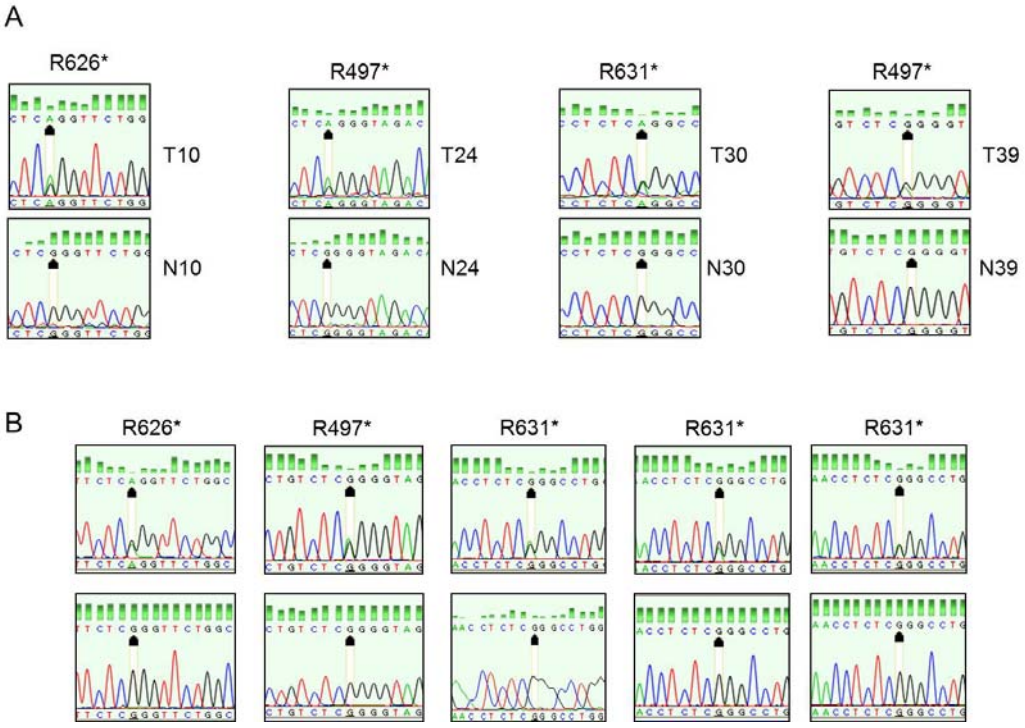
Supplementary Figure 5:



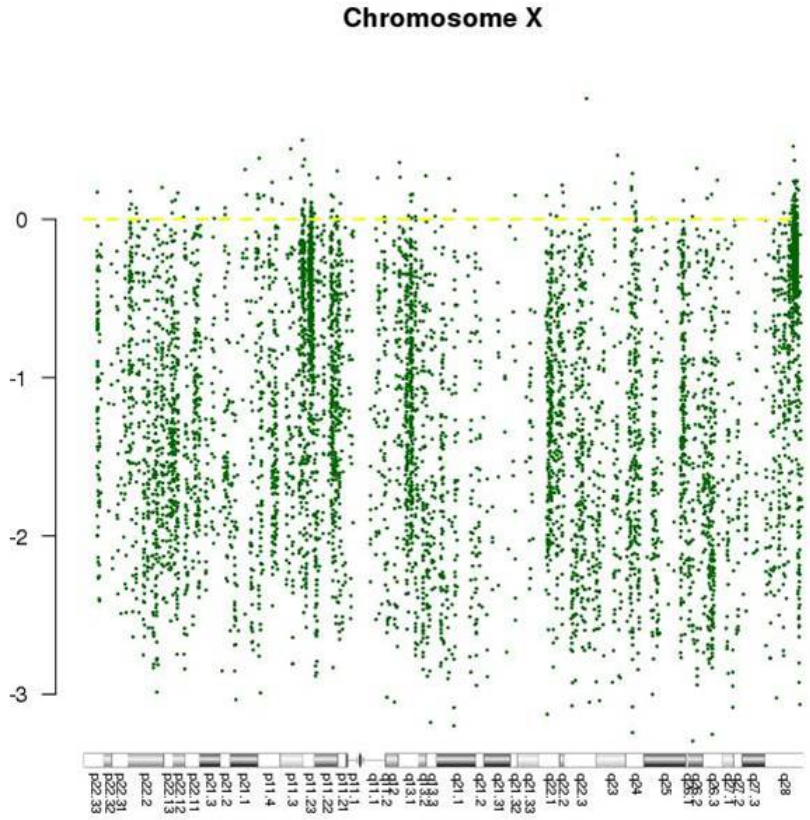
Supplementary Figure 6:



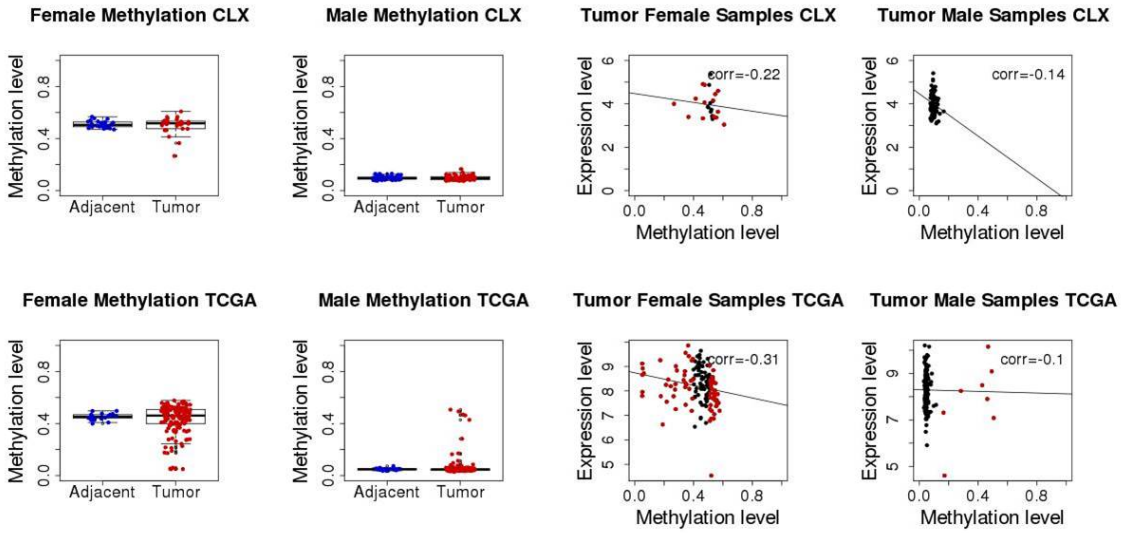
Supplementary Figure 7:



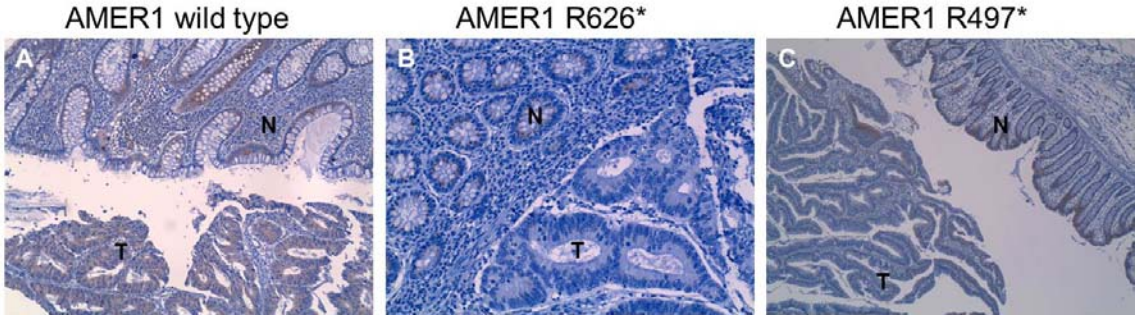
Supplementary Figure 8:



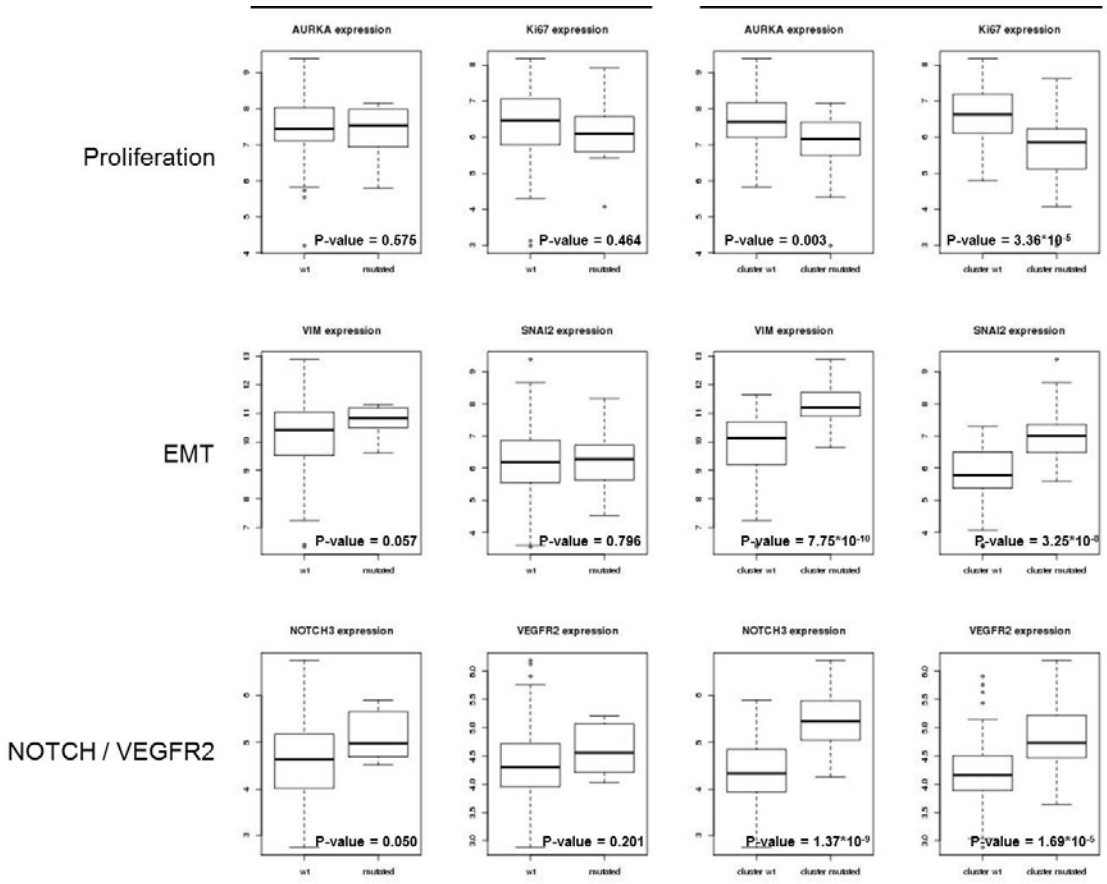
Supplementary Figure 9:



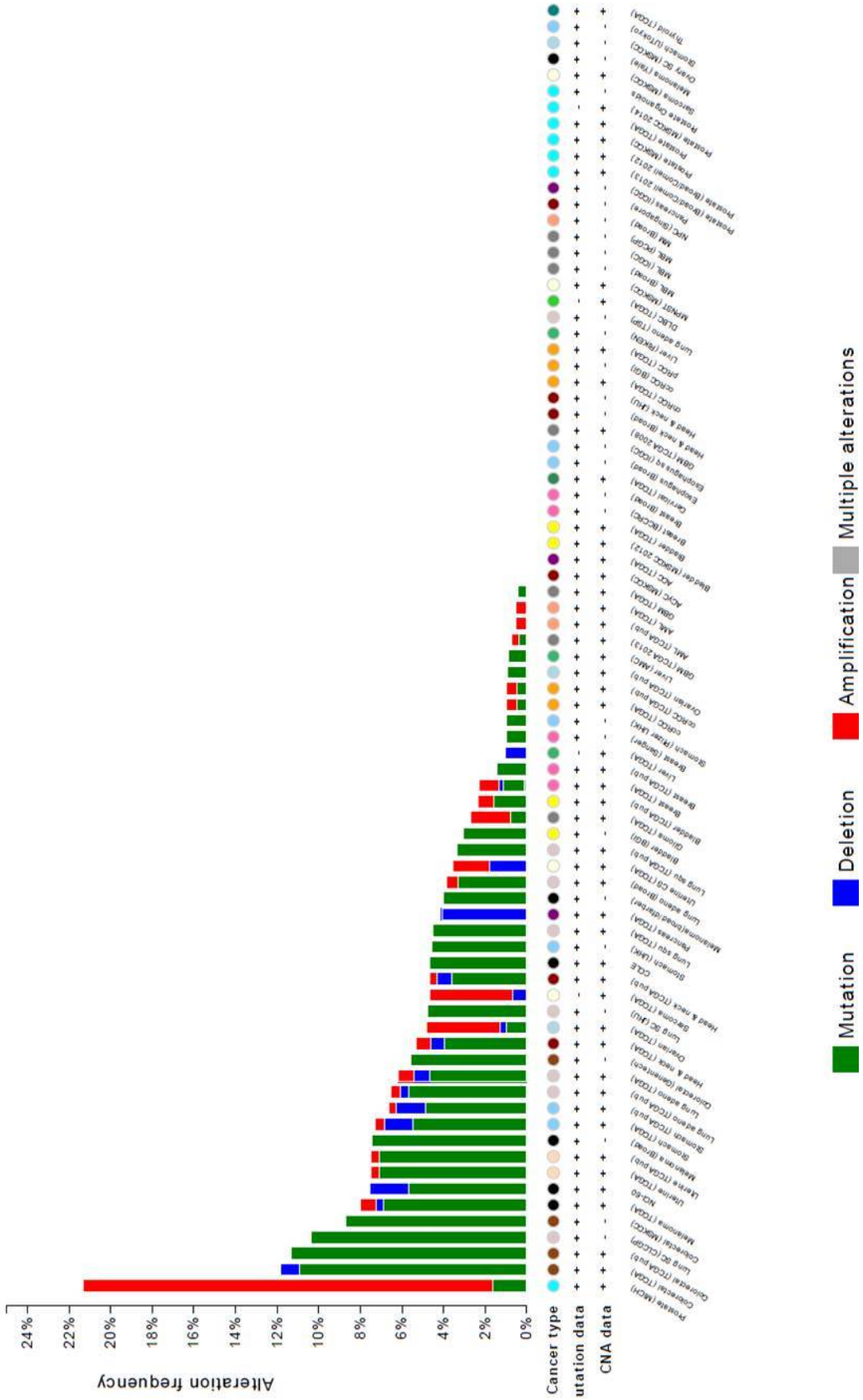
Supplementary Figure 10:



Supplementary Figure 11:



Supplementary Figure 12:



DISCUSSIÓ

L'objectiu de la present tesi doctoral ha estat la implementació a nivell d'anàlisi bioinformàtica de les noves tecnologies de seqüenciació massiva en dos camps concrets: el diagnòstic genètic del càncer hereditari i la recerca translacional centrada en l'estudi mutacional tumoral en càncer colorectal.

1. Algoritme d'anàlisi bioinformàtica per al diagnòstic de càncer hereditari mitjançant NGS

En els inicis de les tecnologies NGS, aquestes van estar pensades per a l'estudi de genomes, exomes o transcriptomes complets. Les primeres plataformes tenien una gran capacitat i un elevat cost, cosa que les feia poc convenients per a la rutina d'un laboratori clínic de diagnòstic genètic, on se solen analitzar un nombre escàs de gens, en un grup de mostres no massa gran i on cal un retorn dels resultats ràpid. Actualment això ha canviat i la majoria de les companyies del sector han desenvolupat equips de capacitat mitjana per abordar aquest tipus d'anàlisi. La Unitat de Diagnòstic Molecular del Programa de Càncer Hereditari de l'ICO va adquirir a finals de 2010 el primer seqüenciador massiu d'aquestes característiques llançat al mercat: el GS Junior de Roche. L'objectiu del grup era adaptar les possibilitats d'aquest instrument a l'estudi de mutacions germinals realitzat a la Unitat.

Els dos primers articles de la tesi descriuen el procés de desenvolupament d'un protocol de laboratori i d'un algoritme bioinformàtic per analitzar les dades de seqüenciació provinents del GS Junior, pel diagnòstic de càncer de mama i ovari hereditari i pel diagnòstic de càncer de colon hereditari, des de la prova de concepte fins al desenvolupament d'una eina de lliure accés que permet analitzar aquest tipus de dades.

1.1 Prova de concepte

Com hem vist al primer article i a altres estudis publicats, s'han fet moltes proves de concepte per avaluar la validesa de la tecnologia de NGS i de la seva anàlisi bioinformàtica, especialment per arribar a l'especificitat i sobretot, a la sensibilitat que requereix el diagnòstic genètic (Choi et al. 2009, Shearer et al. 2010, Hansen et al. 2014, Swanson et al. 2014).

El primer pas per a posar en marxa nous protocols d'anàlisi és decidir per quina tecnologia apostar, és a dir, quin instrument pot ser més cost-eficient, tenint en compte la quantitat de mostres que cal seqüenciar, les particularitats dels gens que es pretenen seqüenciar, i quin tipus de variants s'espera identificar. En la nostra unitat de diagnòstic es va apostar per la plataforma 454 GS Junior perquè va ser la primera plataforma de mitjà rendiment en sortir al mercat, per la longitud de les lectures, ja que en aquell moment Roche produïa les lectures més llargues, i perquè a més aportava la capacitat de multiplexar i introduir vàries mostres en una mateixa carrera.

Van ser crucials els estudis de potència estadística per a determinar el nombre de mostres que s'introduirien en una carrera per tal d'obtenir una cobertura mínima de 38x, acceptant un marge d'error, i trobant les variants amb una freqüència mínima del 25% (De Leeneer et al. 2011). Durant la prova de concepte van sorgir petites dificultats tant en la posada a punt dels protocols de laboratori, com en la determinació del *software* per a realitzar les anàlisis bioinformàtiques. L'elecció dels programes és complexa i en la majoria de casos no hi ha un *software* òptim per a un experiment

específic. És per això que cal realitzar proves de concepte i anar avaluant avantatges i inconvenients de cada *software* per tal d'implementar el protocol d'anàlisi més adient a la tecnologia i experiment que es realitza.

1.2 Anàlisi bioinformàtica en la prova de concepte. El perquè de l'algoritme "VIP+R amb visualitzacions a l'AVA"

En el moment d'analitzar les primeres carreres del GS Junior només hi havia un *software* no comercial disponible, el VIP (De Schrijver et al. 2010), específic per a les lectures llargues de la tecnologia 454, per a retallar els adaptadors, per a l'alineament, i per a la detecció de variants. Altres *software* d'alineament i detecció de variants no comercials com el Bowtie, el BWA o el GATK ja s'havien publicat i podien adaptar-se a les lectures de 454 mitjançant el canvi d'alguns paràmetres, però tenien algunes limitacions. L'alineador Bowtie per exemple, no permetia forats en l'alineament i això descartava la possibilitat de detectar insercions i delecions, la versió posterior Bowtie2 ja ho permet (Langmead and Salzberg 2012, John Hopkings University 2016). A més, aquests programes estaven més enfocats a l'anàlisi de lectures de fins a 70 bases i només permetien alinear lectures amb poques bases discordants, també, per a lectures llargues eren molt menys eficients en termes de bases alineades per unitat de temps (Li and Durbin 2010). Amb tot, l'avantatge principal del VIP era que ja tenia programat de manera eficient el processament de les lectures per a demultiplexar i tallar els adaptadors de les lectures seqüenciades amb la tecnologia 454, i a més, reportava correctament tot tipus de variacions, el que proporcionava una anàlisi més robusta. Tot i això, el VIP també presentava algunes limitacions, per exemple, no permetia visualitzar els alineaments fàcilment, i en les zones properes a homopolímers les bases de les lectures en cadena positiva i negativa s'alineaven diferent, deixant algunes variants reportades en cada cadena en posicions diferents. Aquest problema produeix molts falsos positius que cal visualitzar abans de descartar-los.

