

DETECCIÓ DE COPY NUMBER VARIANTS MITJANÇANT SEQÜENCIACIÓ D'ALT RENDIMENT EN LA MORT SOBTADA CARDÍACA HEREDITÀRIA

Jesús Matés Ramírez

Per citar o enllaçar aquest document:

Para citar o enlazar este documento:

Use this url to cite or link to this publication:

<http://hdl.handle.net/10803/482044>

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



TESI DOCTORAL

**DETECCIÓ DE COPY NUMBER VARIANTS MITJANÇANT
SEQÜENCIACIÓ D'ALT RENDIMENT EN LA MORT SOBTADA
CARDÍACA HEREDITÀRIA**

**Jesús Matés Ramírez
2017**



TESI DOCTORAL

**DETECCIÓ DE COPY NUMBER VARIANTS MITJANÇANT
SEQÜENCIACIÓ D'ALT RENDIMENT EN LA MORT SOBTADA
CARDÍACA HEREDITÀRIA**

**Jesús Matés Ramírez
2017**

Programa de Doctorat en Biologia Molecular, Biomedicina i Salut

Dirigida pel Dr. Ramon Brugada i Terradellas
Tutoritzada pel Dr. Ramon Brugada i Terradellas

Codirigida pel Dr. Carles Ferrer Costa

Codirigida per la Dra. Catarina Allegue Toscano

Memòria presentada per optar al títol de doctor per la Universitat de Girona

5 Annexes

Llista de publicacions derivades de la tesi

1. Irene Mademont-Soler*, **Jesus Mates***, Raquel Yotti, Maria Angeles Espinosa, Alexandra Pérez-Serra, Ana Isabel Fernández-Avila, Monica Coll, Irene Méndez, Anna Iglesias, Bernat del Olmo, Helena Riuró, Sofía Cuenca, Catarina Allegue, Oscar Campuzano, Ferran Picó, Carles Ferrer-Costa, Patricia Álvarez, Sergio Castillo, Pablo Garcia-Pavia, Esther Gonzalez-Lopez, Laura Padron-Barthe, Aranzazu Díaz de Bustamante, María Teresa Darnaude, José Ignacio González-Hevia, Josep Brugada, Francisco Fernández-Aviles, Ramon Brugada. *Additional value of screening for minor genes and copy number variants in hypertrophic cardiomyopathy*. PLoS ONE. 2017. 12(8): e0181465.
2. Irene Mademont-Soler, Mel-lina Pinsach-Abuin, Helena Riuró, **Jesus Mates**, Alexandra Pérez-Serra, Mònica Coll, José Manuel Porres, Bernat del Olmo, Anna Iglesias, Elisabet Selga, Ferran Picó, Sara Pagans, Carles Ferrer-Costa, Geòrgia Sarquella-Brugada, Elena Arbelo, Sergi Cesar, Josep Brugada, Oscar Campuzano, Ramon Brugada. *Large Genomic Imbalances in Brugada Syndrome*. PLoS ONE 11(9): e0163514.
3. Oscar Campuzano, Georgia Sarquella-Brugada, Irene Mademont-Soler, Catarina Allegue, Sergi Cesar, Carles Ferrer-Costa, Monica Coll, **Jesus Mates**, Anna Iglesias, Josep Brugada, Ramon Brugada. *Identification of genetic alterations, as causative genetic defects in long QT syndrome, using next generation sequencing technology*. PLoS ONE 9(12): e114894.

Certs capítols d'aquesta tesi estan pendents de ser publicats. Es llisten a continuació.

4. **Jesus Mates***, Irene Mademont-Soler*, Bernat del Olmo, Carles Ferrer-Costa, Monica Coll, Alexandra Pérez-Serra, Ferran Picó, Catarina Allegue, Anna Fernandez-Falgueres, Patricia Álvarez, Raquel Yotti, Maria Angeles Espinosa, Georgia Sarquella-Brugada, Sergi Cesar, Ester Carro, Josep Brugada, Elena Arbelo, Pablo Garcia-Pavia, Mar Borregan, Eduardo Tizzano, Amador López-Granados, Francisco Mazuelos, Aranzazu Díaz de Bustamante, Maria Teresa Darnaude, José Ignacio González-Hevia, Felicitas Diaz-Flores, Francisco Trujillo, Anna Iglesias, Francisco Fernandez-Aviles, Oscar Campuzano, Ramon Brugada. *Role of copy number variants in sudden cardiac death and related diseases: genètic anàlisis and translation into clinical practice*. 2017. (Enviat a l'*European Journal of Human Genetics*).
5. **Jesus Mates**, Irene Mademont-Soler, Bernat del Olmo, Carles Ferrer-Costa, Catarina Allegue, Oscar Campuzano, Ramon Brugada. *Copy number variants in sudden cardiac death and related diseases. A review*. Manuscrit en preparació.

Llista d'abreviatures

- aCGH:** *Array comparative genomic hybridization*
- ASCII:** *American standard code for information interchange*
- ATP:** *Trifosfat d'adenosina (Adenosine triphosphate)*
- BA:** *Benigna*
- BAC:** *Bacterial artificial chromosome*
- BAM:** *Binary Alignment / Map*
- BE:** *Benignitat elevada*
- BED:** *Browser Extensible Data format*
- BLAST:** *Basic local alignment search tool*
- BS:** *Benignitat sospitada*
- BWA:** *Burrows-Wheeler Aligner*
- BWT:** *Burrows-Wheeler Transform*
- CCDS:** *Consensus coding sequence*
- CGH:** *Comparative genomic hybridization*
- CMA:** *Chromosomal microarray-based analysis*
- CNV:** *Copy number variant*
- DAI:** *Desfibril·lador Automàtic Implantable*
- DATA:** *Dissecció i aneurisma toràctic i d'aorta*
- dbSNP:** *Single Nucleotide Polymorphism database*
- DGV:** *Database of genomic variants*
- DNA:** *Àcid desoxiribonucleic (Deoxyribonucleic acid)*
- DNA-PK:** *Proteïna Cinasa dependent de DNA*
- DNA-PKcs:** *Subunitat catalítica de la proteïna cinasa dependent de DNA*
- dNTP:** *Desoxinucleòtid trifosfat*
- ECG:** *Electrocardiograma*
- emPCR:** *Reacció en cadena de la polimerasa en emulsió (emulsion Polymerase chain reaction)*
- ENCODE:** *Encyclopedia of DNA elements*
- ExAC:** *Exome Aggregation Consortium*
- FA:** *Fibril·lació auricular*
- FISH:** *Fluorescent in situ hybridization*
- FoSTeS:** *Fork stalling and template switching*
- GATK:** *Genome Analysis ToolKit*
- Gb:** *Gigabase*
- GC:** *Guanines i Citosines (Contingut GC)*
- GRCh37:** *Genome reference consortium human genome build 37*

GWAS: *Genome-wide association study*

HGMD: *Human Gene Mutation Database*

HGSC: Consorci per la Seqüenciació del Genoma Humà (*Human Genome Sequencing Consortium*)

Hg19: *Human genome version 19*

IGV: *Integrative genomic viewer*

IMLCFC: Institut de Medicina Legal i Ciències Forenses de Catalunya

Indel: Inserció o deleció

K/mm²: Clústers per mil·límetre quadrat

Kb: Kilobase

LCR: *Low copy repeat*

LINE: *Long interspersed nuclear element*

LTR: *Long Tandem Repeat*

MAF: *Minor Allele Frequency*

Mb: Megabase

MCA: Miocardiopatia arritmogènica

MCD: Miocardiopatia dilatada

MCH: Miocardiopatia hipertròfica

MCR: Miocardiopatia restrictiva

MLPA: *Multiplex Ligation-dependent Probe Amplification*

MOSCAT: Mort sobtada a Catalunya (Projecte)

MS: Mort sobtada

ms: Mil·lisegon

MSC: Mort sobtada cardíaca

MSI: Mort sobtada inexplicada

MSIU: Mort sobtada intrauterina

MSL: Mort sobtada del lactant

MSR: Mort sobtada recuperada

NCBI: *National center for biotechnology information*

NCVE: No compactació del ventricle esquerre

ng: Nanogram

NGS: *Next generation sequencing*

nM: Nanomolar

ORF: Marc obert de lectura (*Open reading frame*)

pb: Parells de bases

PCR: Reacció en cadena de la polimerasa (*Polymerase chain reaction*)

PDF: *Portable Document Format*

PE: Patogenicitat elevada

PM: Patogenicitat moderada
pM: Picomolar
PME: Patogenicitat molt elevada
PNG: *Portable Network Graphic format*
PPI: Anió pirofosfat inorgànic
PS: Patogenicitat sospitada
Q: Qualitat en escala Phred
qPCR: PCR quantitativa
RAG: *Recombination-activating gene*
RHA: Recombinació homòloga al·lèlica
RHNA: Recombinació homòloga no al·lèlica
RNA: Àcid ribonucleic (*Ribonucleic acid*)
SAM: *Sequence Alignment / Map*
SBr: Síndrome de Brugada
SIDA: Síndrome de la immunodeficiència adquirida
SINE: *Short interspersed nuclear element*
SM: Síndrome de Marfan
SMRT: *Single molecule real-time sequencing*
SNP: *single nucleotide polymorphism*
SNV: *single nucleotide variant*
SOLID: *Sequencing by oligo ligation and detection*
SQTC: Síndrome del QT curt
SQTL: Síndrome del QT llarg
SVA: Element genètic mòbil format per SINE, VNTR i Alu
TDC: Trencament de doble cadena
TVPC: Taquicàrdia ventricular polimòrfica catecolaminèrgica
UENH: Unió d'extrems no homòlegs
USB: *Universal Serial Bus*
UTR: *Untranslated region*
VB: Variant benigna
VCF: *Variant Call Format*
VIH: Virus de la immunodeficiència humana
VNTR: *Variable number tandem repeat*
VP: Variant patogènica
VPB: Variant probablement benigna
VPP: Variant probablement patogènica
VSI: Variant de significat incert

WES: *Whole Exome Sequencing*

WGS: *Whole Genome Sequencing*

ZMW: *Zero-mode waveguide*

µm: Micròmetre

µg: Microgram

Nota: al llarg de la tesi els gens es citen amb majúscula i cursiva. Les proteïnes apareixen citades amb majúscula, sense cursiva.

Índex de figures

Figura I-1. Autoradiografia d'un gel de seqüenciació. S'hi pot apreciar la seqüència de DNA, de 64 pb provinents de dues cadenes complementàries. Adaptada de Maxam & Gilbert 1977	4
Figura I-2. Evolució del cost (en dòlars) de seqüenciació d'un milió de bases de DNA al <i>National Genome Research Institute</i> . Adaptada de Wetterstrand 2015.	6
Figura I-3. Diferents estratègies d'immobilització de DNA motlle: A) emPCR; B) Amplificació en fase sòlida d'Illumina, mitjançant PCR en pont. Adaptada de Metzker 2009.	8
Figura I-4. Representació esquemàtica de dos mètodes de seqüenciació per síntesi. A) La reacció de piroseqüenciació, del mètode 454; B) La terminació reversible cíclica en 4 colors del mètode de seqüenciació d'Illumina. Adaptada de Metzker 2009.	10
Figura I-5. Procés de seqüenciació per lligació d'octàmers, utilitzat en el sistema SOLiD™. Adaptada de Metzker 2009.	13
Figura I-6. Procés de seqüenciació SMRT. Adaptada de Metzker 2009.	17
Figura I-7. Composició del genoma humà. Adaptada de Patrushev i Minkevich 2008.....	20
Figura I-8. Representació de la mida i la freqüència de les principals categories de variants genètiques. Adaptada de Girirajan et al. 2011.	24
Figura I-9. A) Mecanisme de reparació de TDC; B) Hibridació de cadena dependent de síntesi. Adaptada de Hastings et al 2003.	28
Figura I-10. Representació esquemàtica d'una unió de Holliday; l'estructura es resol amb creuament genètic. Adaptada de Berg et al 2012.	29
Figura I-11. Mecanisme de replicació induïda per trencament. Adaptada de Hastings et al 2003.	30
Figura I-12. Mecanismes de formació de creuaments genètics: A) RHNA; B) Hibridació de cadena única. Adaptada de Hastings et al 2003.	32
Figura I-13. A) Mecanisme d'UENH; B) UENH mediada per homologia. Adaptada de Berg et al 2012	34
Figura I-14. Cicle de trencament – fusió – pont. Adaptada de Hastings et al 2003.	36
Figura I-15. Esquema del mecanisme de A) Lliscament durant la replicació i B) FoStES. Adaptada de Hastings et al 2003 i de Zhou et al. 2014.	37
Figura I-16. Mecanisme de replicació induïda per trencament intervinguda per microhomologia. Adaptada de Hastings et al 2009.	39
Figura I-17. Transmissió de fenotips clínics mitjançant CNVs. Adaptada de Lupski 1998.	41
Figura I-18. A) Cariotip tradicional; B) Cariotip marcat amb FISH; C) Duplicació de 3 exons a <i>PKP2</i> per MLPA; D) Detecció de CNVs mitjançant un <i>array</i> de SNPs. Adaptada de Koboldt 2017 i Shapiro 2017.	47
Figura I-19. Diferents aproximacions de detecció de CNVs mitjançant mètodes bioinformàtics: A) Comparació de cobertures; B) Informació de seqüències aparellades; C) Mitjançant <i>split-reads</i> . Adaptada de Tattini et al 2015	51
Figura 3-1. Distribució dels pacients de la cohort d'estudi en funció del diagnòstic.	78
Figura 3-2. Gràfics de control de qualitat generats en un <i>run</i> de MiSeq.	85

Figura 3-3. Captures de pantalla del visor genòmic IGV (<i>Broad Institute</i> , Cambridge, USA). S'hi mostren les seqüències alineades a una regió exònica del cromosoma X. A) Seqüències de 151 pb sense processar, amb adaptadors i bases de baixa qualitat incloses B) Les mateixes seqüències mostrades a A, un cop processades. C) Seqüències de 76 pb amb cartutxos de versió 3, sense processar.	86
Figura 3-4. Representació esquemàtica de dos fragments, amb llargades d'insert diferents, immobilitzats a la cel·la de flux del seqüenciador. Els adaptadors es representen en groc. En vermell i blau es representen les seqüències universals per la hibridació dels adaptadors. La polimerasa es representa de color verd.....	87
Figura 3-5. Exemples de formats A) fastq i B) SAM. C) Representació esquemàtica de la seqüència de B i la seva parella.	92
Figura 3-6. Captures de pantalla del visor genòmic IGV (<i>Broad Institute</i>) en les que es mostra: A) Una SNV heterozigota. B) Una SNV homozigota. C) Una inserció heterozigota. D) Una deleció heterozigota.....	95
Figura 4-1. A) Cobertures mitjanes assolides per les mostres de la cohort. Els números informen de la quantitat total de mostres seqüenciades per panell. B) <i>Call rates</i> a 30x de les mostres seqüenciades. Cada panell és representat per un color diferent.	107
Figura 4-2. Panell de 55 gens. A) Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertura mitjana. A les regions amb continguts GC elevats es distribueixen una major quantitat de sondes. B) Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). Es pot apreciar com les regions amb continguts GC elevats assoleixen molta menys cobertura que la resta. C) Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.	109
Figura 4-3. Panell de 78 i 118 gens, respectivament. A-B) Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertura mitjana. S'aprecia l'increment de la densitat de sondes en les regions amb continguts GC elevats. C-D) Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb els panells de 78 i 118 gens, respectivament, en funció del contingut GC (groc). Les cobertures assolides per les regions d'elevat contingut GC són acceptables, però les regions amb un GC reduït presenten menys cobertura que la resta. E-F) Mitjanes d'exons perduts a 30x per les mostres seqüenciades amb el panell.	110
Figura 4-4. Panell de 85 gens. A) Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertura mitjana. En aquest disseny s'han redistribuït sondes cap a les regions amb GCs extrems ($\leq 35\%$ i $\geq 60\%$). B) Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). Pot apreciar-se una adquisició de cobertura homogènia per totes les regions seqüenciades, independentment del GC. C) Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.	111
Figura 4-5. Panell de 147 gens. A) Densitat de sondes per regió en funció del contingut GC (groc). Es segueix la mateixa estratègia que pel panell de 85 gens, però amb un número considerablement superior de regions per seqüenciar. B) Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). S'intueix una lleugera davallada de cobertura a les regions amb GCs centrals, símptoma de que el disseny es troba al límit de la seva capacitat. C) Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.	112
Figura 4-6. Mostres seqüenciades amb el panell de dislipèmia. A) Número (en milions) de seqüències totals (blau) i solapants amb les regions d'interès (vermell) per mostra. L'enriquiment es mostra en verd.	