D'altra banda, els programes comercials tampoc no resolien les anàlisis bioinformàtiques correctament i fallaven en la detecció d'algunes variants patogèniques en la prova de concepte, especialment insercions i delecions, tal com es va provar amb el CLC-Workbench, actualment modificat i millorat. L'AVA (Roche 454 sequencing 2016), el *software* comercial de Roche incorporat en el GS Junior, tampoc no proporcionava uns resultats acurats, fallava en la detecció de delecions d'una base, i tampoc no proporcionava un mètode clar per controlar les zones que podien haver quedat amb baixa cobertura. Actualment Roche ha modificat i solventat en gran part aquestes limitacions de l'AVA, però encara és poc customitzable a l'hora de controlar les zones de baixa cobertura. A favor de l'AVA però, una de les seves fortaleses és el visualitzador dels alineaments, que permet veure clarament les zones que envolten les variants i comparar-les amb altres mostres per tal de descartar possibles falsos positius.

A fi d'arribar a la sensibilitat i especificitat requerides pel diagnòstic, vam posar en marxa el protocol d'anàlisi amb NGS complementant les anàlisis bioinformàtiques del VIP amb mètodes de visualització i

confirmació més manuals mitjançant l'AVA, i amb un kit de seqüenciació d'homopolímers, que cobrien les limitacions corresponents.

1.3 Limitacions de l'algoritme "VIP+R amb visualitzacions a l'AVA"

Malgrat l'assoliment d'una sensibilitat del 100% per a variants d'alta dificultat tècnica en la prova de concepte (article 1), durant l'ús de l'algoritme en rutina es va detectar un fals negatiu. Aquest fou causat per l'alineament erroni del VIP d'una deleció GT, degut a la proximitat d'un homopolímer i a la simetria de la seqüència on es troba (T-GT o TG-T), que situa en la deleció en diferents posicions a les lectures *forward* i *reverse*. Aquest fals negatiu va ser detectat en l'anàlisi específica d'homopolímers. Tal com es mostra a la taula 5, el VIP reporta la variant de *BRCA2* c.3847_3848delGT com una deleció de TGT a baixa freqüència, la qual cosa fa que posteriorment es descarti pels filtres. Per aquesta raó ens vam centrar en millorar l'anàlisi bioinformàtica i així evitar aquest tipus d'errors que es podrien repetir. A la figura 5 es pot veure com l'algoritme de la Web tool, explicat posteriorment, que utilitza un alineador i detector de variants diferents (BWA-MEM i VarScan), alineen les delecions sempre el màxim a l'esquerra possible de l'homopolímer.

Taula 5. Deleció reportada erròniament pel VIP i correctament pel VarScan utilitzat a l'aplicació web.

	Chrom	Position	Ref	Real	F_Cov	R_Cov	T_Cov	F_Var	R_Var	T_Var	T_Var_R_Freq	F_Surround	R_Surround
VIP	chr13	32912337	tgt	---	140	0	140	33	0	33	0.23571	ATGATAAAAC TGT AAGTGAAAAA	TTTTTCACTT ACA GTTTTATCAT
Web tool	chr13	32912337	C	-TG	215	203	418	110	112	222	0.53110	ATGATAAAACT GT AAGTGAAAAA	TTTTTCACTT AC AGTTTTATCAT

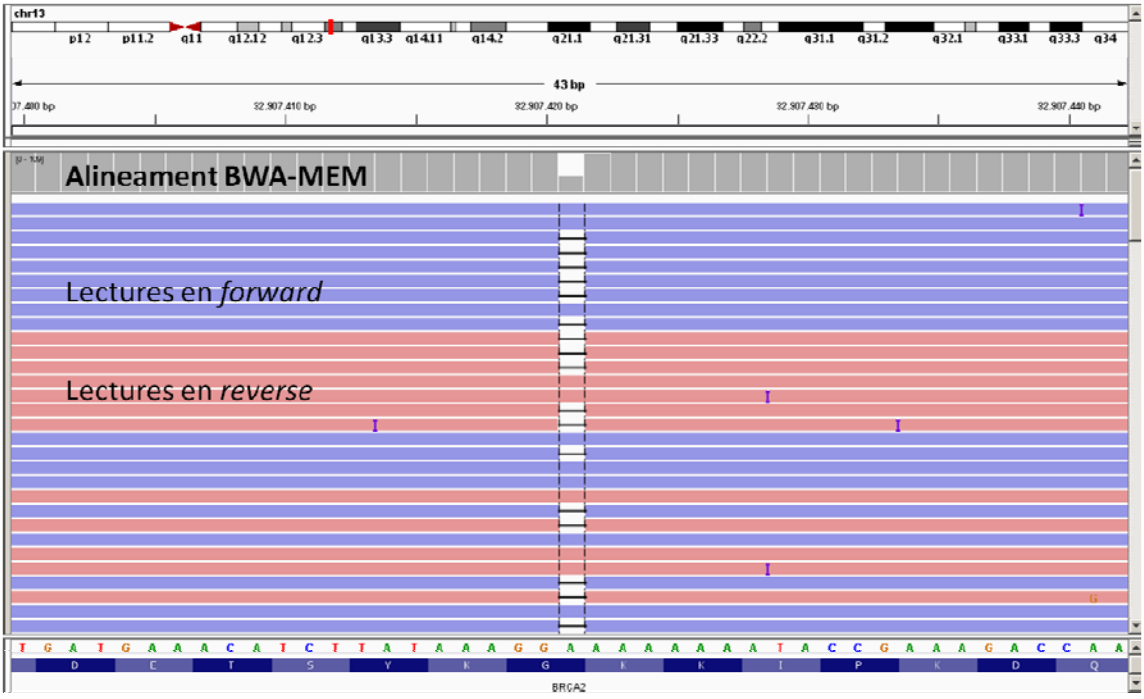


Figura 5. Visualització de la deleció d'un nucleòtid en un homopolímer després de l'alineament amb BWA-MEM. Tant en les lectures en *forward* (color blau), com les lectures en *reverse* (color vermell), alineen la deleció a l'esquerra de l'homopolímer.

1.4 Evolució de l'algoritme bioinformàtic a "BWA-MEM + VarScan amb CDR + R" i desenvolupament de l'aplicació web

Alineador BWA-MEM

En poc temps s'havien publicat alineadors nous que permetien alinear lectures llargues amb un alt rendiment, un d'ells era el BWA-SW, algoritme eficient per alinear lectures llargues però que falla en l'alineament dels extrems de les lectures. Això representa un problema important en les genoteques d'amplicons, on en la majoria de casos totes les lectures acaben en el mateix punt, doncs fàcilment pot provocar falsos negatius als extrems dels amplicons. El BWA-MEM, versió publicada poc després, a més de l'alineament global estàndard del BWA i l'adaptació a lectures llargues del BWA-SW, implementa un algoritme d'alineament local que permet alinear correctament els extrems dels amplicons tot i haver alguna variant en les primeres o últimes bases de l'amplicó. A la figura 6 es pot observar la diferència d'alineament entre el BWA-SW i el BWA-MEM a l'extrem d'un amplicó que conté una variant: el BWA-MEM alinea correctament l'extrem de l'amplicó detectant la variant mentre que el BWA-SW realitza un alineament global on es perden les primeres dues bases de les lectures amb la variant.

Figura 6. Visualització de l'alineament d'un amplicó que conté una variant en la segona base. S'observa com el BWA-MEM alinea correctament la primera base i marca la variant, en canvi, el BWA-SW retalla les primeres bases dels amplicons amb la variant i aparenta una baixada de cobertura, característica que la majoria de detectors de variants no tenen en compte.

Després de comprovar la robustesa de l'alineador BWA-MEM, es va canviar l'algoritme d'alineament en el protocol d'anàlisi, passant del BLAT que utilitza el VIP a l'algoritme BWA-MEM.

Detector de variants VarScan

Després de canviar l'algoritme d'alineament, el *software* per a la detecció de variants també convenia actualitzar-lo per facilitar la compatibilitat de formats i per adaptar els paràmetres als nous alineaments. Les aplicacions per a la detecció de variants basades en mètodes Bayesians permeten eliminar molts falsos positius, però també descarten algun veritable positiu, ja que les probabilitats es veuen afectades per cobertures molt altes, o per contaminació en les mostres (Koboldt et al. 2012). Tot i estar basat en models probabilístics, el programa de detecció de variants GATK donava bons resultats, arribant a sensibilitats altes properes al 100%. GATK, però, és un *software* més enfocat a l'anàlisi d'exomes seqüenciats amb Illumina, i tot i que és possible adaptar les funcions al format de les lectures de 454, és molt exigent amb el format d'entrada i aquestes restriccions fan que la seva execució sigui complicada (McKenna et al. 2010, DePristo et al. 2011). Degut a que pel diagnòstic clínic la sensibilitat ha de ser el més alta possible, no es pot permetre cap font confirmada de falsos negatius. És per això que es va triar un mètode de detecció de variants empíric, el VarScan2 (Koboldt et al. 2012), que reporta totes les variants presents. Així, s'ha focalitzat més l'anàlisi en el filtratge posterior basat en paràmetres com la cobertura, la qualitat de les bases o la freqüència de la variant. En relació amb altres detectors de variants, Xu et al. en la seva comparativa situen l'algoritme de VarScan2 amb una alta sensibilitat en l'anàlisi de detecció de variants en amplicons on la puresa de les mostres és alta, la millora de VarScan2 respecte la primera versió de VarScan recau principalment en la possibilitat d'executar-se en qualsevol sistema operatiu i en un algoritme més eficient per a la detecció de variants somàtiques i estructurals. Altres *software* com MuTect o Strelka també mostren una alta precisió, però aquests tenen en compte la correlació amb els controls normals quan es tracta de cercar mutacions somàtiques en teixit tumorals. En aquest cas de detecció de variants en línia germinal, aquesta correcció no s'aplica i, per tant, l'algoritme de VarScan2 dóna resultats molt satisfactoris per al nostre objectiu (Xu et al. 2014).

Coverage Difference Ratio (CDR)

Quan l'alineament no es realitza sobre cada amplicó, on es pot forçar l'alineament del encebadors i tallar-los després, sino sobre tot el gen tal com fa el BWA-MEM en el nostre protocol d'anàlisi, es dóna el cas que delecions llargues situades al final o al principi d'amplicons no es detecten amb el VarScan, ja que es considera com si l'amplicó fos més curt (Figura 7), és per això que es va treballar un algoritme programat en R per a detectar canvis de cobertura considerables en posicions on no acaba ni comença un amplicó i així detectar possibles delecions o insercions als extrems de les regions seqüenciades. L'algoritme, que vam anomenar CDR (*Coverage Difference Ratio*), ha mostrat ser consistent amb les carreres que s'han anat analitzant de rutina. Tot i que reporta els mateixos falsos positius que Varscan en zones d'homopolímers llargs o properes a aquests (són les pròpies lectures les que porten les falses delecions, degut a les limitacions de la tècnica), confirma veritables delecions també detectades pel

VarScan. Fins a dia d'avui el CDR no ha detectat cap deleció en el diagnòstic de rutina que no hagués reportat el VarScan, però s'espera que en cas d'aparèixer en alguna mostra la reporti correctament tal com va fer amb la deleció trobada a la prova de concepte.

Figura 7. Visualització de l'alineament d'una mostra amb una deleció de 19 nucleòtids (adalt) i d'una mostra sense deleció. En la figura s'aprecia com es tallen els amplicons en diferents punts i com la funció CDR pot detectar-la per la diferència no esperada de cobertura.

R per a presentar els resultats

L'entorn i llenguatge R és una eina molt potent que permet programar les funcions que requereixi l'anàlisi per anotar els resultats amb un format personalitzat. Amb R s'han pogut programar les funcions per generar els diagrames de barres normalitzats en base a la cobertura dels amplicons (Figura 3D de l'article 2), les funcions per al filtratge de les variants, així com per a l'anotació de les variants segons la nomenclatura en cDNA requerida per la Human Genome Variation Society (HGVS 2016), amb significat biològic i directament exportable als informes genètics.

Pàgina web

La pàgina web es va desenvolupar amb PHP i HTML, que són els llenguatges de programació més utilitzats per a aquest context. Amb el desenvolupament de l'eina a través de la pàgina web es permet l'accés i execució a altres usuaris de tot el món.

L'aplicació està programada amb *software* lliure, aquest mostra com evolucionen els algoritmes i permet un desenvolupament més ràpid. El fet que els codis siguin accessibles a tota la comunitat investigadora fa que es puguin adaptar a les necessitats dels diferents experiments. Així, en el disseny de l'aplicació, hem pogut utilitzar codi de diferents programes i adaptar-los als nostres requeriments. D'aquesta manera hem pogut desenvolupar noves funcions com la CDR o funcions per a detectar les regions mal cobertes, que permeten trobar mutacions amb tècniques complementàries en regions on la

sensibilitat estaria més limitada i proporcionar un diagnòstic més sensible. De la mateixa manera que hem utilitzat el codi de programes lliures, a la documentació de l'aplicació web presentada al segon article està disponible el codi utilitzat, de manera que altres usuaris poden adaptar-lo a les seves necessitats.

1.5 Utilització de l'aplicació web dos anys després de la seva publicació

Després de dos anys des de la publicació de l'aplicació web, s'havien analitzat des de fora de l'ICO un total de 158 carreres, 89% per a l'anàlisi de BRCAs, 2% per a l'anàlisi de FAP i 9% per a l'anàlisi d'HNPCC. El percentatge d'anàlisis finalitzades amb èxit és del 73%, mentre que 43 anàlisis han donat errors, principalment pel format amb què els usuaris van codificar els identificadors, que ha de ser exactament com el del model proporcionat. En la majoria dels casos d'error, els usuaris s'han posat en contacte amb l'administrador de la web a través de la pàgina de contacte i se'ls han donat instruccions de com solucionar-los. En algun cas l'usuari ha comès el mateix error en el format però ell mateix l'ha corregit.
D'altra banda, els usuaris que han obtingut els resultats satisfactòriament han repetit amb l'anàlisi de noves carreres. No s'ha pogut identificar l'origen dels usuaris, però excepte en dos casos, tots els usuaris que s'han posat en contacte amb nosaltres a través de la web, per a consultar diversos temes, treballaven a Espanya, només un usuari escrivia des de Croàcia i un altre des de Turquia. A la figura 8 es pot observar la distribució del nombre de carreres analitzades per mes, des del moment de la publicació fins a la finalització de l'escriptura d'aquesta tesi, sent el mes de maig de 2014 quan més anàlisis s'han realitzat a l'aplicació.

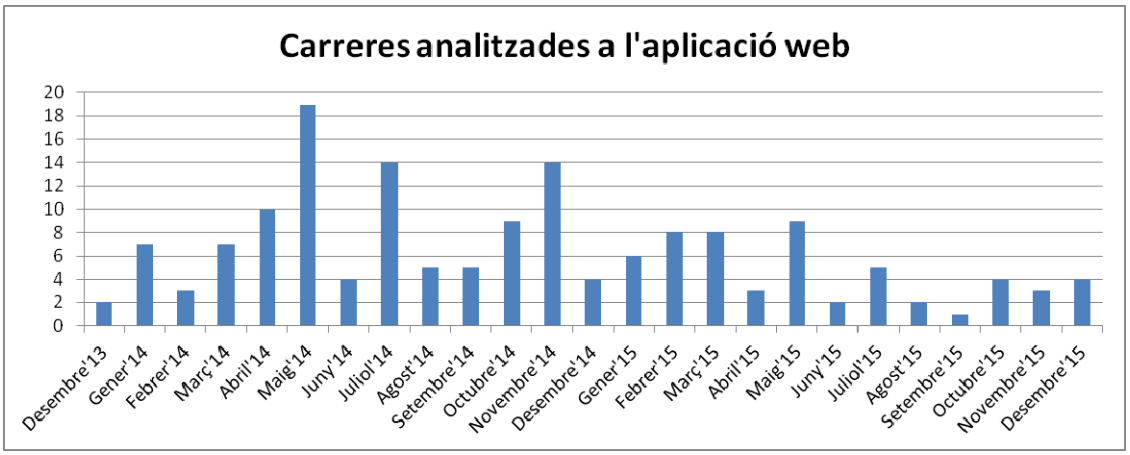


Figura 8. Gràfic de barres on cada barra representa el número de carreres analitzades a l'aplicació web per cada mes.

La companyia Roche, a nivell nacional, es va posar en contacte amb nosaltres al febrer de 2013 per avaluar el protocol d'anàlisi, i després d'avaluar-lo positivament, ens va demanar poder incorporar la documentació de l'aplicació en la formació de nous usuaris del GS Junior, després de la publicació, per a donar una alternativa al seu programa AVA.

També es va sol·licitar per part d'un usuari l'adaptació de l'aplicació a l'anàlisi de dades d'altres tecnologies, Ion Torrent i MiSeq, en aquests casos se'ls va facilitar el codi modificat per a adaptar l'anàlisi a les seves dades i es van proporcionar els resultats d'una carrera. Aquestes modificacions però, no es van implementar a la pàgina web.

1.6 Carreres analitzades en rutina a la Unitat de Diagnòstic Molecular de l'ICO

A més de l'eina pública, s'ha adaptat l'aplicació a les necessitats específiques de la nostra unitat de diagnòstic. Els principals canvis han estat: afegir la informació dels encebadors necessaris per confirmar mitjançant seqüenciació Sanger les variants trobades per NGS i els fragments mal coberts, i també afegir l'identificador del DNA. L'aplicació específica per a la nostra unitat s'ha anomenat "ICO Amplicon NGS Data Analysis v2". Amb aquesta eina, equivalent a l'aplicació oberta al públic, s'han analitzat totes les carreres de la unitat: 83 de BRCAs, 4 de FAP, 6 de HNPCC i 3 de FAP+HNPCC combinades. A la figura 9 es representa la distribució de carreres analitzades a la unitat. Des del desembre de 2012 fins al desembre de 2013 es van analitzar amb el protocol basat en VIP, posteriorment es van analitzar amb l'aplicació web específica per a l'ICO fins al desembre del 2014. Ja a partir del setembre de 2014 es van començar a estudiar les diferents opcions per a l'aplicació dels panells de gens i la corresponent anàlisi bioinformàtica. A partir de gener de 2015, quan es va migrar a la tecnologia Illumina, utilitzant la plataforma MiSeq, les anàlisis bioinformàtiques es van realitzar amb l'aplicació comercial SeqNext, mentre es dissenyava una aplicació que permetés analitzar de manera òptima el panell de gens, actualment ja en ús.