B) Cobertura acumulada en les regions codificants del gen <i>LDLR</i> . La regió 3' UTR apareix indicada amb una fletxa.	114
Figura 4-7. Exemple d'arxiu de configuració necessari per l'execució de l'algoritme de detecció de CNVs.	121
Figura 4-8. Diagrama de flux de l'algoritme. En blau es representen els arxius de partida, en gris els arxius intermedis i en verd els arxius resultants de l'anàlisi. Els processos es representen en vermell.	122
Figura 4-9. Correlacions de cobertures entre mostres analitzades. A) <i>Run</i> de 6 mostres amb bons coeficients de correlació. B) <i>Run</i> de 6 mostres; la primera mostra del grup té una duplicació de 18 exons que fa disminuir el seu coeficient de correlació. C) <i>Run</i> de 6 mostres amb diferències d'enriquiment importants. No s'aconsella la seva anàlisi.	124
Figura 4-10. Gràfics de cobertures en funció del contingut GC. La línia vermella reflexa la tendència de la cobertura. A) Cobertures medianes per regió d'un run realitzat amb el panell de 85 gens. B) Cobertures medianes per regió procedents de mostres d'un projecte no relacionat. El disseny de sondes per aquest panell no va ser optimitzat.	126
Figura 4-11. Desviació de les ràtios per regió al llarg de la cobertura (log2) en funció del número de mostres incloses a l'anàlisi.	127
Figura 4-12. Diagrames de caixa en els que es representa la desviació absoluta de les ràtios (log2) per regió a les diferents etapes de normalització de cobertures. Cada color correspon a un panell de gens diferent.	128
Figura 4-13. A) Distribució normal de les ràtios generades en una anàlisi de 6 mostres. B) Gràfics Q-Q de normalitat per les 6 mostres incloses a l'anàlisi.	129
Figura 4-14. Mosaic detectat al gen <i>SCN5A</i> , abastant els exons 15 - 28. A) Gràfic generat per l'algoritme en el que s'aprecia com les ràtios no arriben al límit de detecció de duplicació heterozigota. B) Resultats de MLPA per la validació del mosaic.	130
Figura 4-15. Exemples de CNVs trobades al gen <i>LDLR</i> en pacients d'hipercolesterolèmia familiar. A) <i>Run</i> en el que s'identifica les delecions dels exons 3-18 + 3'UTR (marró, mostra VAL_55), de l'exó 5 (groc, mostra VAL_61) i dels exons 11 i 12 (lila, mostra VAL_71). B) <i>Run</i> en el que s'identifica les delecions de promotor + 5'UTR + exons 1-2 (mostassa, mostra VAL_32), la delecio dels exons 13-14 (groc, mostra VAL_38) i la dels exons 16-18 + 3'UTR (verd fosc, mostra VAL_40).	134
Figura 4-16. Comparació de l'exactitud i de la sensibilitat de l'algoritme (vermell) amb el <i>software</i> CNVKIT v.0.8.6 (verd) i el <i>software</i> CONTRA v.2.0.8 (blau).	134
Figura 4-17. Caracterització de 3 CNVs. Representació dels punts de trencament i la caracterització precisa per seqüenciació Sanger. A) La delecio de l'exó 27 de <i>MYBPC3</i> del pacient P2. B) La delecio de l'exó 4 al 12 a <i>MYBPC3</i> del pacient P1. C) La delecio de la regió codificant de <i>PLN</i> , del pacient P4. Adaptada de Mademont-Soler et al. 2017.	138
Figura 4-18. Duplicació dels exons 5 – 10 a <i>LMNA</i> (A) i pedigrí (B) del pacient P14. El pacient s'assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol "+" indica que l'individu és portador de la variant, i el "-", que no ho és portador.	145
Figura 4-19. (A) Duplicació dels exons 8 – 10 a <i>PKP2</i> , detectada als pacients P19, P20 i P21. L'asterisc marca l'exó 6 de <i>PKP2</i> , una regió amb marcada variabilitat en l'homogeneïtat de la seqüenciació, tal i com queda resumit a la Taula 4-2. (B) Delecio dels exons 9 – 24 a <i>DSP</i> , detectada al pacient P24.	147
Figura 4-20. Duplicació de 176,8 Kb (exons 45 – 275) a <i>TTN</i> , detectada en el pacient P26.	149

Figura 4-21. **(A)** Deleció dels exons 7 i 8 a *KCNQ1* detectada als pacients P30, P31 i P32. **(B)** Pedigrí del pacient P32. El pacient s'assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol "+" indica que l'individu és portador de la variant, i el "-", que no ho és portador.151

Figura 4-22. Deleció dels exons 12 i 13 a *CTNNA3* **(A)** i del gen *GAA* **(B)** detectada al pacient P41. **(C)** Pedigrí del pacient P41. El pacient s'assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol "+" indica que l'individu és portador de la variant.155

Índex de taules

Taula 1-1. Prevalença, percentatge de casos resolts i gens associats a les principals malalties cardíaques arritmogèniques.	54
Taula 3-1. Llista dels panells de gens utilitzats per la seqüenciació de mostres.	81
Taula 3-2. Camps d'informació presents a la línia identificadora d'una seqüència fastq.	88
Taula 3-3. Camps d'informació presents en un alineament en format SAM.	92
Taula 3-4. Taula de classificació de variants en funció de l'evidència de patogenicitat acumulada.	97
Taula 4-1. Mètriques descriptives relatives a la seqüenciació de la cohort d'estudi.	105
Taula 4-2. Regions amb homologia de seqüència.	116
Taula 4-3. CNVs detectades en la validació de l'algoritme amb mostres de pacients d'hipercolesterolèmia familiar.	133
Taula 4-4. Resum de les CNVs identificades classificades per malaltia.	136
Taula 4-5. Resum de les dades clíniques i els resultats genètics dels pacients portadors de CNVs. Totes les variants genètiques van ser detectades en heterozigosi o hemizigosi.	139
Taula A-2. Resum dels gens (i les isoformes) inclosos en els diferents panells de gens utilitzats.	207

A Rosario, por su cariño (y sus croquetas).

que fuera epitafio del hombre más sabio
un "yo sólo pasé por aquí"

Santi Balmes

Los males pasajeros –Love of Lesbian–

Agraïments

Se'n recorda de la Comarca, senyor Frodo? Aviat arribarà la primavera. Els horts seran tots en flor i a l'avellaneda els ocells tindran a punt els seus nius. Començarà la sembra estival de l'ordi als bancals. La degustació de les primeres maduixes amb nata. El gust de les maduixes... El recorda?

Bé, doncs no.