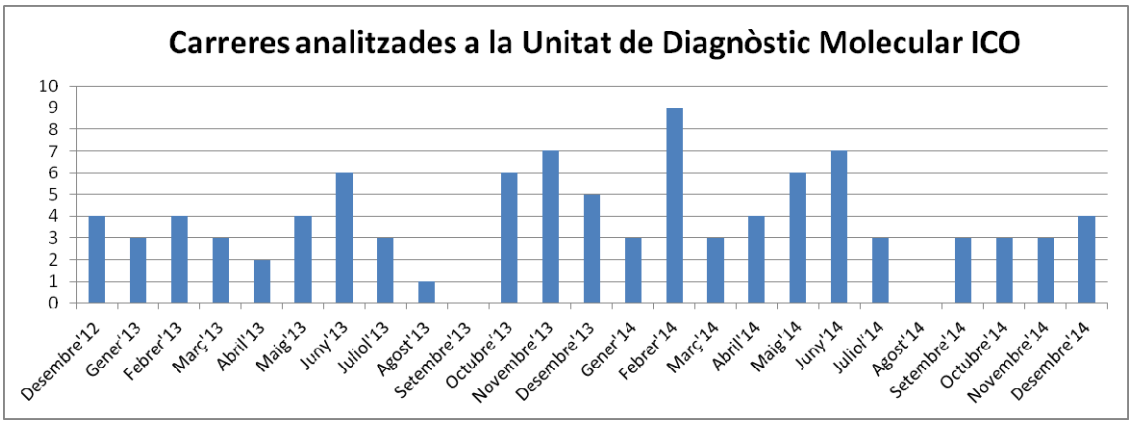


Figura 9. Gràfic de barres representant el número de carreres analitzades a la Unitat de Diagnòstic Molecular de l'ICO cada mes durant el període de Desembre de 2012 fins al Desembre de 2014.

1.7 Avaluació interna de l'aplicació web a l'ICO

Com a control de qualitat en la rutina diagnòstica, s'ha realitzat l'anàlisi bioinformàtica mitjançant els protocols publicats tant al primer article (VIP+R) com al segon (Aplicació web), d'un total de 50 carreres tant dels gens de la síndrome de HBOC com dels gens responsables de CCR hereditari. En total s'han analitzat 350 mostres per les dues metodologies. De les 13186 variants identificades, 6667 (51%) eren comunes en els dos protocols, 3114 (24%) eren específiques del VIP+R, i 3405 (26%) només les reportava l'aplicació web. De les variants només reportades pel VIP+R, el 100% eren falsos positius (segons la posterior inspecció visual i, en algun cas, la seqüenciació Sanger), mentre que l'aplicació web ha reportat dos veritables positius que el VIP no reportava correctament i que són:

- A) *MLH1* c.1210dupC
- B) *BRCA2* c.3847_3848delGT

Ambdues són a causa del mal alineament del VIP, ja que alinea en diferents posicions les lectures en *forward* i en *reverse*. En el primer cas, el VIP reporta la variant com una inserció ACC però amb una freqüència de 0,20 i, per tant, es filtra. En el segon cas, explicat en un apartat anterior, l'error té la mateixa causa i el VIP reporta una deleció TGT però amb una freqüència de 0,23. L'alineador BWA-MEM, quan hi ha una deleció o inserció, alinea totes les bases a l'esquerra, així les variants es reporten correctament i amb la freqüència real.

En relació al nombre de falsos positius, l'aplicació web reporta un 2% més de falsos positius, però entre ells estan incloses les variants reportades per la funció CDR que són un 8% (276) del total, aquesta funció dóna un gran nombre de falsos positius al voltant dels homopolímers ja que aquests s'alineen tots al principi, i quan perden senyal sembla que hi hagi un fals canvi de cobertura. L'error no és de la funció CDR sinó de la piroseqüenciació d'homopolímers, que genera un elevat nombre de delecions que reporten tant VarScan com la CDR. Per tant, descartant aquests falsos positius, fàcils de detectar per ser comuns en quasi totes les mostres i trobar-se al voltant d'homopolímers, el nombre de falsos positius de l'aplicació és menor que els que reporta el VIP, és a dir, millora l'especificitat. Quan mirem els falsos positius comuns, bàsicament es tracta d'errors de seqüenciació a causa dels homopolímers. Amb tot, l'aplicació dóna la sensibilitat i especificitat requerides per al diagnòstic i millora el primer algoritme basat en el VIP. La figura 10 representa la situació de les variants en la comparació entre els dos protocols d'anàlisi.

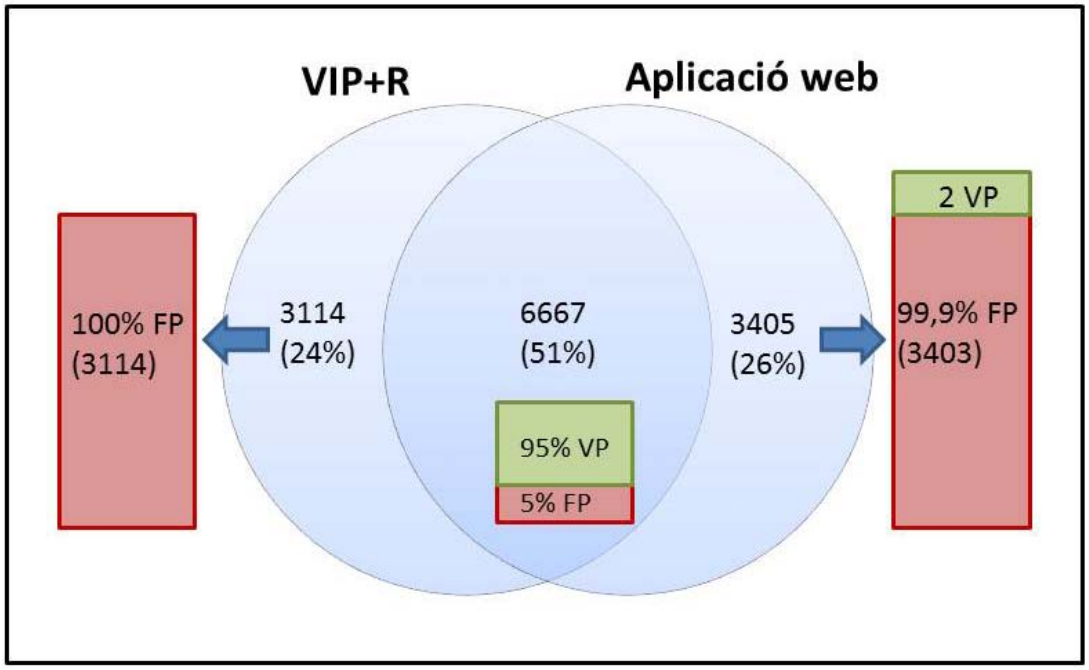


Figura 10. Diagrama de Venn representant les variants comunes i les específiques trobades amb l'anàlisi del VIP+R i l'aplicació web. Amb requadres s'especifica el número i percentatge de falsos positius (FP) i veritables positius (VP) per a cada secció del diagrama.

1.8 Limitacions i futur de l'aplicació

Una de les limitacions del nostre protocol d'anàlisi presentat en el segon article és que permet l'anàlisi d'uns gens particulars seqüenciats a partir d'unes genoteques específiques de Multiplicom per a la generació d'amplicons. L'eina presentada podria evolucionar, no només amb l'adaptació a altres plataformes com Illumina o IonTorrent, sinó també amb l'opció de poder analitzar altres gens i adaptar la nomenclatura completant l'anotació amb altres camps com l'efecte predit, la patogenicitat segons bases de dades clíniques, o la freqüència de les variants a poblacions generals, entre d'altres, segons els interessos de l'usuari.

S'ha valorat l'opció de treballar en aquesta direcció, però amb la gran velocitat amb què avancen aquestes tecnologies i la bona sortida que estan tenint els panells de gens, s'ha optat per no invertir esforços en aquesta eina, i a canvi, dedicar-los a desenvolupar una eina més potent que permeti analitzar les dades resultants de seqüenciar panells de múltiples gens. Tot i això, gran part del codi programat, i sobre tot l'experiència i aprenentatge, ens han donat la capacitat d'abordar anàlisis més complexes. Tot i ser necessària l'adaptació de les funcions i dels paràmetres a les propietats de les lectures, no ha sigut molt laboriós adaptar el protocol d'anàlisi per a tractar dades d'Illumina, tal com s'ha fet per a l'anàlisi del panell Trusight Cancer, comentat en l'apartat 2.1, utilitzant tant les aplicacions disponibles al BaseSpace d'Illumina com l'aplicació comercial SeqNext (JSI).

1.9. Limitacions actuals del diagnòstic genètic amb NGS mitjançant la tecnologia 454

Com hem vist a la secció de "NGS per al diagnòstic genètic" que inclou els dos primers articles, el diagnòstic genètic amb NGS no és trivial, i requereix una anàlisi molt minuciosa. Encara calen millores per assegurar una fiabilitat més alta en els resultats, i per evitar la necessitat de cobrir les limitacions amb altres procediments o anàlisis. Amb això fem referència especialment a alguns artefactes de la tecnologia 454 com l'alta imprecisió de lectura en la longitud dels homopolímers i els errors en les bases posteriors, que necessiten millores substancials per a poder seguir avançant en l'optimització de la tècnica. La sensibilitat ha de seguir acostant-se al 100%, però l'especificitat també hauria de ser propera al 100% abans de la validació per una altra tècnica com la seqüenciació Sanger; això reduiria molt els costos econòmics i de supervisió manual.

Els falsos positius durant la detecció de variants, tant per substitucions com per insercions i delecions, generalment són causats per dos fenòmens. El primer és l'error de seqüenciació. En les plataformes d'Illumina l'error de seqüenciació correlaciona positivament amb la posició en la lectura, els errors tendeixen a estar cap al final de les lectures (Ozsolak 2012). En les plataformes de 454, els errors de seqüència acostumen a estar durant i després dels homopolímers, que fallen en la piroseqüenciació. L'altra font de falsos positius són els artefactes en l'alineament, ja que alinear lectures curtes sobre una seqüència de referència complexa com és la del genoma humà no és trivial. Un exemple de gen problemètic és *PMS2*, que té molts pseudogens d'altíssima homologia. Les lectures curtes del pseudogen s'alinien a *PMS2*, de manera que es poden cridar variants dels pseudogens asignant-les a *PMS2*, i perdre variants del gen funcional degut a la disminució de la fracció de les seves lectures en barrejar-se amb les del pseudogen (Brea-Fernandez et al. 2014, Wimmer and Wernstedt 2014). Aquesta dificultat es podria millorar amb un enriquiment de la llibreria més específic del gen funcional mitjançant PCRs més llarques, o també afegint a l'anàlisi protocols d'alineament més astringents només per al gen específic.

Una altra dificultat en l'anàlisi de les dades de NGS és la detecció de variacions estructurals, que requereix una anàlisi específica i acurada. Actualment s'està treballant bioinformàticament per identificar aquest tipus de variants emprant dades de NGS. Alguns programes com VarScan2, Pindel, o una nova llibreria d'R anomenada *ExomeDepth* permeten detectar canvis en el número de còpies, però no hi ha un programa que permeti la detecció de tots els tipus de variants estructurals amb una alta precisió. Encara hi ha molts reptes en la detecció d'aquestes variants, ja sigui per les limitacions de les tecnologies de NGS, les dificultats en la seva reconstrucció o els mètodes per inferir aquest tipus de variants (Liu et al. 2015).

En el primer article d'aquesta tesi es proposa una normalització de la cobertura dels amplicons. La normalització es basa en dividir la cobertura de cada amplicó pel número total de lectures de cada MID, per la mitjana de lectures dels amplicons de la seva multiplex i per la mitjana de cobertura dels amplicons dels altres gens del kit. Així es pretén detectar les possibles diferències en el número de lectures entre amplicons degudes a variacions estructurals. En la prova de concepte es va veure que era possible detectar les delecions i insercions correctament però que quedaven molts falsos positius, i per

tant es segueix realitzant el test de *Multiplex Ligation Probe Amplification* (MLPA) per a detectar variacions estructurals. Possiblement amb una seqüenciació més homogènia i un algoritme d'anàlisi més refinat, incorporant models probabilístics basats en els resultats de moltes carreres, es podrien aconseguir uns resultats més precisos.

2. Futur del diagnòstic genètic del càncer hereditari amb les noves possibilitats de la NGS

En els dos primers articles s'han analitzat dades de NGS provinents de la tecnologies 454, en el breu apartat d'altres contribucions sobre panells de gens ja s'utilitza la tecnologia Illumina i es veu que l'ha superat en gairebé tots els aspectes. Ara per ara Illumina domina el mercat de la NGS, a petita escala amb instruments de mitjà rendiment com el MiSeq, i a gran escala amb instruments com el HiSeq, que permet seqüenciar exomes i genomes complets o l'X-ten, que seqüencia genomes per 1000$ (Illumina 2016). La feblesa de la tecnologia d'Illumina, que era la poca llargada de les lectures, ha evolucionat positivament i actualment es seqüencien lectures de 300 bases, tot i que encara més curtes que la llargada actual del GS FLX, de lectures de 1000 bases. Però el principal punt fort d'Illumina és que gairebé no presenta el problema d'errors de seqüència al voltant dels homopolímers, que causen molts falsos positius i algun fals negatiu, i que a les plataformes de Roche i Ion Torrent disminueixen la sensibilitat i especificitat o fan necessàries tècniques complementàries per a cobrir aquesta limitació.

2.1 Panells de gens per al diagnòstic genètic

Amb l'evolució dels mètodes de preparació de genoteques i de la NGS, cada cop es fa més factible i barat seqüenciar més gens de diversos individus en una sola carrera. A més cada vegada hi ha més gens identificats associats a risc de càncer i es defineixen millor aquests riscos, permetent l'aplicació clínica de la seqüenciació de molts més gens. Així, les anàlisis actuals dels gens *BRCA1* i *BRCA2* per a pacients amb susceptibilitat al càncer hereditari de mama i ovari, o els corresponents gens per a càncer de còlon hereditari, s'estan incloent en anàlisis més àmplies que engloben un conjunt més gran de gens. Tal com es mostra en el breu apartat d'altres contribucions en la secció de resultats, el diagnòstic genètic està evolucionant cap a l'anàlisi de panells de gens, això és, analitzar una sèrie de gens relacionats amb vàries malalties, i que es pugui aplicar la mateixa ruta analítica (amb la mateixa regió d'interès i per tant els mateixos reactius) de manera rutinària a qualsevol pacient de risc que acudeixi a la unitat de consell genètic, compartint així una sola tanda de treball i optimitzant recursos i temps de resposta (Hall et al. 2014, Lapunzina et al. 2014). Les cases comercials, així com diversos laboratoris privats com Myriad, Ambry Genetics o Sistemas Genómicos, entre d'altres, estan apostant pel disseny de panells amb desenes de gens associats a determinades malalties (Easton et al. 2015).

En el treball sobre panells d'aquesta tesi s'avalua el rendiment del panell "Trusigth Cancer" d'Illumina, i tant la cobertura com la detecció de variants mostren resultats satisfactoris. Per a l'anàlisi bioinformàtica s'ha utilitzat el suport de BaseSpace (Illumina BaseSpace 2016), un núvol de computació que permet dissenyar l'experiment i organitzar les mostres, està vinculat al seqüenciador, i permet emmagatzemar les dades i analitzar-les amb els *software* que tenen implementats. També s'ha testat el *software* comercial SeqNext (JSI): els resultats són molt acurats, essent el cost econòmic la principal limitació pel seu ús. L'anàlisi bioinformàtica amb BaseSpace és fàcil i intuïtiva però cal que es realitzi pas a pas tal com està pensat, no permet fer anàlisis complementàries ni canviar paràmetres o filtres. Tampoc no permet fàcilment canviar l'algoritme d'algun dels passos establerts, és a dir, en cas que l'usuari vulgui canviar l'algoritme d'alineament o de detecció de variants, per exemple, no existeix un mètode fàcil per a descarregar els arxius de seqüències i tornar a carregar els arxius d'alineaments o de variants i seguir amb el protocol establert. Aquesta limitació fa que l'anàlisi de validació de variants posterior sigui més costós ja que el nombre de variants reportades és més alt, degut als filtres i paràmetres menys específics. L'anàlisi amb SeqNext requereix una petita formació però també és intuïtiva. En l'anàlisi realitzada en aquesta tesi s'han considerat totes les variants reportades, però com s'ha comentat prèviament, SeqNext classifica les variants amb un indicador del nivell de confiança de ser falsos positius, si es realitza un estudi amb més mostra seria possible validar el nivell de confiança adequat per a descartar falsos positius, això reduiria considerablement el nombre de variants a validar mitjançant visualització i Sanger. Actualment s'està treballant a nivell bioinformàtic per a dissenyar un protocol d'anàlisi que permeti analitzar de manera acurada i personalitzada, qualsevol tipus de panell de gens així com l'exoma sencer.

2.2 Organització dels serveis de diagnòstic genètic amb la implementació de la NGS

En molts casos els avenços tecnològics superen la capacitat i els recursos que una institució hi pugui destinar. Avui dia per exemple, Illumina ha tret al mercat el seqüenciador HiSeq X Ten, que té el rendiment de 10 seqüenciadors HiSeq i que permet seqüenciar 18.000 genomes en un any amb una cobertura raonable, i un preu de $1000 per genoma (Pennisi 2014, Illumina 2016), però també amb una inversió considerable. És per això que el model que sembla més adient tendeix a l'externalització de la seqüenciació a centres especialitzats que puguin concentrar l'activitat de seqüenciació i excel·lir en la tècnica. A més, en la majoria de laboratoris, els costos d'externalitzar la seqüenciació compensen els costos d'amortització i de manteniment que requereixen aquests aparells de seqüenciació cars i complexos. També cal tenir present que la capacitat per a analitzar les dades de manera òptima podria ser una limitació. Cal que els avenços en els diferents camps, tant tecnològics com bioinformàtics, es produeixin en paral·lel. Per això, en el camp de la bioinformàtica, és important que les grans bases de dades que guarden informació sobre seqüències, estructures, variants, genotips, etc., segueixin evolucionant per a poder extraure'n informació de qualitat fàcilment. També hauran d'evolucionar el *software* i protocols d'anàlisi, per a que continguin potents controls de qualitat i que reportin allò realment rellevant segons l'interès de l'estudi.

És important la formació d'equips multidisciplinaris amb diferents habilitats i coneixements. La seqüenciació d'una major quantitat de gens implica la identificació d'un gran nombre de variants amb significat desconegut i és necessari el treball en equip de clínics, biòlegs, informàtics i bioinformàtics per a la correcta interpretació.

Com s'ha comentat anteriorment, actualment el nostre grup treballa en el desenvolupament d'una aplicació per a l'anàlisi de les dades resultants de la seqüenciació dels panells de gens. Un dels objectius de l'aplicació, a més de l'anàlisi de la cobertura i la qualitat i la detecció de les variants, és donar informació sobre la variant recollida en altres bases de dades, així com la recurrència amb que s'ha anat detectant en les mostres analitzades segons la sospita clínica i la classificació que se n'ha fet a la Unitat de Diagnòstic. El fet d'analitzar un gran nombre de gens requereix una aplicació complexa i robusta.