No recordo el gust de les maduixes. Acabar aquesta tesi m'ha deixat sense energies. És una sensació que m'agradaria recordar quan agafi aquest document polsegós d'allà on el tingui arxivat (cosa que probablement no passi mai, però ves). Potser utilitzar aquest paral·lelisme és fer gala d'un estil massa dramàtic, però que se'm permeti la llicència. No tinc pensat escriure una altra tesi mai més a la vida. Hi ha hagut llums i ombres, però en general m'ho he passat de conya. En aquesta etapa he conegut gent clau per ser la persona que sóc ara mateix, entre elles l'Eva, la meva parella. Però no avancem esdeveniments i anem de cap a les formalitats. Seguiu-me!

Tothom comença els agraïments de la tesi donant les gràcies als seus directors de projecte. Sembla lògic, i aquí no seré una excepció. Preferiria, però, deixar constància del que he après de cada un d'ells, que al cap i a la fi és el que compta. En Ramon m'ha ensenyat (perquè predica amb l'exemple) que no importa si avui vas de congrés a la Xina i en quatre dies has de mantenir una reunió important a Suècia, sempre que tornis a casa i puguis tenir cura de l'hort. En Carles ha estat un exemple de versatilitat i una demostració de com anar entomant les coses a mesura que van venint. Amb calma, però de manera implacable. I la Catarina m'ha ensenyat el valor de l'honestedat en una conversa (o entrevista de treball o... el que sigui): fa que escriguis tesis. Gràcies als tres.

La Irene ha estat una persona clau en la meva formació durant aquests anys. La seva coherència i honestedat, tant científica com personal, han estat una guia. I la seva paciència i el seu temps els responsables d'ensenyar-me a fer les coses com s'han de fer. Moltes gràcies Irene, de tot cor.

Els següents de la llista (veureu que l'ordre és una mica 'així', però l'impuls mana i jo ja estic cansat) són en Pau, en Joan i en Xavi. Hem format un tàndem formidable, que malauradament serà impossible repetir. A no ser, és clar, que ens divorciem tots als 40 i tornem a compartir pis. En fi... el temps dirà. Però per si de cas això no passés, m'ha encantat compartir amb vosaltres el temps viscut a Girona. Sou genials. Amb en Prujà (també ets genial, *tontu*) em temo que no ens hem vist tant com caldria. És un punt que penso compensar a partir de ja mateix (més o menys).

Tornant als de Gencardio (i reafirmant la falta absoluta d'estructura d'aquest apartat) és el torn d'en Bernat i en Javi. Si es pogués resoldre el món des del bar, amb els amics, nosaltres ho hauríem fet un cop cada quinze dies. I això és dir poc però molt al mateix temps. La confiança que porta implícita aquesta frase és digna de menció, encara que sigui pel meu propi record. Suposo que el que vull dir és que enyoro a en Javi. A en Bernat no l'enyoro perquè seu vuit hores diàries al meu costat. Però mola

treballar amb ell. I amb la Mel·lina, és clar, que ja comença a donar senyals de contagi de la nostra bogeria. Somrius més i millor que abans, i jo no podria resumir de cap millor manera com de content em fa això que posant-ho aquí.

Amb la resta del grup també ens ho hem passat bé. Hem celebrat comiats de solter, embarassos i jubilacions. Hem regalat més arbres dels que serien necessaris i, en general, hem consumit més cafè que tots els regiments de la Segona Guerra Mundial junts. L'Alexandra, en Ferran, la Mònica (i des de fa poc la Marta i la Laura) són les mans del grup. I els peus. I un ronyó. L'Anna I, la Sara, l'Eli S, l'Eli C, la Fabiana, en Marcel, l'Oscar i en Guillermo sempre disposats a compartir els seus coneixements i a donar un cop de mà (gràcies!). Un record pel *Komando Gin-tonic*, de gloriosa actuació (encara que efímera) format per l'Anna T, la Cris, en Javi (*again*) i un servidor. Suposo que el nom ho diu tot. L'Anna F (la *Consiliere*), la Mireia, l'Olallo, l'Helena i les recents adquisicions (valuoses per la sang fresca. Mmmh...): en David, l'Eric, el *torbellino pelirrojo* de la Rebecca i la Marta. Tots plegats fem un bon grup. O no? I si em deixo algú doncs, bé. Que aixequi el dit.

Tal i com tenia pensat estructurar aquest apartat, ara vindria la part en la que parlo de les coses dolentes, en com de frustrant és que la teva feina depengui de tercers i en la quantitat formidable de temps, esforç i il·lusió malgastats durant el camí. Però al final he decidit no deixar ni rastre de tot això, així que no. No us explicaré la quantitat formidable de temps, esforç i il·lusió malgastada. No. De veritat que no.

No podria deixar-me a la Sandra, una companya de camí inigualable. Estic molt orgullós de tu, Sandra, i molt content de que estiguem acabant *això* al mateix temps. Ja tinc ganes que trobem feina els dos al McDonald's®. Al teu costat segur que serà molt divertit, l'única cosa que importa.

Per anar acabant, que ja seria hora, vull agrair als meus pares i a la meva germana la comprensió rebuda i l'esforç realitzat en educar-me, mantenir-me i aguantar-me. Però sobretot agrair que m'estimin com ho fan.

I a l'Eva. Aquesta tesi és teva. No en faràs res, però és teva. Per totes les hores que li he dedicat a ella i no a tu. Per les males estones que ens ha fet passar i per les vegades que t'has enfadat amb mi per culpa d'ella (en algunes tenies raó i tot). És en les situacions límit quan es coneix a les persones, i aquestes planes ens hi han portat (al límit). I el que hi he trobat és preciós. I em sento molt afortunat de poder caminar al teu costat. T'estimo.

Sense vosaltres no hauria estat possible. Gràcies.

Més de 4 milions de persones moren arreu del món anualment a causa de la Mort Sobtada Cardíaca –MSC–. Les malalties cardíques arritmogèniques suposen la principal causa de MSC entre la població menor de 35 anys. Tot i les millores en el diagnòstic genètic, a causa de la implementació de les tècniques de seqüenciació d'alt rendiment en la pràctica clínica, el percentatge de casos que resulten sense causa després de l'anàlisi genètica continua sent elevat. Diversos estudis han associat les *Copy Number Variants* –CNVs– com les causants de malalties cardíques associades a MSC. Tot i així, encara no s'ha realitzat mai el cribratge exhaustiu d'un grup important de gens en una gran cohort de pacients diagnosticats amb aquestes malalties. Aquesta falta d'informació és la que ha motivat la realització d'aquesta tesi.

El primer objectiu d'aquesta tesi ha estat desenvolupar un algoritme de detecció de CNVs per l'anàlisi de dades provinents de la seqüenciació d'alt rendiment de les regions clínicament rellevants associades a la MSC i a les malalties relacionades (miocardiopaties i canalopaties). Mitjançant l'optimització dels dissenys de sondes de captura s'han pogut seqüenciar unes mostres que exhibeixen tant una alta qualitat com una elevada homogeneïtat de cobertura al llarg de totes les regions seqüenciades. Aquestes mostres han permès la posada a punt de l'algoritme de detecció de CNVs, dissenyat a mida per aquest tipus de mostres, que ha demostrat una alta sensibilitat, especificitat i precisió.

Els dissenys de sondes i l'algoritme són els dos components del mètode de detecció de CNVs utilitzat per dur a terme el segon objectiu: la realització d'un cribratge exhaustiu que englobi la detecció de CNVs i variants puntuals en una gran cohort de 2073 pacients de MSC, pacients diagnosticats amb malalties associades a la MSC i casos de Mort Sobtada Inexplicada –MSI– (cohort per la qual mai s'ha realitzat un estudi d'aquest tipus). El cribratge ha inclòs els principals gens relacionats a les malalties associades a la MSC, així com una àmplia varietat de gens minoritaris i candidats.

Les recomanacions internacionals actuals es centren en la interpretació de grans reorganitzacions cromosòmiques que sovint involucren diversos gens adjacents. No existeixen recomanacions detallades per la interpretació de CNVs intragèniques, tot i que sota el nostre punt de vista requereixen d'una consideració especial. El tercer objectiu que s'ha plantejat en aquesta tesi és el de fer la translació a la clínica dels resultats obtinguts en base a uns criteris de classificació de CNVs propis, amb la intenció de facilitar una avaluació clínica i/o forense més concisa, que permeti millorar l'assessorament genètic i les mesures preventives per als pacients i als seus familiars. Els resultats obtinguts revelen que les CNVs expliquen una petita porció (encara que no negligible) dels casos de MSI, dels pacients de MSC i dels diagnosticats amb malalties associades a la MSC. Així doncs, es proporciona evidència a favor de la inclusió del cribratge de CNVs en les anàlisis genètiques rutinàries per aquests pacients, per tal d'adoptar estratègies preventives i terapèutiques per ells i els seus familiars.

Summary

More than 4 million people die every year around the world because of Sudden Cardiac Death –SCD–. Arrhythmogenic cardiac diseases are the main cause of SCD among the population under 35 years of age. Despite the improvements in genetic diagnosis, due to the implementation of high-throughput sequencing techniques in clinical practice, the percentage of cases that remain unexplained after genetic analysis is still high. Several studies have identified Copy Number Variants –CNVs– as causative of cardiac diseases associated with SCD, but exhaustive analysis of multiple genes in large cohorts of patients has never been performed for most SCD-related diseases. This lack of information motivated the realization of this thesis.

The first objective of this thesis was to develop a CNV detection algorithm for the analysis of high-throughput sequencing data of the clinically relevant genomic regions associated with SCD and related diseases (myocardiopathies and canalopathies). Through the optimization of the capture probe designs we have been able to sequence samples that exhibit high quality and coverage homogeneity throughout all sequenced regions. These samples allowed the development of the CNV detection algorithm, specifically designed for this type of samples, which has shown a high sensitivity, specificity and precision.

The probe designs and the algorithm are the two components of the CNV detection method used to carry out the second objective: to perform an exhaustive screening that includes the detection of CNVs and specific variants in a large cohort of 2073 SCD patients, others diagnosed with SCD-related diseases and cases of Sudden Unexplained Death –SUD– (a cohort for which such a study has never been performed). The screening included the main genes associated with SCD-related diseases, as well as a wide variety of minor genes and candidates.

Current CNV interpretation guidelines are focused on the interpretation of large genomic rearrangements, generally involving multiple contiguous genes. Guidelines for the interpretation of intragenic CNVs do not exist, although they need special considerations. The third objective considered in this thesis is to do the translation into clinics of the obtained results according on our own CNV classification criteria. Our intention is to facilitate a more concise clinical or forensic evaluation, which allows improving the genetic counselling and the preventive measures for patients and their relatives.

The obtained results reveal that CNVs explain a small (although non-neglectable) portion of cases of SUD, SCD patients and those diagnosed with SCD-related diseases. Therefore, evidence is provided in favor of the inclusion of the CNV screening in routine genetic analysis for these patients, in order to adopt preventative and therapeutic strategies for them and their relatives.

Más de 4 millones de personas mueren cada año en todo el mundo a causa de la Muerte Súbita Cardíaca –MSC–. Las enfermedades cardíacas arritmogénicas suponen la principal causa de MSC entre la población menor de 35 años. A pesar de las mejoras en el diagnóstico genético, debidas a la implementación de las tecnologías de secuenciación de alto rendimiento en la práctica clínica, el porcentaje de casos que resultan sin causa después del análisis genético continua siendo elevado. Varios estudios han asociado las *Copy Number Variants* –CNVs– como las causantes de enfermedades cardíacas asociadas a la MSC. A pesar de ello, aún no se ha realizado nunca un cribado exhaustivo de un grupo importante de genes en una gran cohorte de pacientes diagnosticados con estas enfermedades. Esta falta de información es la que ha motivado la realización de esta tesis.

El primer objetivo de esta tesis ha sido el de desarrollar un algoritmo de detección de CNVs para el análisis de los datos procedentes de la secuenciación de alto rendimiento de las regiones clínicamente relevantes asociadas a la MSC y a las enfermedades relacionadas (miocardiopatías y canalopatías). Mediante la optimización de los diseños de sondas de captura se han podido secuenciar unas muestras que exhiben tanto una alta calidad como una elevada homogeneidad de coberturas a lo largo de todas las regiones secuenciadas. Estas muestras han permitido la puesta a punto de un algoritmo de detección de CNVs diseñado a medida para este tipo de muestras, que ha demostrado una alta sensibilidad, especificidad y precisión.

Los diseños de sondas y el algoritmo son los dos componentes del método de detección de CNVs utilizado para llevar a cabo el segundo objetivo: la realización de un cribado exhaustivo que englobe la detección de CNVs y de variantes puntuales en una gran cohorte de 2073 pacientes de MSC, pacientes diagnosticados con enfermedades asociadas a la MSC y casos de Muerte Súbita Inexplicada –MSI– (cohorte para la cual nunca se ha realizado un estudio de este tipo). El cribado ha incluido los principales genes relacionados con las enfermedades asociadas a la MSC, así como una amplia variedad de genes minoritarios y candidatos.

Las recomendaciones internacionales actuales se centran en la interpretación de grandes reorganizaciones cromosómicas, que a menudo involucran diferentes genes adyacentes. No existen recomendaciones detalladas para la interpretación de las CNVs intragénicas, aunque bajo nuestro punto de vista estas requieren de una consideración especial. El tercer objetivo que se ha planteado en esta tesis es el de hacer la translación a la clínica de los resultados obtenidos en base a unos criterios de clasificación de CNVs propios, con la intención de facilitar una evaluación clínica y/o forense más concisa, que permita mejorar el asesoramiento genético y las medidas preventivas para los pacientes y sus familiares. Los resultados obtenidos revelan que las CNVs explican una pequeña porción (aunque no despreciable) de los casos de MSI, de pacientes de MSC y de aquellos diagnosticados con enfermedades asociadas a la MSC. Se proporciona evidencia a favor de la inclusión del cribado de CNVs en los análisis genéticos rutinarios para estos pacientes, con tal de adoptar tales estrategias preventivas y terapéuticas para ellos y sus familiares.

Índex

Llista de publicacions.....	i
Llista d'abreviatures.....	iii
Índex de figures.....	vii
Índex de taules.....	x
Agraïments.....	xv
Resum.....	xvii
Summary.....	xix
Resumen.....	xxi
Índex.....	xxiii
1 – INTRODUCCIÓ.....	1
1.1 – L'evolució de la tecnologia de seqüenciació de DNA.....	3
1.1.1 – La tecnologia de seqüenciació de primera generació.....	3
1.1.2 – La seqüenciació d'alt rendiment (la segona generació)	6
I – La seqüenciació 454.....	7
II – El mètode de seqüenciació d'Illumina.....	10
III – El sistema SOLiD™.....	12
IV – La seqüenciació paired-end.....	14
V – La revolució dels seqüenciadors de sobretaula.....	14
1.1.3 – La <i>Next-Generation Sequencing</i>	15
I – PacBio RS System.....	15
II – MinION Nanopore Sequencing.....	17
1.2 – El codi genètic humà.....	19
1.2.1 – Les variants genètiques i l'evolució.....	22
1.2.2 – Mecanismes moleculars de generació de variants estructurals.....	26
I – Mecanismes mediat per recombinació homòloga.....	26
a. Reparació de trencaments de doble cadena.....	27
b. Replicació induïda per trencament.....	29
c. Recombinació homòloga no al·lèlica.....	31
II – Mecanismes de reparació no homòloga.....	33
a. Unió d'extrems no homòlegs.....	33
b. Cicle de trencament – pont – fusió.....	35
c. Lliscament durant la replicació.....	36
d. FoSTeS.....	37
e. Replicació induïda per trencament intervinguda per microhomologia.....	38
III – Retrotransposició.....	38
1.2.3 – Mecanismes moleculars de transmissió de fenotips clínics mitjançant CNVs....	40
1.3 – La biologia computacional i la bioinformàtica.....	42
1.3.1 – La bioinformàtica en l'anàlisi de dades de seqüenciació d'alt rendiment.....	44
1.3.2 – Detecció de variants estructurals.....	46
I – Mètodes tradicionals.....	46
II – Mètodes bioinformàtics.....	48