3. Anàlisi bioinformàtica dels exomes

En la segona secció de la tesi doctoral, "NGS aplicat a la recerca del CCR esporàdic", s'ha utilitzat la tecnologia NGS per a la seqüenciació d'exomes amb l'objectiu de detectar noves mutacions recurrents implicades en la tumorigènesi en tumors colorectals d'estadiatge II. Per a aquest treball s'han seqüenciat els exomes de 42 tumors i les corresponents mucoses normals amb la plataforma de gran capacitat Illumina HiSeq 2000. Els exomes de tumors es van seqüenciar amb una cobertura de 60X i els exomes de les mucoses normals a 40X**.** La raó per la qual es va seqüenciar el tumor amb més cobertura és que la proporció amb que s'espera trobar les variants és molt heterogènia: així com en sang s'espera trobar les variants aproximadment en un 100% o un 50% de les lectures, segons si és una variant en homozigosi o en heterozigosi respectivament, les mutacions somàtiques en teixit tumoral poden ser

presents en diferent proporció de cèl·lules i, per tant, es poden trobar en diferents percentatges. El mateix succeeix amb els teixits de mucosa normal, però s'espera menor heterogeneïtat que en el tumor.
La caracterització de les mutacions en els gens *KRAS, APC, BRAF* i *TP53* dels 42 tumors mitjançant altres tècniques ens va proporcionar controls positius i negatius per validar el protocol d'anàlisi i per poder optimitzar els filtres aconseguint la màxima sensibilitat i especificitat. Així, per a l'anàlisi bioinformàtica es va realitzar l'alineament amb el Bowtie2. Es va decidir utilitzar el Bowtie2 ja que la llargada de les lectures era menor que la del GS Junior (70 parells de bases), i aquest algoritme és ràpid, eficient en memòria, i robust per a lectures curtes. Després de l'alineament global del Bowtie2 es va realitzar un realineament local al voltant de les insercions i delecions conegudes en el dbSNP i en el 1000G, projecte que d'entre altra informació reporta totes les variants trobades en 1000 genomes (EMBL-EBI 2008), i també al voltant de les insercions i delecions trobades en les 42 mostres analitzades. Aquest pas es va realitzar per evitar falsos positius al voltant de les delecions i insercions causades per un mal alineament. Tot i això, després del realineament local no es van observar grans diferències en el número de variants. Per a la detecció de variants es va utilitzar el GATK, ja que s'ha vist que proporciona una gran sensibilitat i especificitat en la detecció de mutacions somàtiques (Xu et al. 2014). Després es va fer un filtratge exhaustiu, es van filtrar els milers d'insercions i delecions trobades ja que introduïen molt soroll als resultats i era molt difícil abordar aquesta informació, el nombre de variants detectades era massa alt per a poder-les confirmar mitjançant tècniques alternatives i es va optar per perdre els possibles veritables positius. També es van filtrar totes les variacions presents en línia germinal, és a dir, que estaven en les mucoses normals, ja que es buscaven només variants somàtiques, i totes aquelles reportades al projecte dels 1000G i que es podrien haver obviat en el filtratge anterior per haver quedat amb baixa cobertura en les mucoses normals. D'aquesta manera es va reduir considerablement el nombre de variants focalitzant l'atenció a aquelles que podrien tenir més rellevància. Un estudi pendent és analitzar totes les insercions i delecions amb mètodes estadístics que permetin descartar un gran nombre de falsos positius.
En el centre on es va realitzar la seqüenciació, també es va efectuar l'anàlisi bioinformàtica emprant l'alineador GEM i la detecció de variants amb Samtools, però els filtres aplicats eren molt restrictius i no es reportaven algunes de les variants control conegudes. Per això es va optar pel nostre protocol d'anàlisi que va resultat més acurat.

3.1 Capacitat informàtica per a l'anàlisi d'exomes

El servidor de computació on s'han realitzat les anàlisis bioinformàtiques compta amb un total de 164 nuclis repartits en 13 nodes. A la figura 11 es representa el clúster de computació i les seves prestacions.

Figura 11. Pàgina de control del servidor de la unitat on es representa l'ocupació dels nodes. Es veu com el node 10 està treballant a més del 75% de la seva capacitat.

Per a l'alineament d'un exoma a 40X amb Bowtie2 es requereixen aproximadament 30 minuts utilitzant un processador AMD Opteron 6272 amb 16 nuclis a 2,1 GHz. Amb aquestes mateixes prestacions, els processos intermitjos de transformació de formats de *.sam* a *.bam* i ordenament de l'alineament amb Samtools requereixen uns 30 minuts més per exoma. Per a la detecció de variants d'un exoma amb GATK es requereixen aproximadament 3 hores utilitzant 30 nuclis i 120Gb de RAM, tot i que si no es disposa de tants nuclis es pot realitzar l'anàlisi correctament però amb més temps. Tant els algoritmes d'alineament com de detecció de variants estan evolucionant per a optimitzar el temps d'anàlisi, per a permetre paral·lelitzar els processos i així millorar el rendiment. En centres on es realitza l'anàlisi de centenars de mostres diàriament, l'optimització del temps d'anàlisi és molt important. Si no es disposa de bones prestacions informàtiques, les anàlisis de dades de seqüenciació es veuen molt limitades.

3.2 El gen *AMER1* recurrentment mutat en càncer colorectal esporàdic

Els resultats de l'anàlisi bioinformàtica dels exomes mostren una alta heterogeneïtat mutacional entre tots els tumors. Els resultats indiquen que la gran majoria de variants són úniques, amb un percentatge molt baix (3%) de mutacions recurrents en diferents tumors. Tot i l'alta heterogeneïtat mutacional, les mutacions en el gen *AMER1* (també anomenat *FAM123B* o *WTX*) apareixen com a recurrents (en un 10% dels tumors), proposant que el rol com a possible gen *driver* (aquell que pot contenir mutacions que confereixin un avantatge selectiu a un clon en el seu microambient) en el càncer colorectal pot ser rellevant. El *software* MutSig, que analitza llistes de mutacions per a identificar gens que estan mutats

més freqüentment del que s'esperaria per atzar (Lawrence et al. 2013), situava el gen *AMER1* (*APC* Membrane Recruitment Protein 1) entre els primers gens funcionals juntament amb els ja coneguts *TP53, APC* i *KRAS*. La proteïna codificada per aquest gen interacciona amb diverses proteïnes i regula positivament l'activació transcripcional mitjançant la proteïna del tumor de Wilms (*WT1*). Aquest estudi ha trobat que els tumors amb mutacions en el gen supresor *AMER1* presenten un fenotip mesenquimal caracteritzat per la inhibició de la via de Wnt canònica. Tot i que en altres estudis ja s'havia associat el gen *AMER1* a CCR (Seshagiri 2013), amb aquest treball es remarca la importància que pot tenir aquest gen supressor de tumors en el desenvolupament d'alguns tipus de càncer colorectal esporàdic.

3.3 Anàlisi bioinformàtica per la validació dels resultats dels exomes amb mostres del TCGA

La validació *in silico* de la freqüència de les mutacions d'*AMER1* es va realitzar amb l'anàlisi de 239 exomes de tumors de càncer colorectal i 100 mucoses normals del projecte TCGA. Aquesta anàlisi va ser més costosa computacionalment i es van haver d'utilitzar al màxim les prestacions del servidor de computació que tenim a la unitat.

A més, els alineaments descarregats del projecte del TCGA estaven fets sobre seqüències de referència del genoma variades segons els centres on s'havien seqüenciat, i per tant tenien zones no homogènies. Es va optar per passar de l'alineament descarregat en format *.bam* de totes les mostres, a les lectures originals en format *fastq,* i analitzar-les de nou amb el mateix protocol que els 42 tumors i mucoses normals de la nostra sèrie.

En total, per a l'anàlisi bioinformàtica de la validació amb les mostres del TCGA es van necessitar 339*1/2 hores per a l'alineament, 339*1/2 hores per al processament de formats, i 339*3 hores per a la detecció de variants, tot suma un total de 1356 hores (56 dies). Tot i això, en el moment de l'anàlisi es disposava de nodes lliures que van possibilitar l'execució de varis processos en paral·lel reduint el temps real de càlcul.

Els resultats de l'anàlisi bioinformàtica de la sèrie de tumors del TCGA mostren que només el 3.6% de les variants somàtiques trobades en la sèrie dels 42 tumors es van trobar també en els tumors del TCGA, confirmant així l'alta heterogeneïtat mutacional del càncer colorectal. En relació al gen *AMER1*, 25 dels 239 (10.5%) tumors analitzats del TCGA acumulaven un total de 26 mutacions somàtiques diferents, això confirmava *AMER1* com a gen candidat associat al càncer colorectal esporàdic.

3.4 La seqüenciació d'exomes per a la recerca del càncer

La seqüenciació d'exomes és una eina útil per a trobar mutacions, gens i vies proteiques encara desconegudes que expliquin millor els mecanismes de la carcinogènesi en general i del càncer colorectal en particular. Tot i això, la tècnica encara té limitacions. Les mutacions es detecten millor en zones amb una alta cobertura i per tant els extrems dels exons capturats poden tenir sensibilitats més baixes. A més, la captura de l'exoma no cobreix tots els exons del genoma amb la mateixa eficiència. Com hem comentat prèviament, l'heterogeneïtat del tumor i la contaminació estromal s'han de tenir presents a l'hora de calcular la freqüència de les variants per a diferenciar les mutacions amb baixa freqüència del

soroll per artefactes tècnics. En el nostre algoritme de detecció de variants era necessari considerar un mínim de cobertura de 10X o freqüències del 10% per a considerar una variant com a real. Es van posar aquests filtres poc restrictius per a poder descartar el màxim possible de variants germinals que sinó no es consideraven en les mucoses normals per tenir una cobertura més baixa. A major cobertura es poden detectar variants amb menor freqüència.

Amb la gran quantitat de variants somàtiques que s'obtenen és necessari un bon filtratge. Tot i el filtratge més o menys acurat que s'apliqui, els diferents protocols de seqüenciació, i posteriorment, els diferents programes d'anàlisi, poden donar resultats molt diferents, Tyler S. Alioto et al. recomanen analitzar les dades amb més d'un programa per tal d'obtenir resultats més acurats i descartar potencials falsos positius (Alioto et al. 2015). Per a l'anàlisi de les insersions i delecions serà important seguir aquesta recomanació com a mesura de filtratge.

4. Reptes de l'ús de les noves tecnologies NGS

Les tecnologies de NGS actuals ens han aportat moltes millores en el diagnòstic i recerca genètiques, tot i que encara existexien problemes en l'anàlisi de seqüències repetides i de baixa complexitat. En són bons exemples les regions repetitives dels telòmers, o els gens com *PMS2* que, com hem vist en l'article 2, té diversos pseudogens altament homòlegs i fa que les lectures puguin mapar erròniament indicant variants falses o perdent-ne de veritables.

S'espera que tecnologies més avançades amb lectures més llargues i menor percentatge d'errors podran solucionar aquestes dificultats. Actualment s'estan desenvolupant tecnologies de seqüenciació de DNA anomenades de tercera generació, que es caracteritzen per la inspecció directa de molècules individuals (Schadt et al. 2010, Zhang et al. 2011). Aquestes tecnologies de seqüenciació de molècules individuals (en anglès, *Single Molecule Sequencing*, SMS) es poden dividir en tres categories: (i) les tecnologies on cada cel·la té una sola molècula de DNA polimerasa que sintetitza una molècula individual de DNA; (ii) tecnologies que utilitzen tècniques de microscòpia avançada per captar directament les imatges de les molècules individuals de DNA de cadena simple; i (iii) les tecnologies de seqüenciació de nanoporus, on molècules individuals de DNA de cadena simple es passen a través de nanoporus i les bases es van identificant a mesura que hi passen. Les tres aproximacions intenten estalviar-se la necessitat de l'amplificació del DNA, que pot introduir artefactes, i generar problemes de desfasament en els clons seqüenciats, que limiten la longitud de lectura. També busquen aconseguir seqüències molt llargues (> 10 kb) per a tractar regions complicades del genoma, i apunten a una seqüenciació molt més ràpida (Figura 12).

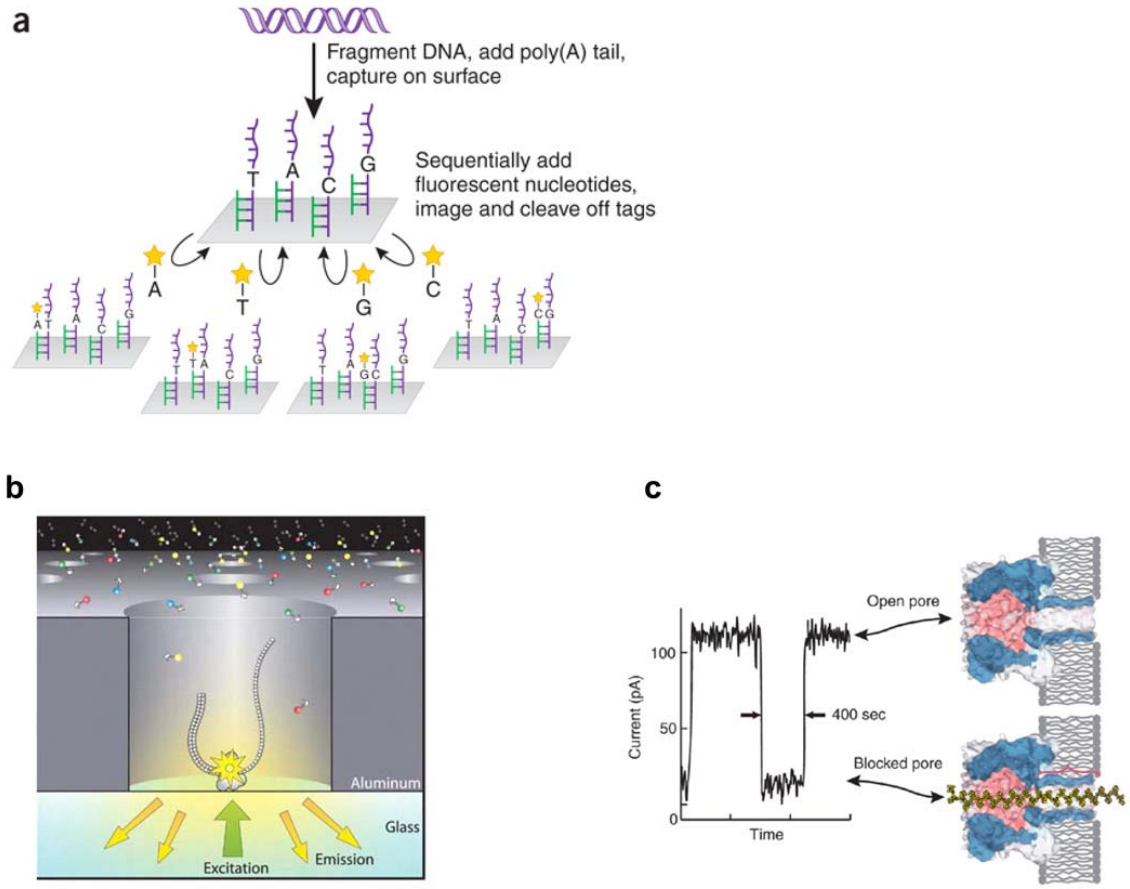


Figura 12. Representació de tres de les principals aproximacions per a la seqüenciació de tercera generació. A) Seqüenciació de molècules individuals amb la tecnologia Helicos SMDS/DRS; B) Sistema de seqüenciació SMRT de Pacific Biosciences; C) Seqüenciació de nanoporus. Extreta de (Ozsolak 2012).

Tot i això, aquestes tecnologies de seqüenciació de molècules aïllades no demostren encara robustesa suficient. Recentment s'ha descrit també la que seria la quarta generació de seqüenciació, basada en mètodes de seqüenciació *in situ,* que exploten la química de la NGS per a llegir la composició dels àcids nucleics directament de cèl·lules i teixits fixats (Ke et al. 2013, Mignardi and Nilsson 2014). Aquest tipus de seqüenciació planteja bones perspectives per a la recerca i el diagnòstic molecular del càncer, però encara està en desenvolupament.

Així com s'estan cercant millores en els instruments de seqüenciació, millores en termes bioinformàtics també seran necessàries per emmagatzemar, analitzar i interpretar les grans quantitats de dades genòmiques que es produiran. Així, per analitzar la gran quantitat de dades que generarà la seqüenciació de tercera o quarta generació s'hauran d'estudiar nous models matemàtics i algoritmes per treure el màxim d'informació de les dades. S'haurà d'implementar una nova generació d'eines i *software* d'anàlisi (Schadt et al. 2010). Un dels aspectes que s'estan estudiant és com adaptar els algoritmes d'alineament a les lectures cada cop més llargues, així com avenços en el *hardware* informàtic que permetin processar les dades més ràpidament (Reinert et al. 2015).

CONCLUSIONS

- S'ha desenvolupat un algoritme complert de generació i anàlisi de seqüencies obtingudes amb la tecnologia NGS 454 que permet detectar substitucions i petites insercions o delecions, amb una sensibilitat i especificitat adequades per al diagnòstic mutacional del gens *BRCA1* i *BRCA2*. Aquesta mateixa aproximació és vàlida pels principals gens responsables del càncer colorectal hereditari (*APC, MUTYH, MLH1, MSH2, MSH6).* El protocol comprèn una anàlisi bioinformàtica acurada de les dades de NGS, una anàlisi específica d'homopolímers i un protocol de complementació i confirmació per seqüenciació Sanger.

- L'anàlisi normalitzada de les dades de la cobertura dels amplicons permet suggerir la presència de variacions en el nombre de còpies, tot i que la precisió en la detecció d'aquest tipus de variants depèn en gran mesura de la homogeneïtat i qualitat de la seqüenciació i encara no és prou acurada per ser utilitzada en un context diagnòstic.

- El *software* lliure permet adaptar les anàlisis bioinformàtiques als interessos de cada experiment i posar a l'abast de tota la comunitat investigadora eines potents per a l'anàlisi de les dades de NGS. El desenvolupament de l'aplicació web "ICO Amplicon NGS Data Analysis" ha mostrat ser útil no solament per la nostra unitat de diagnòstic genètic sinó per altres grups de càncer hereditari.

- Resultats preliminars indiquen que l'anàlisi d'un panell de múltiples gens relacionats amb càncer hereditari és una aproximació de major rendiment diagnòstic que l'anàlisi d'un o pocs gens i que permet una homogenització tècnica i una anàlisi eficient dels resultats de NGS obtinguts.

- L'anàlisi bioinformàtica d'exomes ha permès identificar el gen *AMER1* com a gen important en la carcinogènesi del càncer colorectal, mostrant-se mutat en aproximadament un 10% dels càncers colorectals esporàdics.

BIBLIOGRAFIA

Alioto, T. S., I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge, E. Hovig, L. E. Heisler, T. A. Beck, J. T. Simpson, L. Tonon, A. S. Sertier, A. M. Patch, N. Jager, P. Ginsbach, R. Drews, N. Paramasivam, R. Kabbe, S. Chotewutmontri, N. Diessl, C. Previti, S. Schmidt, B. Brors, L. Feuerbach, M. Heinold, S. Grobner, A. Korshunov, P. S. Tarpey, A. P. Butler, J. Hinton, D. Jones, A. Menzies, K. Raine, R. Shepherd, L. Stebbings, J. W. Teague, P. Ribeca, F. C. Giner, S. Beltran, E. Raineri, M. Dabad, S. C. Heath, M. Gut, R. E. Denroche, N. J. Harding, T. N. Yamaguchi, A. Fujimoto, H. Nakagawa, V. Quesada, R. Valdes-Mas, S. Nakken, D. Vodak, L. Bower, A. G. Lynch, C. L. Anderson, N. Waddell, J. V. Pearson, S. M. Grimmond, M. Peto, P. Spellman, M. He, C. Kandoth, S. Lee, J. Zhang, L. Letourneau, S. Ma, S. Seth, D. Torrents, L. Xi, D. A. Wheeler, C. Lopez-Otin, E. Campo, P. J. Campbell, P. C. Boutros, X. S. Puente, D. S. Gerhard, S. M. Pfister, J. D. McPherson, T. J. Hudson, M. Schlesner, P. Lichter, R. Eils, D. T. Jones and I. G. Gut (2015). "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing." Nat Commun **6**: 10001.

Allard, M. W., Y. Luo, E. Strain, C. Li, C. E. Keys, I. Son, R. Stones, S. M. Musser and E. W. Brown (2012). "High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach." BMC Genomics **13**: 32.

American Society of Clinical Oncology. "ASCO, Making a world of difference in cancer care." 2016, from http://www.asco.org/.