a. Detecció basada en la comparació de cobertures.....	50
b. Detecció basada en les seqüències aparellades.....	51
c. Detecció basada en els alineaments parcials.....	51
1.4 - La Mort Sobtada Cardíaca.....	52
1.4.1 – Definició i epidemiologia.....	52
1.4.2 – Les malalties cardíaques arritmogèniques.....	53
I – La diagnosi genètica de les malalties cardiovasculars arritmogèniques.....	55
II – Les miocardiopaties.....	57
a. La Miocardiopatia Hipertròfica (MCH)	57
b. La Miocardiopatia Dilatada (MCD)	58
c. La Miocardiopatia Arritmogènica (MCA)	60
d. La No Compactació del Ventricle Esquerre (NCVE)	61
e. La Miocardiopatia Restrictiva (MCR)	62
III – Les canalopaties.....	63
a. La Síndrome de Brugada (SBr)	63
b. La Síndrome del QT Llarg (SQTL)	64
c. La Síndrome del QT Curt (SQTC)	65
d. La Taquicàrdia Ventricular Polimòrfica Catecolaminèrgica (TVPC)	66
e. La Fibril·lació Auricular (FA)	67
IV – La Síndrome de Marfan.....	68
2 – JUSTIFICACIÓ DE LA RECERCA, HIPÒTESIS I OBJECTIUS.....	71
3 – MATERIALS I MÈTODES.....	75
3.1 – Declaració de contribució.....	77
3.2 – Consentiment informat.....	77
3.3 – Cohort d'estudi.....	78
3.3.1 – Pacients diagnosticats amb miocardiopaties.....	79
3.3.2 – Pacients diagnosticats amb canalopaties.....	79
3.3.3 – Casos de MSI.....	79
3.3.4 – Pacients diagnosticats amb hipercolesterolèmia familiar.....	80
3.4 – Disseny de panells de captura.....	80
3.4.1 – Selecció de regions d'interès.....	81
3.4.2 – Criteris pel disseny de sondes de captura.....	81
3.5 – Processament de mostres.....	83
3.5.1 – Preparació de llibreries genòmiques.....	83
3.5.2 – <i>Pooling</i> i càrrega al seqüenciador MiSeq.....	83
3.6 – Anàlisi bioinformàtica dels resultats.....	84
3.6.1 – Controls de qualitat.....	84
I – Control de qualitat de la seqüenciació.....	84
II – Control de qualitat de les seqüències.....	85
3.6.2 – Els arxius fastq.....	88
3.6.3 – Alineament de seqüències.....	89
3.6.4 – El format SAM / BAM.....	91
3.6.5 – Eliminació de duplicats òptics i de PCR.....	93
3.6.6 – Detecció de variants puntuals i <i>indels</i>	94
3.6.7 – Anotació de variants puntuals i <i>indels</i>	96
3.6.8 – Detecció i anotació de variants estructurals.....	97

3.7 – Classificació de les variants genètiques.....	97
3.8 – Confirmació de variants.....	100
3.8.1 – Seqüenciació Sanger.....	100
3.8.2 – MLPA.....	100
3.8.3 – PCR quantitativa.....	101
4 – RESULTATS I DISCUSSIÓ.....	103
4.1 – Caracterització i discussió dels resultats de la seqüenciació d'alt rendiment.....	105
4.1.1 – Evolució dels panells de seqüenciació.....	107
4.1.2 – Regions multimap.....	115
4.2 – Desenvolupament d'un algoritme informàtic per la detecció de CNVs.....	119
4.2.1 – Disseny i implementació de l'algoritme.....	120
I – Etapa de preprocessament.....	123
II – Normalització de cobertures.....	124
III – Càlcul de ràtios.....	126
IV – Estimació del número de còpia.....	128
V – Anotació de les senyals.....	130
VI – Avaluació de les senyals mitjançant un <i>score</i> de fiabilitat.....	131
VII –Exportació dels resultats.....	131
4.2.2 – Validació i avaluació comparativa de l'algoritme.....	132
4.3 – Resultats del cribratge genètic.....	135
4.3.1 – CNVs identificades a la cohort de miocardiopaties.....	137
I – Pacients diagnosticats amb MCH.....	137
II – Pacients diagnosticats amb MCD.....	143
III – Pacients diagnosticats amb MCA.....	146
IV – Pacients diagnosticats amb NCVE.....	148
4.3.2 – CNVs identificades a la cohort de canalopaties.....	150
I – Pacients diagnosticats amb SBr.....	150
II – Pacients diagnosticats amb SQT.....	150
III – Pacients diagnosticats amb FA.....	152
4.3.3 – CNVs identificades a la cohort de MSI.....	153
4.3.4 – CNVs identificades als pacients de SM i DATA.....	156
5 – DISCUSSIÓ GENERAL.....	157
6 – CONCLUSIONS.....	169
7 – BIBLIOGRAFIA.....	173
ANNEX I	199
ANNEX II.....	205
ANNEX III.....	209
ANNEX IV.....	235
ANNEX V.....	249

I. Introducció

1.1 – L'evolució de la tecnologia de seqüenciació de DNA

Des del descobriment de l'estructura del DNA i de les implicacions genètiques i de transmissió d'informació biològica que d'ella se'n deriven (1,2), la seqüenciació del material genètic ha estat sempre en el punt de mira de la comunitat científica. Gràcies al descobriment de la DNA polimerasa 1 i del seu protocol d'aïllament i purificació (3,4) va ser possible, durant la dècada dels setanta, el floreixement de les tecnologies de seqüenciació de DNA.

1.1.1 – La tecnologia de seqüenciació de primera generació

Els pioners en la seqüenciació de DNA van ser Sanger i Coulson, amb la publicació de l'anomenat "*plus and minus system*" (5), mètode amb el que l'any 1977 es va publicar el primer genoma de la història, el del bacteriòfag Φ X174 (6). Tot i així, el mètode més popular del moment va ser el publicat per Maxam i Gilbert al 1977, ja que resultava més pràctic i senzill d'executar, generava menys artefactes de lectura i podia seqüenciar DNA de cadena doble (7). El mètode es basava en la fragmentació del DNA purificat i en el marcatge radioactiu dels fragments obtinguts. Aquests es feien córrer a través d'un gel desnaturalitzador d'acrilamida, que posteriorment s'exposava a raigs X per revelar les bandes enfosquides relatives als fragments marcats de DNA. En funció de la presència o l'absència de les bandes es deduïa la seqüència d'interès (Figura 1-1). La millora del mètode publicat per Sanger i Coulson, amb l'addició d'inhibidors específics de la DNA polimerasa 1 (els anomenats "terminadors de cadena"), va convertir-lo en la tècnica *gold standard* de seqüenciació de DNA, passant a conèixer-se com el mètode Sanger (o mètode del '*dideoxi*'). El mètode millorat resultava encara més simple que el mètode de Maxam i Gilbert –que consistia en un protocol de 30 passos a realitzar per un químic ben entrenat–, era més escalable i evitava l'ús de radioisòtops, perillosos per la salut (8).

Al 1986 es va introduir el marcatge d'encebadors per fluoròfors, associant per primera vegada un color a cada una de les 4 bases del DNA. El sistema de codificació per colors va eliminar la necessitat de córrer gels d'electroforesi separats, un per cada una de les 4 bases. També va fer possible la captació informàtica de les seqüències mitjançant un sistema òptic, característica primordial per poder automatitzar el procés de seqüenciació (9,10). Amb aquests avenços, i gràcies al treball previ del grup de Wada et al., que havia dissenyat un robot capaç d'automatitzar el procés de fragmentació del DNA (11), el 1986 va comercialitzar-se el primer seqüenciador automàtic de DNA, l'ABI PRISM® 370A (Applied Biosystems; Foster City, USA), capaç de seqüenciar 300 parells de bases (pb) cada 12 hores.