Babraham Bioinformatics. "FastQC." 2016, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Brea-Fernandez, A. J., J. M. Cameselle-Teijeiro, C. Alenda, C. Fernandez-Rozadilla, J. Cubiella, J. Clofent, J. M. Rene, U. Anido, M. Mila, F. Balaguer, A. Castells, S. Castellvi-Bel, R. Jover, A. Carracedo and C. Ruiz-Ponte (2014). "High incidence of large deletions in the PMS2 gene in Spanish Lynch syndrome families." Clin Genet **85**(6): 583-588.

Broad Institute. "Picard tools." 2016, from http://broadinstitute.github.io/picard/.

Burrows-Wheeler Aligner. "BWA." 2016, from http://bio-bwa.sourceforge.net/.

Choi, M., U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane and R. P. Lifton (2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." Proc Natl Acad Sci U S A **106**(45): 19096-19101.

CLCbio QIAGEN. "End-to-end NGS data analysis solution." 2016, from https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/.

Cloonan, N., A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan and S. M. Grimmond (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing." Nat Methods **5**(7): 613-619.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. Genomes Project Analysis (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

Danzer, M., N. Niklas, S. Stabentheiner, K. Hofer, J. Proll, C. Stuckler, E. Raml, H. Polin and C. Gabriel (2013). "Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics." BMC Genomics **14**: 221.

De Leeneer, K., J. De Schrijver, L. Clement, M. Baetens, S. Lefever, S. De Keulenaer, W. Van Criekinge, D. Deforce, F. Van Nieuwerburgh, S. Bekaert, F. Pattyn, B. De Wilde, P. Coucke, J. Vandesompele, K. Claes and J. Hellemans (2011). "Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics." PLoS One **6**(9): e25531.

De Leeneer, K., J. Hellemans, J. De Schrijver, M. Baetens, B. Poppe, W. Van Criekinge, A. De Paepe, P. Coucke and K. Claes (2011). "Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations." Hum Mutat **32**(3): 335-344.

De Schrijver, J. M., K. De Leeneer, S. Lefever, N. Sabbe, F. Pattyn, F. Van Nieuwerburgh, P. Coucke, D. Deforce, J. Vandesompele, S. Bekaert, J. Hellemans and W. Van Criekinge (2010). "Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline." BMC Bioinformatics **11**: 269.

Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002). "Capturing chromosome conformation." Science **295**(5558): 1306-1311.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler and M. J. Daly (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nat Genet **43**(5): 491-498.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." Nucleic Acids Res **36**(16): e105.

Easton, D. F., P. D. Pharoah, A. C. Antoniou, M. Tischkowitz, S. V. Tavtigian, K. L. Nathanson, P. Devilee, A. Meindl, F. J. Couch, M. Southey, D. E. Goldgar, D. G. Evans, G. Chenevix-Trench, N. Rahman, M. Robson, S. M. Domchek and W. D. Foulkes (2015). "Gene-panel sequencing and the prediction of breast-cancer risk." N Engl J Med **372**(23): 2243-2257.

EMBL-EBI. "IGSR: The International Genome Sample Resource. 2016" from http://www.1000genomes.org/.

Frazer, K. A., S. S. Murray, N. J. Schork and E. J. Topol (2009). "Human genetic variation and its contribution to complex traits." Nat Rev Genet **10**(4): 241-251.

Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman and M. R. Stratton (2004). "A census of human cancer genes." Nat Rev Cancer **4**(3): 177-183.

Hall, M. J., A. D. Forman, R. Pilarski, G. Wiesner and V. N. Giri (2014). "Gene panel testing for inherited cancer risk." J Natl Compr Canc Netw **12**(9): 1339-1346.

Hansen, M. F., U. Neckmann, L. A. Lavik, T. Vold, B. Gilde, R. K. Toft and W. Sjursen (2014). "A massive parallel sequencing workflow for diagnostic genetic testing of mismatch repair genes." Mol Genet Genomic Med **2**(2): 186-200.

Harris, T. D., P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss and Z. Xie (2008). "Single-molecule DNA sequencing of a viral genome." Science **320**(5872): 106-109.

Henson, J., G. Tischler and Z. Ning (2012). "Next-generation sequencing and large genome assemblies." Pharmacogenomics **13**(8): 901-915.

HGVS. "Human Genome Variation Society." 2016, from http://www.hgvs.org/.

Homer, N., B. Merriman and S. F. Nelson (2009). "BFAST: an alignment tool for large scale genome resequencing." PLoS One **4**(11): e7767.

ICGC. "International Cancer Genome Consortium." 2016, from https://icgc.org/.

Illumina. "For all you seek." 2016, from http://www.illumina.com/.

Illumina BaseSpace. "BaseSpace Sequence Hub." 2016, from http://support.illumina.com/sequencing/sequencing_software/basespace.html.

International Cancer Genome, C., T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolas, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P. A. Futreal,

H. Aburatani, M. Bayes, D. D. Botwell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. J. Campbell, P. Flicek, G. Getz, R. Guigo, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, S. M. Jones, Q. Li, N. Lopez-Bigas, R. Luo, L. Muthuswamy, B. F. Ouellette, J. V. Pearson, X. S. Puente, V. Quesada, B. J. Raphael, C. Sander, T. Shibata, T. P. Speed, L. D. Stein, J. M. Stuart, J. W. Teague, Y. Totoki, T. Tsunoda, A. Valencia, D. A. Wheeler, H. Wu, S. Zhao, G. Zhou, L. D. Stein, R. Guigo, T. J. Hubbard, Y. Joly, S. M. Jones, A. Kasprzyk, M. Lathrop, N. Lopez-Bigas, B. F. Ouellette, P. T. Spellman, J. W. Teague, G. Thomas, A. Valencia, T. Yoshida, K. L. Kennedy, M. Axton, S. O. Dyke, P. A. Futreal, D. S. Gerhard, C. Gunter, M. Guyer, T. J. Hudson, J. D. McPherson, L. J. Miller, B. Ozenberger, K. M. Shaw, A. Kasprzyk, L. D. Stein, J. Zhang, S. A. Haider, J. Wang, C. K. Yung, A. Cros, Y. Liang, S. Gnaneshan, J. Guberman, J. Hsu, M. Bobrow, D. R. Chalmers, K. W. Hasel, Y. Joly, T. S. Kaan, K. L. Kennedy, B. M. Knoppers, W. W. Lowrance, T. Masui, P. Nicolas, E. Rial-Sebbag, L. L. Rodriguez, C. Vergely, T. Yoshida, S. M. Grimmond, A. V. Biankin, D. D. Bowtell, N. Cloonan, A. deFazio, J. R. Eshleman, D. Etemadmoghadam, B. B. Gardiner, J. G. Kench, A. Scarpa, R. L. Sutherland, M. A. Tempero, N. J. Waddell, P. J. Wilson, J. D. McPherson, S. Gallinger, M. S. Tsao, P. A. Shaw, G. M. Petersen, D. Mukhopadhyay, L. Chin, R. A. DePinho, S. Thayer, L. Muthuswamy, K. Shazand, T. Beck, M. Sam, L. Timms, V. Ballin, Y. Lu, J. Ji, X. Zhang, F. Chen, X. Hu, G. Zhou, Q. Yang, G. Tian, L. Zhang, X. Xing, X. Li, Z. Zhu, Y. Yu, J. Yu, H. Yang, M. Lathrop, J. Tost, P. Brennan, I. Holcatova, D. Zaridze, A. Brazma, L. Egevard, E. Prokhortchouk, R. E. Banks, M. Uhlen, A. Cambon-Thomsen, J. Viksna, F. Ponten, K. Skryabin, M. R. Stratton, P. A. Futreal, E. Birney, A. Borg, A. L. Borresen-Dale, C. Caldas, J. A. Foekens, S. Martin, J. S. Reis-Filho, A. L. Richardson, C. Sotiriou, H. G. Stunnenberg, G. Thoms, M. van de Vijver, L. van't Veer, F. Calvo, D. Birnbaum, H. Blanche, P. Boucher, S. Boyault, C. Chabannon, I. Gut, J. D. Masson-Jacquemier, M. Lathrop, I. Pauporte, X. Pivot, A. Vincent-Salomon, E. Tabone, C. Theillet, G. Thomas, J. Tost, I. Treilleux, F. Calvo, P. Bioulac-Sage, B. Clement, T. Decaens, F. Degos, D. Franco, I. Gut, M. Gut, S. Heath, M. Lathrop, D. Samuel, G. Thomas, J. Zucman-Rossi, P. Lichter, R. Eils, B. Brors, J. O. Korbel, A. Korshunov, P. Landgraf, H. Lehrach, S. Pfister, B. Radlwimmer, G. Reifenberger, M. D. Taylor, C. von Kalle, P. P. Majumder, R. Sarin, T. S. Rao, M. K. Bhan, A. Scarpa, P. Pederzoli, R. A. Lawlor, M. Delledonne, A. Bardelli, A. V. Biankin, S. M. Grimmond, T. Gress, D. Klimstra, G. Zamboni, T. Shibata, Y. Nakamura, H. Nakagawa, J. Kusada, T. Tsunoda, S. Miyano, H. Aburatani, K. Kato, A. Fujimoto, T. Yoshida, E. Campo, C. Lopez-Otin, X. Estivill, R. Guigo, S. de Sanjose, M. A. Piris, E. Montserrat, M. Gonzalez-Diaz, X. S. Puente, P. Jares, A. Valencia, H. Himmelbauer, V. Quesada, S. Bea, M. R. Stratton, P. A. Futreal, P. J. Campbell, A. Vincent-Salomon, A. L. Richardson, J. S. Reis-Filho, M. van de Vijver, G. Thomas, J. D. Masson-Jacquemier, S. Aparicio, A. Borg, A. L. Borresen-Dale, C. Caldas, J. A. Foekens, H. G. Stunnenberg, L. van't Veer, D. F. Easton, P. T. Spellman, S. Martin, A. D. Barker, L. Chin, F. S. Collins, C. C. Compton, M. L. Ferguson, D. S. Gerhard, G. Getz, C. Gunter, A. Guttmacher, M. Guyer, D. N. Hayes, E. S. Lander, B. Ozenberger, R. Penny, J. Peterson, C. Sander, K. M. Shaw, T. P. Speed, P. T. Spellman, J. G. Vockley, D. A. Wheeler, R. K. Wilson, T. J. Hudson, L. Chin, B. M. Knoppers, E. S. Lander, P. Lichter, L. D. Stein, M. R. Stratton, W. Anderson, A. D. Barker, C. Bell, M. Bobrow, W. Burke, F. S. Collins, C. C. Compton, R. A. DePinho, D. F. Easton, P. A. Futreal, D. S. Gerhard, A. R. Green, M. Guyer, S. R. Hamilton, T. J. Hubbard, O. P. Kallioniemi, K. L. Kennedy, T. J. Ley, E. T. Liu, Y. Lu, P. Majumder, M. Marra, B. Ozenberger, J. Peterson, A. J. Schafer, P. T. Spellman, H. G. Stunnenberg, B. J. Wainwright, R. K. Wilson and H. Yang (2010). "International network of cancer genome projects." Nature **464**(7291): 993-998.

Jiang, Q., T. Turner, M. X. Sosa, A. Rakha, S. Arnold and A. Chakravarti (2012). "Rapid and efficient human mutation detection using a bench-top next-generation DNA sequencer." Hum Mutat **33**(1): 281-289.

John Hopkings University. "Bowtie2, Fast and sensitive read alignment." 2016, from http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.

JSI Medical Systems. "SeqNext, Sequence Pilot software for genetic analysis." 2016, from http://www.jsi-medisys.de/products/seqnext.

Ke, R., M. Mignardi, A. Pacureanu, J. Svedlund, J. Botling, C. Wahlby and M. Nilsson (2013). "In situ sequencing for RNA analysis in preserved tissue and cells." Nat Methods **10**(9): 857-860.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Kharchenko, P. V., M. Y. Tolstorukov and P. J. Park (2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nat Biotechnol **26**(12): 1351-1359.

Koboldt, D. C., L. Ding, E. R. Mardis and R. K. Wilson (2010). "Challenges of sequencing human genomes." Brief Bioinform **11**(5): 484-498.

Koboldt, D. C., D. E. Larson, K. Chen, L. Ding and R. K. Wilson (2012). "Massively parallel sequencing approaches for characterization of structural variation." Methods Mol Biol **838**: 369-384.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding and R. K. Wilson (2012). "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing." Genome Res **22**(3): 568-576.

Kruglyak, L. and D. A. Nickerson (2001). "Variation is the spice of life." Nat Genet **27**(3): 234-236.
LaDuca, H., A. J. Stuenkel, J. S. Dolinsky, S. Keiles, S. Tandy, T. Pesaran, E. Chen, C. L. Gau, E. Palmaer, K. Shoaepour, D. Shah, V. Speare, S. Gandomi and E. Chao (2014). "Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients." Genet Med **16**(11): 830-837.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.
Lapunzina, P., R. O. Lopez, L. Rodriguez-Laguna, P. Garcia-Miguel, A. R. Martinez and V. Martinez-Glez (2014). "Impact of NGS in the medical sciences: Genetic syndromes with an increased risk of developing cancer as an example of the use of new technologies." Genet Mol Biol **37**(1 Suppl): 241-249.

Lawrence, M. S., P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander and G. Getz (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes." Nature **499**(7457): 214-218.

Lee, H., E. Popodi, P. L. Foster and H. Tang (2014). "Detection of structural variants involving repetitive regions in the reference genome." J Comput Biol **21**(3): 219-233.

Lemke, J. R., E. Riesch, T. Scheurenbrand, M. Schubach, C. Wilhelm, I. Steiner, J. Hansen, C. Courage, S. Gallati, S. Burki, S. Strozzi, B. G. Simonetti, S. Grunt, M. Steinlin, M. Alber, M. Wolff, T. Klopstock, E. C. Prott, R. Lorenz, C. Spaich, S. Rona, M. Lakshminarasimhan, J. Kroll, T. Dorn, G. Kramer, M. Synofzik, F. Becker, Y. G. Weber, H. Lerche, D. Bohm and S. Biskup (2012). "Targeted next generation sequencing as a diagnostic tool in epileptic disorders." Epilepsia **53**(8): 1387-1398.

Li, H. (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." Bioinformatics **27**(21): 2987-2993.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics **26**(5): 589-595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, R., Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen and J. Wang (2009). "SNP detection for massively parallel whole-genome resequencing." Genome Res **19**(6): 1124-1132.

Liu, B., J. M. Conroy, C. D. Morrison, A. O. Odunsi, M. Qin, L. Wei, D. L. Trump, C. S. Johnson, S. Liu and J. Wang (2015). "Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives." Oncotarget **6**(8): 5477-5489.

Lo, C. C. and P. S. Chain (2014). "Rapid evaluation and quality control of next generation sequencing data with FaQCs." BMC Bioinformatics **15**: 366.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain and M. J. Pallen (2012). "Performance comparison of benchtop high-throughput sequencing platforms." Nat Biotechnol **30**(5): 434-439.
Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure and D. J. Turner (2010). "Target-enrichment strategies for next-generation sequencing." Nat Methods **7**(2): 111-118.

Manske, H. M. and D. P. Kwiatkowski (2009). "LookSeq: a browser-based viewer for deep sequencing data." Genome Res **19**(11): 2125-2132.

Marco-Sola, S., M. Sammeth, R. Guigo and P. Ribeca (2012). "The GEM mapper: fast, accurate and versatile alignment by filtration." Nat Methods **9**(12): 1185-1188.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-1303.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Mignardi, M. and M. Nilsson (2014). "Fourth-generation sequencing in the cell and the clinic." Genome Med **6**(4): 31.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson and J. Shendure (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

NIH. "The Cancer Genome Atlas, TCGA." 2016, from http://cancergenome.nih.gov/.

Okonechnikov, K., O. Golosova, M. Fursov and U. team (2012). "Unipro UGENE: a unified bioinformatics toolkit." Bioinformatics **28**(8): 1166-1167.

Ozsolak, F. (2012). "Third-generation sequencing techniques and applications to drug discovery." Expert Opin Drug Discov **7**(3): 231-243.

Patel, R. K. and M. Jain (2012). "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data." PLoS One **7**(2): e30619.

Pennisi, E. (2014). "Genomics. DNA sequencers still waiting for the nanopore revolution." Science **343**(6173): 829-830.

Petric, R. C., L. A. Pop, A. Jurj, L. Raduly, D. Dumitrascu, N. Dragos and I. B. Neagoe (2015). "Next generation sequencing applications for breast cancer research." Clujul Med **88**(3): 278-287.
Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." Trends Genet **24**(3): 142-149.

Puente, X. S., M. Pinyol, V. Quesada, L. Conde, G. R. Ordonez, N. Villamor, G. Escaramis, P. Jares, S. Bea, M. Gonzalez-Diaz, L. Bassaganyas, T. Baumann, M. Juan, M. Lopez-Guerra, D. Colomer, J. M. Tubio, C. Lopez, A. Navarro, C. Tornador, M. Aymerich, M. Rozman, J. M. Hernandez, D. A. Puente, J. M. Freije, G. Velasco, A. Gutierrez-Fernandez, D. Costa, A. Carrio, S. Guijarro, A. Enjuanes, L. Hernandez, J. Yague, P. Nicolas, C. M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J. C. Dohm, S. de Sanjose, M. A. Piris, E. de Alava, J. San Miguel, R. Royo, J. L. Gelpi, D. Torrents, M. Orozco, D. G. Pisano, A. Valencia, R. Guigo, M. Bayes, S. Heath, M. Gut, P. Klatt, J. Marshall, K. Raine, L. A. Stebbings, P. A. Futreal, M. R. Stratton, P. J. Campbell, I. Gut, A. Lopez-Guillermo, X. Estivill, E. Montserrat, C. Lopez-Otin and E. Campo (2011). "Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia." Nature **475**(7354): 101-105.

Quer, J., J. Gregori, F. Rodriguez-Frias, M. Buti, A. Madejon, S. Perez-del-Pulgar, D. Garcia-Cehic, R. Casillas, M. Blasi, M. Homs, D. Tabernero, M. Alvarez-Tejado, J. M. Munoz, M. Cubero, A. Caballero, J. A. del Campo, E. Domingo, I. Belmonte, L. Nieto, S. Lens, P. Munoz-de-Rueda, P. Sanz-Cameno, S. Sauleda, M. Bes, J. Gomez, C. Briones, C. Perales, J. Sheldon, L. Castells, L. Viladomiu, J. Salmeron, A. Ruiz-Extremera, R. Quiles-Perez, R. Moreno-Otero, R. Lopez-Rodriguez, H. Allende, M. Romero-Gomez, J. Guardia, R. Esteban, J. Garcia-Samaniego, X. Forns and J. I. Esteban (2015). "High-resolution hepatitis C virus subtyping using NS5B deep sequencing and phylogeny, an alternative to current methods." J Clin Microbiol **53**(1): 219-226.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-454.

Reinert, K., B. Langmead, D. Weese and D. J. Evers (2015). "Alignment of Next-Generation Sequencing Reads." Annu Rev Genomics Hum Genet **16**: 133-151.

Richter, B. G. and D. P. Sexton (2009). "Managing and analyzing next-generation sequence data." PLoS Comput Biol **5**(6): e1000369.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). "Integrative genomics viewer." Nat Biotechnol **29**(1): 24-26.

Roche 454 sequencing. "Analysis software." 2016, from http://www.454.com/products/analysis-software/.

Roche. "454 Sequencing." 2016, from http://www.454.com/.

Rumble, S. M., P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow and M. Brudno (2009). "SHRiMP: accurate mapping of short color-space reads." PLoS Comput Biol **5**(5): e1000386.

Schadt, E. E., S. Turner and A. Kasarskis (2010). "A window into third-generation sequencing." Hum Mol Genet **19**(R2): R227-240.

SeattleSeq Annotation 137. "SeattleSeq Annotation 137" from http://snp.gs.washington.edu/SeattleSeqAnnotation137/HelpAbout.jsp.

Seshagiri, S. (2013). "The burden of faulty proofreading in colon cancer." Nat Genet **45**(2): 121-122.

Shankar, R. (2011). "The bioinformatics of next generation sequencing: a meeting report." J Mol Cell Biol **3**(3): 147-150.

Shearer, A. E., A. P. DeLuca, M. S. Hildebrand, K. R. Taylor, J. Gurrola, 2nd, S. Scherer, T. E. Scheetz and R. J. Smith (2010). "Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing." Proc Natl Acad Sci U S A **107**(49): 21104-21109.

Shen, Y., Z. Wan, C. Coarfa, R. Drabek, L. Chen, E. A. Ostrowski, Y. Liu, G. M. Weinstock, D. A. Wheeler, R. A. Gibbs and F. Yu (2010). "A SNP discovery method to assess variant allele probability from next-generation resequencing data." Genome Res **20**(2): 273-280.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." Nat Genet **38**(11): 1348-1354.

Swanson, A., E. Ramos and H. Snyder (2014). "Next Generation sequencing is the impetus for the next generation of laboratory-based genetic counselors." J Genet Couns **23**(4): 647-654.

Tarpey, P. S., R. Smith, E. Pleasance, A. Whibley, S. Edkins, C. Hardy, S. O'Meara, C. Latimer, E. Dicks, A. Menzies, P. Stephens, M. Blow, C. Greenman, Y. Xue, C. Tyler-Smith, D. Thompson, K. Gray, J. Andrews, S. Barthorpe, G. Buck, J. Cole, R. Dunmore, D. Jones, M. Maddison, T. Mironenko, R. Turner, K. Turrell, J. Varian, S. West, S. Widaa, P. Wray, J. Teague, A. Butler, A. Jenkinson, M. Jia, D. Richardson, R. Shepherd, R. Wooster, M. I. Tejada, F. Martinez, G. Carvill, R. Goliath, A. P. de Brouwer, H. van Bokhoven, H. Van Esch, J. Chelly, M. Raynaud, H. H. Ropers, F. E. Abidi, A. K. Srivastava, J. Cox, Y. Luo, U. Mallya, J. Moon, J. Parnau, S. Mohammed, J. L. Tolmie, C. Shoubridge, M. Corbett, A. Gardner, E. Haan, S. Rujirabanjerd, M. Shaw, L. Vandeleur, T. Fullston, D. F. Easton, J. Boyle, M. Partington, A. Hackett, M. Field, C. Skinner, R. E. Stevenson, M. Bobrow, G. Turner, C. E. Schwartz, J. Gecz, F. L. Raymond, P. A. Futreal and M. R. Stratton (2009). "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation." Nat Genet **41**(5): 535-543.

Thermo Fisher Scientific. "Applied Biosystems." 2016, from http://www.thermofisher.com/es/es/home/brands/applied-biosystems.html.

Thermo Fisher Scientific. "Ion Torrent." 2016, from https://www.thermofisher.com/es/es/home/brands/ion-torrent.html.

Thompson, E. R., M. A. Doyle, G. L. Ryland, S. M. Rowley, D. Y. Choong, R. W. Tothill, H. Thorne, kConFab, D. R. Barnes, J. Li, J. Ellul, G. K. Philip, Y. C. Antill, P. A. James, A. H. Trainer, G. Mitchell and I. G. Campbell (2012). "Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles." PLoS Genet **8**(9): e1002894.

Thorvaldsdottir, H., J. T. Robinson and J. P. Mesirov (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Brief Bioinform **14**(2): 178-192.

Turner, E. H., S. B. Ng, D. A. Nickerson and J. Shendure (2009). "Methods for genomic partitioning." Annu Rev Genomics Hum Genet **10**: 263-284.

Voelkerding, K. V., S. Dames and J. D. Durtschi (2010). "Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology." J Mol Diagn **12**(5): 539-551.

Wang, K., M. Li and H. Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Weber, M., J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam and D. Schubeler (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." Nat Genet **37**(8): 853-862.

Wimmer, K. and A. Wernstedt (2014). "PMS2 gene mutational analysis: direct cDNA sequencing to circumvent pseudogene interference." Methods Mol Biol **1167**: 289-302.

Xu, H., J. DiCarlo, R. V. Satya, Q. Peng and Y. Wang (2014). "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data." BMC Genomics **15**: 244.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler and Z. Ning (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." Bioinformatics **25**(21): 2865-2871.

Zhang, J., R. Chiodini, A. Badr and G. Zhang (2011). "The impact of next-generation sequencing on genomics." J Genet Genomics **38**(3): 95-109.

ANNEX

A continuació es presenten altres publicacions on la doctoranda ha contribuït com a coautora. En totes elles la doctoranda ha participat en l'anàlisi estadística i/o bioinformàtica de les dades. Totes les publicacions s'emmarquen dins de les línies de recerca del càncer. Es presenta la llista en ordre cronològic invers, i posteriorment la primera pàgina de cadascun d'ells en el format de la revista on ha estat publicat.

Llista de publicacions:

1. Mutanome and expression of immune response genes in microsatelite stable colon cancer.
Sanz-Pamplona R, Gil-Hoyos R, **López-Doriga A**, Alonso MH, Aussó S, Molleví DG, Santos C, Sanjuán X, Salazar R, Alemany R, Moreno V.
Oncotarget. 2016 Feb 9.

2. Mutations in JMJD1C are involved in Rett syndrome and intellectual disability.
Sáez MA, Fernández-Rodríguez J, Moutinho C, Sanchez-Mut JV, Gomez A, Vidal E, Petazzi P, Szczesna K, Lopez-Serra P, Lucariello M, Lorden P, Delgado-Morales R, de la Caridad OJ, Huertas D, Gelpí JL, Orozco M, **López-Doriga A**, Milà M, Perez-Jurado LA, Pineda M, Armstrong J, Lázaro C, Esteller M.
Genet Med. 2015 Jul 16.

3. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair.
Seguí N, Mina LB, Lázaro C, Sanz-Pamplona R, Pons T, Navarro M, Bellido F, **López-Doriga A**, Valdés-Mas R, Pineda M, Guinó E, Vidal A, Soto JL, Caldés T, Durán M, Urioste M, Rueda D, Brunet J, Balbín M, Blay P, Iglesias S, Garré P, Lastra E, Sánchez-Heras AB, Valencia A, Moreno V, Pujana MÁ, Villanueva A, Blanco I, Capellá G, Surrallés J, Puente XS, Valle L.
Gastroenterology. 2015 Sep;149(3):563-6.

4. Comprehensive establishment and characterization of orthoxenograft mouse models of malignant peripheral nerve sheath tumors for personalized medicine.
Castellsagué J, Gel B, Fernández-Rodríguez J, Llatjós R, Blanco I, Benavente Y, Pérez-Sidelnikova D, García-Del Muro J, Viñals JM, Vidal A, Valdés-Mas R, Terribas E, **López-Doriga A**, Pujana MA, Capellá G, Puente XS, Serra E, Villanueva A, Lázaro C.
EMBO Mol Med. 2015 Mar 25;7(5):608-27.

5. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis.
Closa A, Cordero D, Sanz-Pamplona R, Solé X, Crous-Bou M, Paré-Brunet L, Berenguer A, Guino E, **Lopez-Doriga A**, Guardiola J, Biondo S, Salazar R, Moreno V.
Carcinogenesis. 2014 Sep;35(9):2039-46.

6. Exome sequencing identifies MUTYH mutations in a family with colorectal cancer and an atypical phenotype.
Seguí N, Navarro M, Pineda M, Köger N, Bellido F, González S, Campos O, Iglesias S, Valdés-Mas R, **López-Doriga A**, Gut M, Blanco I, Lázaro C, Capellá G, Puente XS, Plotz G, Valle L.
Gut. 2015 Feb;64(2):355-6.

7. Clinicopathological risk factors of Stage II colon cancer: results of a prospective study.
Santos C, **López-Doriga A**, Navarro M, Mateo J, Biondo S, Martínez Villacampa M, Soler G, Sanjuan X, Paules MJ, Laquente B, Guinó E, Kreisler E, Frago R, Germà JR, Moreno V, Salazar R.
Colorectal Dis. 2013 Apr;15(4):414-22.

8. Susceptibility genetic variants associated with early-onset colorectal cancer.
Giráldez MD, **López-Dóriga A**, Bujanda L, Abulí A, Bessa X, Fernández-Rozadilla C, Muñoz J, Cuatrecasas M, Jover R, Xicola RM, Llor X, Piqué JM, Carracedo A, Ruiz-Ponte C, Cosme A, Enríquez-Navascués JM, Moreno V, Andreu M, Castells A, Balaguer F, Castellví-Bel S; Gastrointestinal Oncology Group of the Spanish Gastroenterological Association.
Carcinogenesis. 2012 Mar;33(3):613-9.

9. Hepatic carcinoma-associated fibroblasts promote an adaptative response in colorectal cancer cells that inhibit proliferation and apoptosis: nonresistant cells die by nonapoptotic cell death.
Berdiel-Acer M, Bohem ME, **López-Doriga A**, Vidal A, Salazar R, Martínez-Iniesta M, Santos C, Sanjuan X, Villanueva A, Molleví DG.
Neoplasia. 2011 Oct;13(10):931-46.

10. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer.
Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, **Lopez-Doriga A**, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ, Tollenaar R.
J Clin Oncol. 2011 Jan 1;29(1):17-24.

Online ISSN: 1949-2553

Oncotarget

Research Papers:

Mutanome and expression of immune response genes in microsatellite stable colon cancer

PDF | HTML | Supplementary Files

DOI: 10.18632/oncotarget.7293

Metrics: HTML *5 views*

Rebeca Sanz-Pamplona[1], Raúl Gil-Hoyos[2], Adriana López-Doriga[1], M. Henar Alonso[1], Susanna Aussó[1], David G. Molleví[2], Cristina Santos[2,3], Xavier Sanjuán[4], Ramón Salazar[2,3], Ramón Alemany[2], Víctor Moreno[1,5]

[1]Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL) and CIBERESP, L'Hospitalet de Llobregat, Barcelona, Spain

[2]Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

[3]Department of Medical Oncology, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

[4]Pathology Service, University Hospital Bellvitge (HUB – IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain

[5]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona (UB), Barcelona, Spain

Correspondence to: Victor Moreno, e-mail: v.moreno@iconcologia.net

Keywords: *colorectal cancer, neoantigens, prognosis, antigen presentation, immune response*

Received: *December 18, 2015* Accepted: *January 26, 2016* Published: *February 9, 2016*

ABSTRACT

The aim of this study was to analyze the impact of the mutanome in the prognosis of microsatellite stable stage II CRC tumors. The exome of 42 stage II, microsatellite stable, colon tumors (21 of them relapse) and their paired mucosa were sequenced and analyzed. Although some pathways accumulated more mutations in patients exhibiting good or poor prognosis, no single somatic mutation was associated with prognosis. Exome sequencing data is also valuable to infer tumor neoantigens able to elicit a host immune response. Hence, putative neoantigens were identified by combining information about missense mutations in each tumor and HLAs genotypes of the patients. Under the hypothesis that neoantigens should be correctly presented in order to activate the immune response, expression levels of genes involved in the antigen presentation machinery were also assessed. In addition, CD8A level (as a marker of T-cell infiltration) was measured. We found that tumors with better prognosis showed a tendency to generate a higher number of immunogenic epitopes, and up-regulated genes involved in the antigen processing machinery. Moreover, tumors with higher T-cell infiltration also showed better prognosis. Stratifying by consensus molecular subtype, CMS4 tumors showed the highest association of expression levels of genes involved in the antigen presentation machinery with prognosis. Thus, we hypothesize that a subset of stage II microsatellite stable CRC tumors are able to generate an immune response in the host via MHC class I antigen presentation, directly related with a better prognosis.

PII: *7293*

ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Open

Mutations in JMJD1C are involved in Rett syndrome and intellectual disability

Mauricio A. Sáez, PhD[1], Juana Fernández-Rodríguez, PhD[2], Catia Moutinho, PhD[1], Jose V. Sanchez-Mut, PhD[1], Antonio Gomez, PhD[1], Enrique Vidal, PhD[1], Paolo Petazzi, PhD[1], Karolina Szczesna, PhD[1], Paula Lopez-Serra, PhD[1], Mario Lucariello, PhD[1], Patricia Lorden, PhD[1], Raul Delgado-Morales, PhD[1], Olga J. de la Caridad, PhD[1], Dori Huertas, PhD[1], Josep L. Gelpí, PhD[2,3], Modesto Orozco, PhD[2–4], Adriana López-Doriga, PhD[5], Montserrat Milà, PhD[6,7], Luís A. Perez-Jurado, MD, PhD[7,8], Mercedes Pineda, MD, PhD[7,9], Judith Armstrong, PhD[7,9], Conxi Lázaro, PhD[5], and Manel Esteller, MD, PhD[1,4,10]

Purpose: Autism spectrum disorders are associated with defects in social response and communication that often occur in the context of intellectual disability. Rett syndrome is one example in which epilepsy, motor impairment, and motor disturbance may co-occur. Mutations in histone demethylases are known to occur in several of these syndromes. Herein, we aimed to identify whether mutations in the candidate histone demethylase JMJD1C (jumonji domain containing 1C) are implicated in these disorders.

Methods: We performed the mutational and functional analysis of JMJD1C in 215 cases of autism spectrum disorders, intellectual disability, and Rett syndrome without a known genetic defect.

Results: We found seven JMJD1C variants that were not present in any control sample (~ 6,000) and caused an amino acid change involving a different functional group. From these, two *de novo* JMJD1C germline mutations were identified in a case of Rett syndrome and in a patient with intellectual disability. The functional study of the JMJD1C mutant Rett syndrome patient demonstrated that the altered protein had abnormal subcellular localization, diminished activity to demethylate the DNA damage-response protein MDC1, and reduced binding to MECP2. We confirmed that JMJD1C protein is widely expressed in brain regions and that its depletion compromises dendritic activity.

Conclusions: Our findings indicate that mutations in JMJD1C contribute to the development of Rett syndrome and intellectual disability.

Genet Med advance online publication 16 July 2015

Key Words: autism; intellectual disability; mutational screening; Rett syndrome

INTRODUCTION

Autism spectrum disorders are a heterogeneous clinical and genetic group of neurodevelopmental defects that are characterized by impaired social communication functions and inappropriate repetitive behavior.[1] This family of disorders is characterized by enormous phenotypic variability, from mild primary deficits in language pragmatics[2] to major neurological phenotypes, such as that of Rett syndrome (OMIM 312750), where it co-occurs with epilepsy, motor impairment, and sleep disturbance.[3] The disabilities associated with autism spectrum disorders are often so severe that affected individuals do not generally reach parenthood, thereby preventing comprehensive familial genetic studies from being undertaken. However, genetic alterations are already recognized as major etiological factors. In this regard, concordance with autism spectrum disorders is higher than with any other cognitive or behavioral disorder.[4,5] In addition to the contribution of polymorphic variants that confer low or moderate risk of the appearance of these neurodevelopmental defects, a cause of autism spectrum disorders can be the occurrence of de novo mutations affecting genes in a number of cellular pathways.[6–8] A similar scenario can be proposed for the genetic contribution to the even more heterogeneous group of disorders classified as intellectual disabilities.[9]

Among the described genetic defects associated with intellectual disabilities, our attention was caught by a single case report of an autistic patient carrying a de novo balanced paracentric inversion 46, XY in (10)(q11.1;q21.3) in which the distal breakpoint disrupted what was at that time known as the TRIP8 gene,[10] which has been characterized as a member of the jmJC domain-containing protein family involved in the methyl

[1]Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Catalonia, Spain; [2]Joint Biomedical Research Institute-Barcelona Supercomputing Center (IRB-BSC) Computational Biology Program, Barcelona, Catalonia, Spain; [3]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona, Catalonia, Spain; [4]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain; [5]Hereditary Cancer Program, Catalan Institute of Oncology-Bellvitge Institute for Biomedical Research (ICO-IDIBELL), Barcelona, Catalonia, Spain; [6]Biochemistry and Molecular Genetics Department, Hospital Clínic, Barcelona, Catalonia, Spain; [7]CIBERER (Biomedical Network Research Centre on Rare Diseases, Instituto de Salud Carlos III), Barcelona, Spain; [8]Genetics Unit, University Pompeu Fabra, Barcelona, Catalonia, Spain; [9]Department of Neurology, Hospital Sant Joan de Déu (HSJD), Barcelona, Catalonia, Spain; [10]Department of Physiological Sciences II, School of Medicine, University of Barcelona, Barcelona, Catalonia, Spain. Correspondence: Manel Esteller (mesteller@idibell.cat)

Submitted 5 November 2014; accepted 9 June 2015; advance online publication 16 July 2015. doi:10.1038/gim.2015.100

Gastroenterology 2015;149:563–566

BRIEF REPORT

Germline Mutations in *FAN1* Cause Hereditary Colorectal Cancer by Impairing DNA Repair

Nuria Seguí,[1,*] **Leonardo B. Mina**,[2,*] Conxi Lázaro,[1] Rebeca Sanz-Pamplona,[3] Tirso Pons,[4] Matilde Navarro,[1] Fernando Bellido,[1] Adriana López-Doriga,[3] Rafael Valdés-Mas,[5] Marta Pineda,[1] Elisabet Guinó,[3] August Vidal,[6] José Luís Soto,[7] Trinidad Caldés,[8] Mercedes Durán,[9] Miguel Urioste,[10] Daniel Rueda,[11] Joan Brunet,[12] Milagros Balbín,[13] Pilar Blay,[14] Silvia Iglesias,[1] Pilar Garré,[8] Enrique Lastra,[15] Ana Beatriz Sánchez-Heras,[16] Alfonso Valencia,[4] Victor Moreno,[3,17] Miguel Ángel Pujana,[18] Alberto Villanueva,[18] Ignacio Blanco,[1] Gabriel Capellá,[1] Jordi Surrallés,[2] Xose S. Puente,[5] and Laura Valle[1]

[1]Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL, Hospitalet de Llobregat; [2]Genome Instability and DNA Repair Group, Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, and Center for Biomedical Network Research on Rare Diseases (CIBERER), Barcelona; [3]Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology, IDIBELL and CIBERESP, Hospitalet de Llobregat; [4]Structural Biology and Biocomputing Program, Spanish National Cancer Research Center (CNIO), Madrid; [5]Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología del Principado de Asturias, Universidad de Oviedo, Oviedo; [6]Department of Pathology, Bellvitge University Hospital, IDIBELL, Hospitalet de Llobregat; [7]Molecular Genetics Laboratory, Elche University Hospital, Elche; [8]Laboratorio de Oncología Molecular, Servicio de Oncología Médica, Hospital Clínico San Carlos, Madrid; [9]Instituto de Biología y Genética Molecular, IBGM-UVA-CSIC, Valladolid; [10]Familial Cancer Clinical Unit, Human Cancer Genetics Program, Spanish National Cancer Centre and Center for Biomedical Network Research on Rare Diseases; [11]Molecular Biology Laboratory, 12 de Octubre University Hospital, Madrid; [12]Hereditary Cancer Program, Catalan Institute of Oncology, IDIBGi, Girona; [13]Laboratorio de Oncología Molecular; [14]Familial Cancer Unit, Department of Medical Oncology, Instituto Universitario de Oncología del Principado de Asturias, Hospital Universitario Central de Asturias, Oviedo; [15]Department of Oncology, Hospital General Yagüe, Burgos; [16]Unit of Genetic Counseling in Cancer, Elche University Hospital, Elche; [17]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona; and [18]Translational Research Laboratory, Catalan Institute of Oncology, IDIBELL, Hospitalet de Llobregat, Spain

See related article, Yurgelun et al, on page 604.