Poc a poc, la informació genètica començava a sortir dels laboratoris, per trobar aplicacions pràctiques. La comunitat científica i la indústria biotecnològica va prendre consciència de que amb la millora dels mapes físics i genètics, i amb la identificació de cada cop més gens humans, el potencial dels resultats de la recerca en la tecnologia de seqüenciació podien ser enormes. La seqüenciació del DNA era estratègicament essencial i, de manera molt especial, la del DNA humà. Una visió global del

genoma humà acceleraria la recerca biomèdica, permetent als investigadors atacar els problemes de manera exhaustiva i objectiva. Era evident, però, que un objectiu de tal magnitud requeriria un esforç logístic i de recursos sense precedents en el camp de la recerca biomèdica. Per aquest motiu, a finals dels anys vuitanta, la idea de seqüenciar el genoma humà esperonada per l'aparició de les millores significatives abans esmentades en els mètodes de seqüenciació, va començar a cristal·litzar en la comunitat científica.

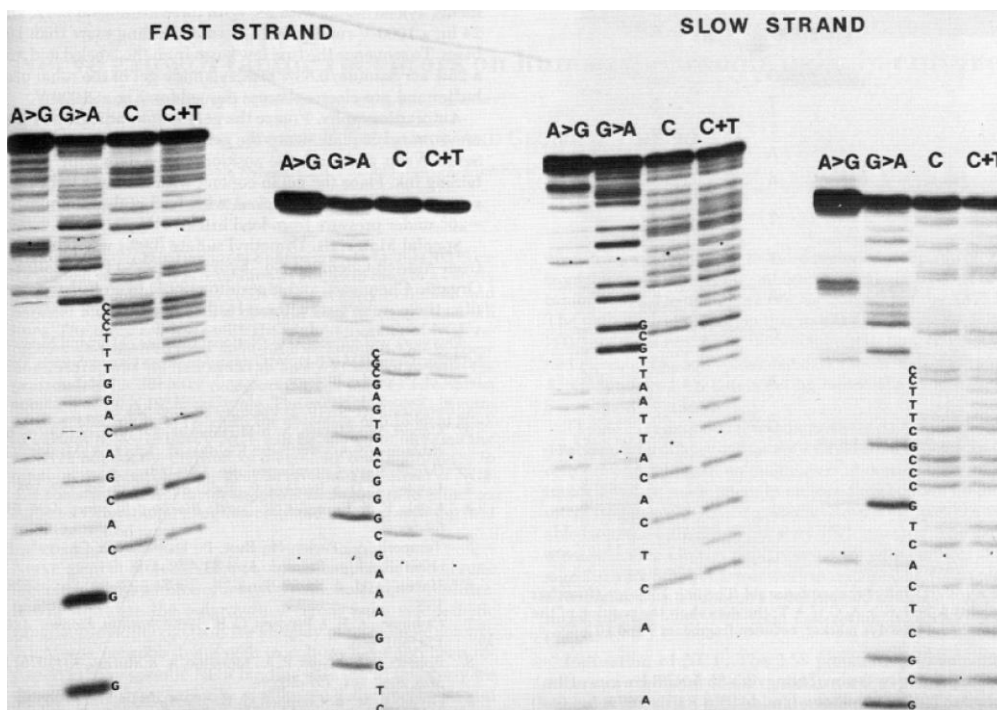


Figura I-1 | Autoradiografia d'un gel de seqüenciació. S'hi pot apreciar la seqüència de DNA, de 64 pb provinents de dues cadenes complementàries (7).

Al 1988 va aparèixer el primer aparell automàtic per realitzar la reacció en cadena de la polimerasa (de l'anglès, *Polymerase Chain Reaction* –PCR–), combinant així l'amplificació exponencial del DNA de la que era capaç la tècnica ideada al 1983 pel Dr. Kary Mullis (12) amb la capacitat de producció d'un sistema automàtic (13). La reducció del temps i dels costos va obrir el camí cap a la seqüenciació totalment automatitzada i al Projecte Genoma Humà. No va ser fins l'any 1990 que el Departament d'Energia i d'Institut Nacional de Salut dels Estats Units van publicar el pla d'acció estratègic del Projecte Genoma Humà, de quinze anys de durada. L'objectiu era desenvolupar la tecnologia necessària per la seqüenciació, l'anàlisi i l'ensamblatge de quantitats ingents de DNA. Assolits aquests objectius, l'obtenció de la seqüència completa del genoma humà podria ser una realitat. Per tirar endavant el projecte es va crear el Consorci per la Seqüenciació del Genoma Humà (de l'anglès, *Human Genome Sequencing Consortium* –HGSC–), una entitat finançada per subvencions públiques internacionals.

Al llarg dels anys noranta, els terminis previstos pel Projecte Genoma Humà van accelerar

l'optimització del procés de seqüenciació, que encara tenia un coll d'ampolla important. Era necessari augmentar les pistes de migració de molècules de DNA, així com la velocitat d'aquestes al llarg del suport per on migraven. Al 1990, Luckey et al. van proposar un sistema d'electroforesi capil·lar del DNA, en substitució del gel laminar tradicional. D'aquesta manera es podrien assolir camps elèctrics d'elevat voltatge que accelerarien el procés de migració de les molècules a través de la matriu, i com el suport afavoriria la dissipació del calor, s'evitaria la destrucció del gel o la degradació del DNA per efecte de l'escalfament (14). Al 1998 es van comercialitzar els dos primers seqüenciadors de 96 capil·lars, l'ABI 3700, d'*Applied Biosystems*, i el MegaBace d'*Amersham Pharmacia Biotech* (Little Chalfont, UK) (15).

Gràcies a l'esforç col·laboratiu internacional entre més de 600 científics, el 1996 es va publicar el primer genoma eucariota, el del llevat *Saccharomyces cerevisiae*, de 12'5 megabases (Mb) de DNA (16). Per primera vegada s'havia establert una xarxa d'informació per internet en forma de base de dades. Al finalitzar el dia, els científics del projecte enviaven les dades a un repositori web, on eren processades. Al finalitzar el projecte, tota la informació es va vessar a la base de dades GenBank, fent-les d'accés lliure i gratuït. L'èxit del projecte va fer decidir-se als científics de l'HGSC a publicar les seqüències que anaven obtenint en una base de dades pública, cada vint-i-quatre hores, per una millor optimització del ritme de treball.

Finalment, a l'any 2001 es va publicar el primer esborrany i anàlisi de la seqüència del genoma humà. La competició que es va establir entre l'HGSC i la corporació privada *Celera Genomics* (California, USA), dirigida pel Dr. Craig Venter, va catalitzar la innovació tecnològica, els esforços en el desenvolupament d'algoritmes bioinformàtics fonamentals pel processament de dades, l'abaratiment de la tecnologia de seqüenciació (amb un rendiment cada cop més elevat) i les iniciatives de col·laboració internacional, fent que el projecte finalitzés abans del que s'havia previst i amb un estalvi en el pressupost inicial (17). Els dos grups van publicar una versió pròpia del genoma en paral·lel, amb la respectiva anàlisi (18,19). La quantitat d'informació publicada pels dos grups va ser comparable, però la qualitat de l'ensamblatge de l'HGSC va resultar ser de millor qualitat. El sistema de predicció de gens que havien utilitzat es basava en homologia GENSCAN (20) i en mètodes de predicció basats en models ocults de Markov, generant resultats més realistes que la competència. Aquestes prediccions van resumir-se i publicar-se en la primera versió de la base de dades ENSEMBL. El mètode de predicció de gens utilitzat per Celera, l'anomenat Otto (19), era més complex i donava més èmfasi a les comparacions entre espècies i en l'homologia amb un set de dades curades. Van obtenir una predicció sobreestimada del número de gens (uns 40000, en comparació amb els 30000 d'ENSEMBL) i, en general, s'apreciava un solapament molt baix dels resultats entre els dos grups (21). L'any 2003 es va publicar la seqüència completa del genoma eucromàtic haploide humà (22).