Identification of genes associated with hereditary cancers facilitates management of patients with family histories of cancer. We performed exome sequencing of DNA from 3 individuals from a family with colorectal cancer who met the Amsterdam criteria for risk of hereditary nonpolyposis colorectal cancer. These individuals had mismatch repair-proficient tumors and each carried nonsense variant in the FANCD2/FANCI-associated nuclease 1 gene (*FAN1*), which encodes a nuclease involved in DNA inter-strand cross-link repair. We sequenced *FAN1* in 176 additional families with histories of colorectal cancer and performed in vitro functional analyses of the mutant forms of FAN1 identified. We detected *FAN1* mutations in approximately 3% of families who met the Amsterdam criteria and had mismatch repair-proficient cancers with no previously associated mutations. These findings link colorectal cancer predisposition to the Fanconi anemia DNA repair pathway, supporting the connection between genome integrity and cancer risk.

Keywords: Lynch Syndrome; Genetic Risk Factor; Susceptibility; DNA Mismatch Repair.

Familial aggregation of colorectal cancer (CRC) is one of the strongest risk factors for CRC. Germline mutations in the DNA mismatch repair (MMR) genes, *EPCAM*, *APC*, *MUTYH*, *POLE*, *POLD1*, *GREM1*, *SMAD4*, *BMPR1A*, *STK11*, and *PTEN* cause hereditary forms of CRC.[1–3] However, part of the observed heritability and familial aggregation of the disease is yet to be explained.

With the aim of identifying new hereditary CRC genes, we sequenced the exomes of 3 cancer-affected members of a high-risk, Amsterdam I MMR-proficient, CRC family (Figure 1*A*, Family 1). Of 32 unreported or rare (minor allele frequency <1%) nonsynonymous variants shared by all affected relatives (Supplementary Table 1), a nonsense mutation in *FAN1*, c.141C>A (p.C47*) deserved our attention, as the coded protein, FANCD2/FANCI-associated nuclease 1 (MIM# 613534), is involved in interstrand cross-link repair (Fanconi anemia [FA]) and interacts with MMR components, such as MLH1, PMS2 and PMS1, thus playing a role in maintaining genome integrity.[4–8] The identified *FAN1* mutation had not been reported previously (NHLBI GO Exome Sequencing Project [ESP], 1000 Genomes Project) or found in 1648 alleles of Spanish origin, including

*Authors share co-first authorship.

Abbreviations used in this paper: **CRC, colorectal cancer; FA, Fanconi anemia; MMC, mitomycin C; MMR, DNA mismatch repair; TCGA, The Cancer Genome Atlas.**

Most current article

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2015.05.056

Research Article

TRANSPARENT PROCESS OPEN ACCESS

EMBO Molecular Medicine

Comprehensive establishment and characterization of orthoxenograft mouse models of malignant peripheral nerve sheath tumors for personalized medicine

Joan Castellsagué[1,2,†], Bernat Gel[3,†], Juana Fernández-Rodríguez[1,2,†], Roger Llatjós[4], Ignacio Blanco[1], Yolanda Benavente[5], Diana Pérez-Sidelnikova[6], Javier García-del Muro[7], Joan Maria Viñals[6], August Vidal[4], Rafael Valdés-Mas[8], Ernest Terribas[3], Adriana López-Doriga[1,2], Miguel Angel Pujana[2], Gabriel Capellá[1,2], Xose S Puente[8], Eduard Serra[3,***], Alberto Villanueva[2,**] & Conxi Lázaro[1,2,*]

Abstract

Malignant peripheral nerve sheath tumors (MPNSTs) are soft-tissue sarcomas that can arise either sporadically or in association with neurofibromatosis type 1 (NF1). These aggressive malignancies confer poor survival, with no effective therapy available. We present the generation and characterization of five distinct MPNST orthoxenograft models for preclinical testing and personalized medicine. Four of the models are patient-derived tumor xenografts (PDTX), two independent MPNSTs from the same NF1 patient and two from different sporadic patients. The fifth model is an orthoxenograft derived from an NF1-related MPNST cell line. All MPNST orthoxenografts were generated by tumor implantation, or cell line injection, next to the sciatic nerve of nude mice, and were perpetuated by 7–10 mouse-to-mouse passages. The models reliably recapitulate the histopathological properties of their parental primary tumors. They also mimic distal dissemination properties in mice. Human stroma was rapidly lost after MPNST engraftment and replaced by murine stroma, which facilitated genomic tumor characterization. Compatible with an origin in a catastrophic event and subsequent genome stabilization, MPNST contained highly altered genomes that remained remarkably stable in orthoxenograft establishment and along passages. Mutational frequency and type of somatic point mutations were highly variable among the different MPNSTs modeled, but very consistent when comparing primary tumors with matched orthoxenografts generated. Unsupervised cluster analysis and principal component analysis (PCA) using an MPNST expression signature of ~1,000 genes grouped together all primary tumor–orthoxenograft pairs. Our work points to differences in the engraftment process of primary tumors compared with the engraftment of established cell lines. Following standardization and extensive characterization and validation, the orthoxenograft models were used for initial preclinical drug testing. Sorafenib (a BRAF inhibitor), in combination with doxorubicin or rapamycin, was found to be the most effective treatment for reducing MPNST growth. The development of genomically well-characterized preclinical models for MPNST allowed the evaluation of novel therapeutic strategies for personalized medicine.

Keywords MPNST; NF1; patient-derived tumor xenograft; preclinical mouse models; sorafenib

Subject Categories Cancer; Neuroscience

DOI 10.15252/emmm.201404430 | Received 13 August 2014 | Revised 24 February 2015 | Accepted 25 February 2015 | Published online 25 March 2015

EMBO Mol Med (2015) 7: 608–627

1 Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain
2 Translational Research Laboratory ICO-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
3 Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Badalona, Barcelona, Spain
4 Pathology Service, HUB-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
5 Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program ICO-IDIBELL and CIBER Epidemiología y Salud Pública (CIBERESP), L'Hospitalet de Llobregat, Barcelona, Spain
6 Plastic Surgery Service HUB-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
7 Medical Oncology Service ICO-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
8 Instituto Universitario de Oncología del Principado de Asturias (IUOPA), Universidad de Oviedo, Oviedo, Spain
*Corresponding author. Tel: +34 93 2607342; Fax: +34 93 2607466; E-mail: clazaro@iconcologia.net
**Corresponding author. Tel: +34 93 2607952; Fax: +34 93 2607466; E-mail: avillanueva@iconcologia.net
***Corresponding author. Tel: +34 93 5543067; Fax: +34 93 4651472; E-mail: eserra@imppc.org
†These authors contributed equally to this work

Carcinogenesis vol.35 no.9 pp.2039–2046, 2014
doi:10.1093/carcin/bgu092
Advance Access publication April 23, 2014

Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis

Adria Closa[1,2,†], David Cordero[1,2,†], Rebeca Sanz-Pamplona[1,2], Xavier Solé[1,2], Marta Crous-Bou[1,2], Laia Paré-Brunet[1,2], Antoni Berenguer[1,2], Elisabet Guino[1,2], Adriana Lopez-Doriga[1,2], Jordi Guardiola[5], Sebastiano Biondo[3,6], Ramon Salazar[4] and Victor Moreno[1,2,3,*]

[1]Cancer Prevention and Control Program, Catalan Institute of Oncology, and Consortium for Biomedical Research on Epidemiology and Public Health (CIBERESP), Barcelona E08907, Spain, [2]Colorectal Cancer Group, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona E08907, Spain, [3]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona E08907, Spain, [4]Medical Oncology Service, Catalan Institute of Oncology, Barcelona E08907, Spain, [5]Gastroenterology Service, Bellvitge University Hospital, Barcelona E08907, Spain and [6]General and Digestive Surgery Service, Bellvitge University Hospital, Barcelona E08907, Spain

*To whom correspondence should be addressed. Tel: +34 932 607 186; Fax: +34 932 607 188;
Email: v.moreno@iconcologia.net

In this study, we aim to identify the genes responsible for colorectal cancer risk behind the loci identified in genome-wide association studies (GWAS). These genes may be candidate targets for developing new strategies for prevention or therapy. We analyzed the association of genotypes for 26 GWAS single nucleotide polymorphisms (SNPs) with the expression of genes within a 2 Mb region (*cis*-eQTLs). Affymetrix Human Genome U219 expression arrays were used to assess gene expression in two series of samples, one of healthy colonic mucosa (n = 47) and other of normal mucosa adjacent to colon cancer (n = 97, total 144). Paired tumor tissues (n = 97) were also analyzed but did not provide additional findings. Partial Pearson correlation (r), adjusted for sample type, was used for the analysis. We have found Bonferroni-significant *cis*-eQTLs in three loci: rs3802842 in 11q23.1 associated to *C11orf53*, *COLCA1* (*C11orf92*) and *COLCA2* (*C11orf93*; r = 0.60); rs7136702 in 12q13.12 associated to *DIP2B* (r = 0.63) and rs5934683 in Xp22.3 associated to *SHROOM2* and *GPR143* (r = 0.47). For loci in chromosomes 11 and 12, we have found other SNPs in linkage disequilibrium that are more strongly associated with the expression of the identified genes and are better functional candidates: rs7130173 for 11q23.1 (r = 0.66) and rs61927768 for 12q13.12 (r = 0.86). These SNPs are located in DNA regions that may harbor enhancers or transcription factor binding sites. The analysis of *trans*-eQTLs has identified additional genes in these loci that may have common regulatory mechanisms as shown by the analysis of protein–protein interaction networks.

Introduction

Genome-wide association studies (GWAS) have been successful in identifying susceptibility loci for cancer and other diseases, but no progress has been made regarding the functional mechanisms underlying the associations. In colorectal cancer (CRC), 26 susceptibility single nucleotide polymorphisms (SNPs) in 23 different loci have been identified in GWAS to date (Supplementary Table 1, available at *Carcinogenesis* Online) (1–12). Most of them are located in intergenic positions and the genes responsible for the risk modification are unknown. The identification of these genes is important because they may be considered targets for developing new strategies for prevention or therapy (13).

The combination between high throughput genotyping and gene expression profiling technologies allows studying genome-wide associations between genetic polymorphisms and gene expression levels, known as expression quantitative trait loci (eQTL). The identification of eQTL has been proposed as a method to find genes underlying the associations with disease risk (14). The eQTL analysis also has been proposed as a tool to improve the power of GWAS (15) or to engineer genetic-gene expression networks and discover new mechanisms or pathways related to disease (16).

Most analyses of eQTL have used lymphoblastoid cell lines (14), which may not be optimal when the interest is in explaining risk in specific target tissues. Global eQTL analyses of diverse tissues have been done in liver (17), kidney (18) and brain (19), among others. The Genotype-Tissue Expression (GTEx) project (20) aims to create a comprehensive public atlas of gene expression and regulation across multiple human tissues (http://www.broadinstitute.org/gtex). Regarding colon cancer, the interest of analyzing eQTL for GWAS SNPs has been recognized (21) and some of the articles reporting GWAS SNPs have analyzed expression levels in reduced sets of tumors or lymphoblastoid cell lines to document a potential functional role of the SNPs (1,3,5,9). More recently, Loo *et al.* have found interesting associations using expression data assessed in colonic mucosa, either from tumor or normal mucosa adjacent to tumor, though the limited sample size provided low power to identify small associations (22).

In this study, we analyze *cis*- and *trans*-eQTL for GWAS SNPs to identify candidate genes responsible for CRC susceptibility. We combine two series of samples, one of healthy colonic mucosa and other of normal mucosa adjacent to colon cancer. In parallel, we have also analyzed the effect in tumor mucosa, but these data are more difficult to interpret because the expression in tumors is more heterogeneous and is highly altered by diverse mechanisms.

Materials and methods

Subjects and samples

Colon tumor and paired adjacent normal mucosa tissue samples used in this study were selected from a series of cases with a new diagnosis of colon adenocarcinoma attending the University Hospital of Bellvitge in Barcelona between January 1996 and December 2000. Patients included were diagnosed of stage II, microsatellite stable colon cancer, were surgically treated and had not received adjuvant chemotherapy. Adjacent mucosa was obtained from the proximal surgical margins and was at least 10 cm distant from the tumor lesion. Healthy colon mucosa samples were obtained during colonoscopy between February and May 2010. These samples come from a series of unselected patients who underwent a colonoscopy indicated for screening or suspicion of colonic pathology but no colonic lesions were observed. Biopsies were obtained from left and right colon. For this study, we selected randomly approximately half from each site (Supplementary Table 2, available at *Carcinogenesis* Online).

All subjects provided written informed consent to participate in the study and the ethics committee of the hospital cleared the protocol. Additional information about the study can be found at http://www.colonomics.org.

The eQTL analysis was focused on expression data assessed in normal mucosa. Though we initially selected 100 patients and 50 healthy controls, the final sample size after quality control of the data was 144: 47 from healthy donors and 97 adjacent normal mucosa from patients. Gene expression in tumors (n = 97) was also analyzed, and the results compared with those of normal mucosa. Also, for completeness and because we have previously demonstrated in these same samples that the expression in some genes is different between adjacent normal and healthy mucosa (23), we have performed the analyses separated for each tissue (Supplementary File 1, available at *Carcinogenesis* Online).

Abbreviations: CRC, colorectal cancer; GWAS, genome-wide association studies; LD, linkage disequilibrium; PPIN, protein–protein interaction network; SNP, single nucleotide polymorphism.

†These authors contributed equally to this work

Exome sequencing identifies *MUTYH* mutations in a family with colorectal cancer and an atypical phenotype

Ma *et al*[1] comprehensively assessed the association of previously reported genetic variants with colorectal cancer (CRC) risk. The meta-analyses revealed strong evidence for association with rare *MUTYH* variants, even when excluding cases with MUTYH-associated polyposis. An article by Nieuwenhuis *et al*[2] accurately defined the phenotypical features of MUTYH-associated polyposis. However, the study was performed on clinic-based series ascertained based on the inheritance model or the presence of polyps, which may miss additional phenotypes relevant to improve the disease characterisation and therefore, its genetic diagnosis. To illustrate this, we report a family with a clinical phenotype that resembled Lynch syndrome but was caused by *MUTYH* mutations.

To identify novel hereditary CRC genes, we studied an Amsterdam I family (hereditary non-polyposis CRC) with no mutations in the DNA mismatch repair (MMR) genes (figure 1, table 1). By exome sequencing performed on four cancer-affected (II.2, II.6, III.1 and III.6) and one cancer-free (III.5) family members, we identified a total of 11 unreported or rare heterozygous variants present in the cancer-affected individuals (see online supplementary table S1). One of them was *MUTYH* c.1147delC (p.Ala385Profs*23) (NM_001128425.1), an European recurrent mutation.[3] Cancer-affected individuals of the third generation also carried a novel *MUTYH* variant: c.43A>G (p.Met15Val). The presence of biallelic *MUTYH* mutations in II.2 and II.6 that could explain the pseudodominant transmission was discarded.

The functional studies performed for c.43A>G are described in online supplementary material. The variant inactivates the start codon of the two transcripts encoding the nuclear MUTYH isoforms (β and γ), highly relevant in ascending colon,[4] the location of ≥five tumours developed in the family.

KRAS c.34G>T (p.G12C), hallmark of MUTYH-associated carcinomas,[5] was present in the tumours developed by two *MUTYH* biallelic mutation carriers but not in the tumour developed by II.6, carrier of only c.1147delC. The MMR-deficient tumour developed by a *MUTYH* biallelic mutation carrier had a transition in the same *KRAS* codon; c.35G>A (p.G12D) (table 1).

The features of this family suggest that the selection criteria proposed for *MUTYH* testing,[6] might fail to detect a number of mutated families due to infrequent phenotypes. First, *MUTYH* heterozygous mutations may, probably in the presence of other cancer risk factors, provide an increased risk of developing cancer in heterozygous carriers,[5 7] and thereby disguise the *MUTYH* recessive inheritance to look like autosomal-dominant. One should be suspicious when an extreme anticipation in the age of cancer onset is observed between two affected generations. Second, absence or scarcity of polyps, even at relatively advanced ages (early 50 s) and with a prior CRC diagnosis, can occur in biallelic mutation carriers. When this occurs in several cancer-affected mutation carriers within the same family it can lead to a misdiagnosis of hereditary non-polyposis CRC. Third, the presence of MMR-deficient tumours should not be an exclusion criterion for *MUTYH* genetic screening.[5] Finally, as previously

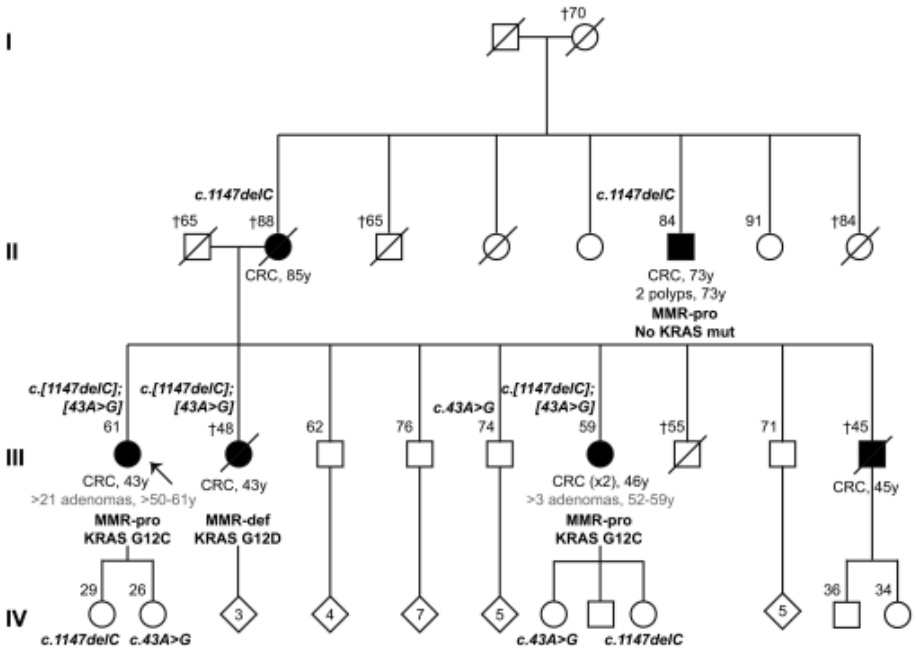


Figure 1 Family pedigree. Ages at information gathering or at death, when available, are indicated on the top left corner of each individual's symbol. Germline *MUTYH* mutations identified are indicated on the top left corner (above the age). Tumour *KRAS* mutations in codon 12 are indicated below the CRC and/or polyposis diagnosis information. In grey, number of adenomas identified at follow-up screenings. Filled symbol, CRC; numbers within the symbols, number of children; arrow, index case. CRC, colorectal cancer; y, years at cancer or polyposis diagnosis; MMR-pro, DNA mismatch repair proficiency in the tumour; MMR-def, DNA mismatch repair deficiency in the tumour.

Colorectal Disease

Original article

doi:10.1111/codi.12028

Clinicopathological risk factors of Stage II colon cancer: results of a prospective study

C. Santos*, A. López-Doriga†, M. Navarro‡, J. Mateo*, S. Biondo§, M. Martínez Villacampa*, G. Soler*, X. Sanjuan¶, M. J. Paules¶, B. Laquente*, E. Guinó†, E. Kreisler§, R. Frago§, J. R. Germà*, V. Moreno† and R. Salazar*

*Department of Medical Oncology, †Bioinformatics and Biostatistics Unit, Department of Epidemiology, and ‡Cancer Genetic Counseling Program, Institut Català d'Oncologia – Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain, §Department of Surgery and ¶Department of Pathology, Hospital Universitari de Bellvitge – IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

Received 28 February 2012; accepted 30 July 2012; Accepted Article online 13 September 2012

Abstract

Aim Adjuvant 5-fluorouracil based chemotherapy has demonstrated benefit in Stage III colon cancer but still remains controversial in Stage II. The aim of this study was to analyse the prognostic impact of clinicopathological factors that may help guide treatment decisions in Stage II colon cancer.

Method Between 1996 and 2006 data from patients diagnosed with colorectal cancer at Hospital Universitari Bellvitge and its referral comprehensive cancer centre Institut Català d'Oncologia/L'Hospitalet were prospectively included in a database. We identified 432 patients with Stage II colon cancer operated on at Hospital Universitari Bellvitge. The 5-year relapse-free survival (RFS) and colon-cancer-specific survival (CCSS) were determined.

Results The 5-year RFS and CCSS were 83% and 88%, respectively. Lymphovascular or perineural invasion was associated with RFS [hazard ratio (HR) 1.84; 95% CI 1.01–3.35]. Gender (women, HR 0.48; 95% CI 0.23–1) and lymphovascular or perineural invasion (HR 3.51; 95% CI 1.86–6.64) together with pT4 (HR 2.79; 95% CI 1.44–5.41) influenced CCSS. In multivariate analysis pT4 and lymphovascular or perineural invasion remained significantly associated with CCSS. We performed a risk index with these factors with prognostic impact. Patients with pT4 tumours and lymphovascular or perineural invasion had a 5-year CCSS of 61% *vs* the 93% (HR 5.87; 95 CI 2.46–13.97) of those without any of these factors.

Conclusion pT4 and lymphatic, venous or perineural invasion are confirmed as significant prognostic factors in Stage II colon cancer and should be taken into account in the clinical validation process of new molecular prognostic factors.

Keywords Colon cancer, prognosis, adjuvant chemotherapy, colorectal surgery

What is new in this paper?

A risk index including T stage and vascular invasion can identify a subgroup of patients with Stage II colon cancer with low risk of recurrence and colon cancer death. We suggest that patients classified as low risk patients could be safely managed without adjuvant chemotherapy.

Introduction

Colorectal cancer is the second cause of cancer related death in Europe [1]. The gold standard treatment in non-metastatic disease is surgery, with 5-year survival rates ranging from 44% to 93% depending on stage [2].

5-Fluorouracil (5-FU) based adjuvant treatment has demonstrated a significant benefit in Stage III colon cancer (CC) [3,4] but controversy still remains for Stage II patients [5,6]. Different consensus and expert opinion reports have been published over the last few years in an effort to define a high risk Stage II subgroup that may benefit from adjuvant 5-FU based chemotherapy [7,8]. The most robust clinicopathological factor associated with poor prognosis in the literature is pT4 [2,9–11]. High histological grade [2,12], number of lymph nodes assessed < 12 [13,14], venous, lymphatic or perineural invasion [9,15], emergency surgery (due to obstruction or perforation) [16,17] and high preoperative carcinoembryonic antigen (CEA) level [18,19] have also been associated with a high risk of relapse and death. In

Correspondence to: Ramon Salazar, Department of Medical Oncology, Institut Català d'Oncologia – Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Avinguda Gran Via 199–203, 08907 L'Hospitalet de Llobregat, Barcelona, Spain. E-mail: ramonsalazar@iconcologia.net

Carcinogenesis vol.33 no.3 pp.613–619, 2012
doi:10.1093/carcin/bgs009
Advance Access publication January 10, 2012

Susceptibility genetic variants associated with early-onset colorectal cancer

María Dolores Giráldez[1], Adriana López-Dóriga[2], Luis Bujanda[3], Anna Abulí[1,4], Xavier Bessa[4], Ceres Fernández-Rozadilla[5], Jenifer Muñoz[1], Miriam Cuatrecasas[1], Rodrigo Jover[6], Rosa M.Xicola[7], Xavier Llor[7], Josep M.Piqué[1], Angel Carracedo[5], Clara Ruiz-Ponte[5], Angel Cosme[3], José María Enríquez-Navascués[3], Victor Moreno[2], Montserrat Andreu[3], Antoni Castells[1], Francesc Balaguer[1], Sergi Castellví-Bel[1,*] and the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association[†]

[1]Department of Gastroenterology, Hospital Clínic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), IDIBAPS, University of Barcelona, Villarroel 170, 08036 Barcelona, Catalonia, Spain, [2]IDIBELL-Institut Català d'Oncologia (ICO), CIBER Epidemiología y Salud Pública (CIBERESP), University of Barcelona, L'Hospitalet de Llobregat, Barcelona, Spain, [3]Department of Gastroenterology, Hospital de Donostia, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), University of Basque Country, San Sebastian, Spain, [4]Gastroenterology Department, Parc de Salut Mar, Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Catalonia, Spain, [5]Galician Public Foundation of Genomic Medicine (FPGMX), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Genomics Medicine Group, Hospital Clínico, Santiago de Compostela, University of Santiago de Compostela, Galicia, Spain, [6]Department of Gastroenterology, Hospital General d'Alacant, Alicante, Spain and [7]Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, USA

*To whom correspondence should be addressed. Tel: 34 93 2275418; Fax: 34 93 2279387; Email: sbel@clinic.ub.es

Colorectal cancer (CRC) is the second most common cancer in Western countries. Hereditary forms only correspond to 5% of CRC burden. Recently, genome-wide association studies have identified common low-penetrant CRC genetic susceptibility loci. Early-onset CRC (CRC<50 years old) is especially suggestive of hereditary predisposition although 85–90% of heritability still remains unidentified. CRC<50 patients (n = 191) were compared with a late-onset CRC group (CRC>65 years old) (n = 1264). CRC susceptibility variants at 8q23.3 (rs16892766), 8q24.21 (rs6983267), 10p14 (rs10795668), 11q23.1 (rs3802842), 15q13.3 (rs4779584), 18q21 (rs4939827), 14q22.2 (rs4444235), 16q22.1 (rs9929218), 19q13.1 (rs10411210) and 20p12.3 (rs961253) were genotyped in all DNA samples. A genotype–phenotype correlation with clinical and pathological characteristics in both groups was performed. Risk allele carriers for rs3802842 [Odds ratio (OR) = 1.5, 95% confidence interval (CI) 1.1–2.05, P = 0.0096, dominant model) and rs4779584 (OR = 1.39, 95% CI 1.02–1.9, P = 0.0396, dominant model) were more frequent in the CRC<50 group, whereas homozygotes for rs10795668 risk allele were also more frequent in the early-onset CRC (P = 0.02, codominant model). Regarding early-onset cases, 14q22 (rs4444235), 11q23 (rs3802842) and 20p12 (rs961253) variants were more associated with family history of CRC or tumors of the Lynch syndrome spectrum excluding CRC. In our entire cohort, sum of risk alleles was significantly higher in patients with a CRC family history (OR = 1.40, 95% CI 1.06–1.85, P = 0.01). In conclusion, variants at 10p14 (rs10795668), 11q23.1 (rs3802842) and 15q13.3 (rs4779584) may have a predominant role in predisposition to early-onset CRC. Association of CRC susceptibility variants with some patient's familiar and personal features could be relevant for screening and surveillance strategies in this high-risk group and it should be explored in further studies.

Abbreviations: CI, confidence interval; CRC, colorectal cancer; LS, Lynch syndrome; MSI, microsatellite instability; OR, odds ratio; SNP, single-nucleotide polymorphism.

[†]All authors are listed in a Supplementary Note.

Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related death in Western countries (1). As other complex diseases, CRC is caused by both genetic and environmental factors. Although environmental factors, such as smoking and diet are undoubtedly major risk factors for CRC, twin studies have shown that 30% of the variation in CRC susceptibility involves inherited genetic differences (2,3). Familial adenomatous polyposis and Lynch syndrome (LS) are the more frequent hereditary CRC syndromes and they are caused by germline mutations in *APC* or *MUTYH* and DNA repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*), respectively. Hereditary CRC forms account only for a minority of the total CRC burden (5%). The genetic components involved in these less frequent hereditary forms were successfully identified in the past two decades and they correspond to rare highly penetrant alleles that predispose to CRC (4).

According to the common disease-common variant hypothesis, a majority of the heritability in CRC may be explained by multiple common genetic variants with a low-moderate effect on cancer susceptibility (5). The common disease-common variant hypothesis has been recently vindicated by genome-wide association studies and, so far, 14 common low-penetrant genetic *loci* have been identified for CRC susceptibility on 8q24.21, 18q21.1, 15q13.3, 8q23.3, 10p14, 11q23.1, 14q22.2, 16q22.1, 19q13, 20p12.3, 1q41, 3q26.2, 12q13.13 and 20q13.33 (6–8). In addition, some of these variants appear to be associated with some clinical and familial features, which could have potential implications for screening and surveillance strategies (9,10).

Early-onset CRC (<50 years old) is especially suggestive of a hereditary predisposition and it can be used in genetic association studies to increase likelihood of finding susceptibility variants. We have shown previously that LS- and *MUTYH*-associated CRC account for only 15–20% of the early-onset CRC cases (11). Therefore, a considerable proportion of heritability still remains unidentified in this high-risk group. Hypothesizing that some of these susceptibility variants could be distinctively related to early-onset CRC, the aims of our study were firstly to assess the prevalence of 10 common, low-penetrance CRC susceptibility variants at 8q23.3 (rs16892766), 8q24.21 (rs6983267), 10p14 (rs10795668), 11q23.1 (rs3802842), 15q13.3 (rs4779584), 18q21 (rs4939827), 14q22.2 (rs4444235), 16q22.1 (rs9929218), 19q13.1 (rs10411210) and 20p12.3 (rs961253) in a group of unselected early-onset CRC patients with no involvement of hereditary syndromes, and secondly, to perform a genotype–phenotype correlation of these variants with clinical and pathological characteristics.

Materials and methods

Study population

We retrospectively recruited all patients <50 years old diagnosed with CRC (early-onset CRC) who were surgically treated at two Spanish centers between 1995 and 2007 (Hospital Clínic of Barcelona and Hospital of Donostia), with available archival formalin-fixed paraffin-embedded samples (11). Patients with personal history of colorectal polyposis, inflammatory bowel disease, biallelic *MUTYH* mutations, LS diagnosis (germline mutation carriers) or LS-suspected (mismatch repair deficiency without *MLH1* promoter methylation) were excluded. A total of 115 patients from both centers were considered for this study. Additionally, 76 early-onset CRC patients were also included from the EPICOLON cohort (12) using the same previous exclusion criteria.

NEOPLASIA
www.neoplasia.com

Volume 13 Number 10 October 2011 pp. 931–946 931

Hepatic Carcinoma–Associated Fibroblasts Promote an Adaptative Response in Colorectal Cancer Cells That Inhibit Proliferation and Apoptosis: Nonresistant Cells Die by Nonapoptotic Cell Death[1]

Mireia Berdiel-Acer*,†, Monika E. Bohem*, Adriana López-Doriga‡, August Vidal§, Ramon Salazar¶, Maria Martínez-Iniesta*, Cristina Santos¶, Xavier Sanjuan§, Alberto Villanueva* and David G. Mollevi*

*Laboratory of Translational Research, Institut Català d'Oncologia-IDIBELL, Hospitalet de Llobregat, Barcelona, Spain; †Department of Medicine, Autonomous University of Barcelona, Bellaterra, Barcelona, Spain; ‡Bioinformatics Unit, Institut Català d'Oncologia-IDIBELL, Hospitalet de Llobregat, Barcelona, Spain; §Pathology Department, Hospital Universitari de Bellvitge-IDIBELL, Hospitalet de Llobregat, Barcelona, Spain; ¶Medical Oncology Department, Institut Català d'Oncologia-IDIBELL, Hospitalet de Llobregat, Barcelona, Spain

Abstract

Carcinoma-associated fibroblasts (CAFs) are important contributors of microenvironment in determining the tumor's fate. This study aimed to compare the influence of liver microenvironment and primary tumor microenvironment on the behavior of colorectal carcinoma. Conditioned medium (CM) from normal colonic fibroblasts (NCFs), CAFs from primary tumor (CAF-PT) or liver metastasis (CAF-LM) were obtained. We performed functional assays to test the influence of each CM on colorectal cell lines. Microarray and gene set enrichment analysis (GSEA) were performed in DLD1 cells cultured in matched CM. In DLD1 cells, CAF-LM CM compared with CAF-PT CM and NCF led to a more aggressive phenotype, induced the features of an epithelial-to-mesenchymal transition more efficiently, and stimulated migration and invasion to a greater extent. Sustained stimulation with CAF-LM CM evoked a transient G_2/M cell cycle arrest accompanied by a reduction of apoptosis, inhibition of proliferation, and decreased viability of SW1116, SW620, SW480, DLD1, HT-29, and Caco-2 cells and provoked nonapoptotic cell death in those cells carrying *KRAS* mutations. Cells resistant to CAF-LM CM completely changed their morphology in an extracellular signal–regulated protein kinase–dependent process and depicted an increased stemness capacity alongside the Wnt pathway stimulation. The transcriptomic profile of DLD1 cells treated with CAF-LM CM was associated with Wnt and mitogen-activated protein kinase pathways activation in GSEA. Therefore, the liver microenvironment induces more efficiently the aggressiveness of colorectal cancer cells than other matched microenvironments do but secondarily evokes cell death. Resistant cells displayed higher stemness capacity.

Neoplasia (2011) 13, 931–946

Abbreviations: CAF-PT, carcinoma-associated fibroblasts from primary tumor; CAF-LM, carcinoma-associated fibroblasts from liver metastasis; NCF, normal colonic fibroblast; CM, conditioned medium; GSEA, gene set enrichment analysis

Address all correspondence to: David G. Mollevi, PhD, Laboratori de Recerca Translacional, Institut Català d'Oncologia-IDIBELL, Hospital Duran i Reynals, 3a planta, Av Gran Via 199-203, 08907 L'Hospitalet de Llobregat, Barcelona, Spain. E-mail: dgmollevi@iconcologia.net

[1]This work was supported by a grant from the Spanish Ministry of Health (FIS PI07-0657). The authors declare that there are no conflicts of interests.
Received 17 May 2011; Revised 16 August 2011; Accepted 23 August 2011

DOI 10.1593/neo.11706

VOLUME 29 · NUMBER 1 · JANUARY 1 2011

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer

Ramon Salazar, Paul Roepman, Gabriel Capella, Victor Moreno, Iris Simon, Christa Dreezen, Adriana Lopez-Doriga, Cristina Santos, Corrie Marijnen, Johan Westerga, Sjoerd Bruin, David Kerr, Peter Kuppen, Cornelis van de Velde, Hans Morreau, Loes Van Velthuysen, Annuska M. Glas, Laura J. Van't Veer, and Rob Tollenaar

From the Institut Català d'Oncologia–Biomedical Research Institute of Bellvitge, L'Hospitalet de Llobregat; University of Barcelona, Barcelona, Spain; Agendia; Netherlands Cancer Institute; Slotervaart Hospital, Amsterdam; Leiden University Medical Center, Leiden, the Netherlands; and University of Oxford, Radcliffe Infirmary, Oxford, United Kingdom.

Submitted May 5, 2010; accepted September 7, 2010; published online ahead of print at www.jco.org on November 22, 2010.

RNA isolation and hybridization of the samples and part of the analysis were performed and funded at Agendia. The training set of the study was partly supported by the Leiden Medical Centre Institutional Grant and by the Dutch Genomics Initiative Cancer Genomics Center in the Netherlands Cancer Institute. The validation of the study was partly supported by the Catalan Institute of Oncology and the Private Foundation of the Biomedical Research Institute of Bellvitge, the Spanish Ministry of Science (Grants No. SAF 06-6084 and SAF 2009-07319), the Instituto de Salud Carlos III (Grants No. PI08-1635 and PI09-01037), Spanish Networks Red Temática de Investigación Cooperativa en Cáncer (Grant No. RD06/0020/1050), Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública G55 and the Acción Transversal del Cancer, the Catalan Government Departament d'Universitats, Recerca i Sociedad de la Información (Grants No. 2009SGR1489 and 2009SGR290), the European Commission (Grant No. FP7-COOP-Health-2007-B), HiperDart, and Fundació Gastroenterologia Dr Francisco Vilardell.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Ramon Salazar, MD, Institut Català d'Oncologia-IDIBELL, L'Hospitalet de Llobregat, Av Gran Via 199-203, Barcelona, Spain 08907; e-mail: ramonsalazar@iconcologia.net.

0732-183X/11/2901-17/$20.00

DOI: 10.1200/JCO.2010.30.1077

ABSTRACT

Purpose
This study aims to develop a robust gene expression classifier that can predict disease relapse in patients with early-stage colorectal cancer (CRC).

Patients and Methods
Fresh frozen tumor tissue from 188 patients with stage I to IV CRC undergoing surgery was analyzed using Agilent 44K oligonucleotide arrays. Median follow-up time was 65.1 months, and the majority of patients (83.6%) did not receive adjuvant chemotherapy. A nearest mean classifier was developed using a cross-validation procedure to score all genes for their association with 5-year distant metastasis–free survival.

Results
An optimal set of 18 genes was identified and used to construct a prognostic classifier (ColoPrint). The signature was validated on an independent set of 206 samples from patients with stage I, II, and III CRC. The signature classified 60% of patients as low risk and 40% as high risk. Five-year relapse-free survival rates were 87.6% (95% CI, 81.5% to 93.7%) and 67.2% (95% CI, 55.4% to 79.0%) for low- and high-risk patients, respectively, with a hazard ratio (HR) of 2.5 (95% CI, 1.33 to 4.73; $P = .005$). In multivariate analysis, the signature remained one of the most significant prognostic factors, with an HR of 2.69 (95% CI, 1.41 to 5.14; $P = .003$). In patients with stage II CRC, the signature had an HR of 3.34 ($P = .017$) and was superior to American Society of Clinical Oncology criteria in assessing the risk of cancer recurrence without prescreening for microsatellite instability (MSI).

Conclusion
ColoPrint significantly improves the prognostic accuracy of pathologic factors and MSI in patients with stage II and III CRC and facilitates the identification of patients with stage II disease who may be safely managed without chemotherapy.

J Clin Oncol 29:17-24. © 2010 by American Society of Clinical Oncology

INTRODUCTION

The American Joint Committee on Cancer TNM staging system is the current standard for determining the prognosis of patients with colorectal cancer (CRC). Patients with stage I CRC have a 5-year survival rate of approximately 93%, which decreases to approximately 80% for patients with stage II disease and to 60% for patients with stage III disease.[1] Despite numerous clinical trials, the benefit of adjuvant chemotherapy for patients with stage II CRC is still debatable.[2-4] In Western countries, official guidelines give suggestions for risk stratification but no clear recommendations on the administration of adjuvant chemotherapy.[5] In contrast, adjuvant treatment is universally recommended for all patients with stage III disease.[6] However, patients with T1-2N1M0 tumors (stage IIIA) have significantly higher survival rates than patients with stage IIB tumors,[1] suggesting that adjuvant chemotherapy selection needs optimization.

To date, substantial effort has been put into the identification of clinicopathologic parameters that predict prognosis of patients with stage II disease. The most important factors for predicting the risk of systemic recurrence (ie, distant metastases) are emergency presentation, poorly differentiated tumor, depth of tumor invasion, and adjacent organ involvement (T4).[5,7] Inadequate sampling of lymph nodes is an additional risk factor.[8] Among the molecular factors investigated as prognostic candidates in early CRCs, microsatellite instability (MSI) is the