La millora del mètode de seqüenciació Sanger va continuar al llarg de la primera dècada del segle XXI, permetent la seqüenciació paral·lelitzada de fins a 384 fragments de DNA, de 1000 pb de llargada, amb una precisió de seqüència del 99.99% (23).

1.1.2 – La seqüenciació d'alt rendiment (la segona generació)

Tot i els grans avenços en l'automatització de la seqüenciació de DNA, les limitacions del mètode Sanger van posar en evidència la necessitat de tecnologies noves i millorades (ja que es perseguia la seqüenciació d'una gran quantitat de genomes, independentment de l'espècie). Per aquest motiu, la finalització del Projecte Genoma Humà va centrar el focus d'atenció en el desenvolupament de metodologies de major rendiment i de cost reduït, que permetessin els estudis genòmics rutinaris.

Durant la primera dècada del segle XXI van començar a comercialitzar-se una miríada de plataformes de seqüenciació basades en tecnologies ideades durant els anys noranta. Aquest avenç va provocar una reducció dràstica del cost de seqüenciació del DNA (24) (Figura 1-2), multiplicant-ne les possibles aplicacions i anant més enllà dels objectius inicials que les havien inspirat. Per aquests motius, *Nature Methods* va seleccionar la *Next-Generation Sequencing* (NGS) com a mètode de l'any 2007 (25).

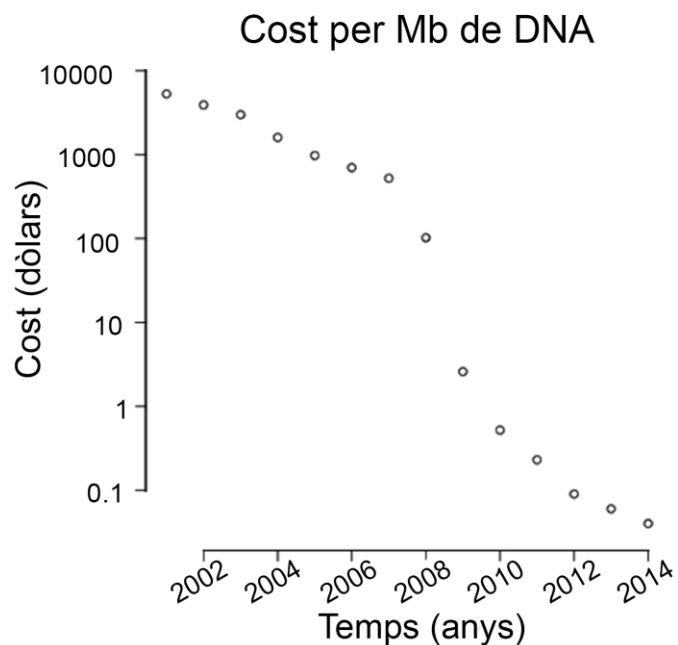


Figura 1-2 | Evolució del cost (en dòlars) de seqüenciació d'un milió de bases de DNA al *National Genome Research Institute* (24).

Tot i la gran varietat de plataformes comercials de seqüenciació d'alt rendiment (aleshores anomenades de NGS), el mercat ha estat dominat durant molts anys per tres tecnologies utilitzades arreu del món: el mètode de seqüenciació 454, de Roche (Basilea, Suïssa), el d'Illumina (Califòrnia, USA) i el SOLiD™ de *Life Technologies* (Califòrnia, USA). No és l'objectiu d'aquesta introducció el revisar el

ventall de mètodes de seqüenciació d'alt rendiment al complet, per tant, la descripció dels mètodes s'ha centrat en aquestes tres tecnologies.

Les estratègies de seqüenciació presenten alguns punts en comú i tenen un flux de treball similar. En l'etapa de construcció de llibreries de DNA genòmic, és necessari el processament del genoma de partida, de manera que els fragments obtinguts siguin representatius de la porció del genoma que es vol seqüenciar. La fragmentació pot ser aleatòria, mitjançant mètodes de fragmentació físics, com la sonicació, o dirigida, mitjançant reaccions enzimàtiques. Aleshores es crea la llibreria de fragments genòmics; uns adaptadors universals es lliguen enzimàticament als extrems dels fragments. Són aquests elements els que permetran la seqüenciació de genomes complexes amb l'ús d'encebadors de PCR convencionals. Després de la lligació i, en funció del sistema triat, les construccions resultants són immobilitzades en una superfície sòlida, ja sigui sobre una superfície plana de vidre o adherides a petites esferes de polímers plàstics encapsulades en magnetita (les anomenades *beads*). La immobilització es dona sota unes condicions estequiomètriques determinades, afavorint que una única molècula sigui capturada pel corresponent sistema. Així, milers de milions de reaccions de seqüenciació tenen lloc de manera simultània. En funció de la plataforma, l'amplificació és per PCR en pont (26) o per PCR en emulsió –emPCR– (27).

Les tres tecnologies utilitzen una estratègia de seqüenciació per síntesi, en la que la detecció de la incorporació de nucleòtids es duu a terme iterativament i en temps real. Les construccions de DNA prèviament processades es disposen en *arrays* de clústers que seran seqüenciats de manera cíclica, mitjançant una polimerasa. Gràcies al procés de pre-amplificació dels fragments capturats, els flaixos generats per l'extensió dels milers de cadenes de DNA i pel fluoròfor adjunt al nucleòtid agregat són captats per sistemes fotogràfics de detecció microscòpica. Un algoritme informàtic analitza seqüencialment les imatges obtingudes de les reaccions de síntesi i les transforma en seqüències, que són emmagatzemades en arxius de text. D'aquesta manera s'estalvia el pas d'electroforesi, el més lent i costós de les plataformes de seqüenciació de primera generació.

Les principals diferències entre els tres mètodes es resumeixen en el sistema de formació d'*arrays* de DNA, en l'amplificació i formació de clústers, i en la seqüenciació basada en l'acció enzimàtica.

I – La seqüenciació 454

L'any 2005, l'empresa 454 *Life Sciences* (l'actual Roche) va comercialitzar la primera plataforma de seqüenciació d'alt rendiment (28). El 454 *Genome Sequencer* estava basat en un sistema de seqüenciació per detecció de l'alliberament de l'anió pirofosfat inorgànic (PPi) publicat al 1998 (29). Va costar deu anys convertir la idea inicial del Dr. Hyman, la piroseqüenciació, en un mètode funcional (30).

Un cop fragmentat el DNA genòmic, es lliguen els adaptadors universals. Les construccions s'uneixen a *beads* recobertes d'estreptavidina gràcies a la biotina de l'extrem 5' d'un dels dos adaptadors. L'estequiometria de la reacció és controlada, per assegurar que la majoria de *beads* no portin més d'una única molècula de DNA motlle de cadena única. L'amplificació clonal dels fragments de DNA es porta a terme per emPCR (Figura 1-3/A): les *beads* són separades en una emulsió d'oli i aigua, i l'amplificació té lloc dins de les gotetes d'oli, que fan la funció de microreactors, on també hauran quedat capturats els reactius necessaris per a que la PCR es pugui dur a terme (27). Després dels cicles corresponents d'amplificació es trenca l'emulsió. Les *beads* són tractades amb un agent desnaturant i enriquides mitjançant una estratègia d'hibridació a fragments de DNA motlle (28). Aquestes *beads* enriquides són carregades en una placa *PicoTiter*, que presenta una gran quantitat de pouets de 28 μm de diàmetre capaços d'acceptar únicament una *bead* (Figura 1-4/A). Això permet l'establiment en una posició fixa, en la que cada reacció de seqüenciació pot ser monitoritzada. També s'hi afegeixen *beads* accessorïes més petites carregades amb els enzims necessaris per la seqüenciació: l'ATP sulfurilasa, la lluciferasa i l'apirasa.

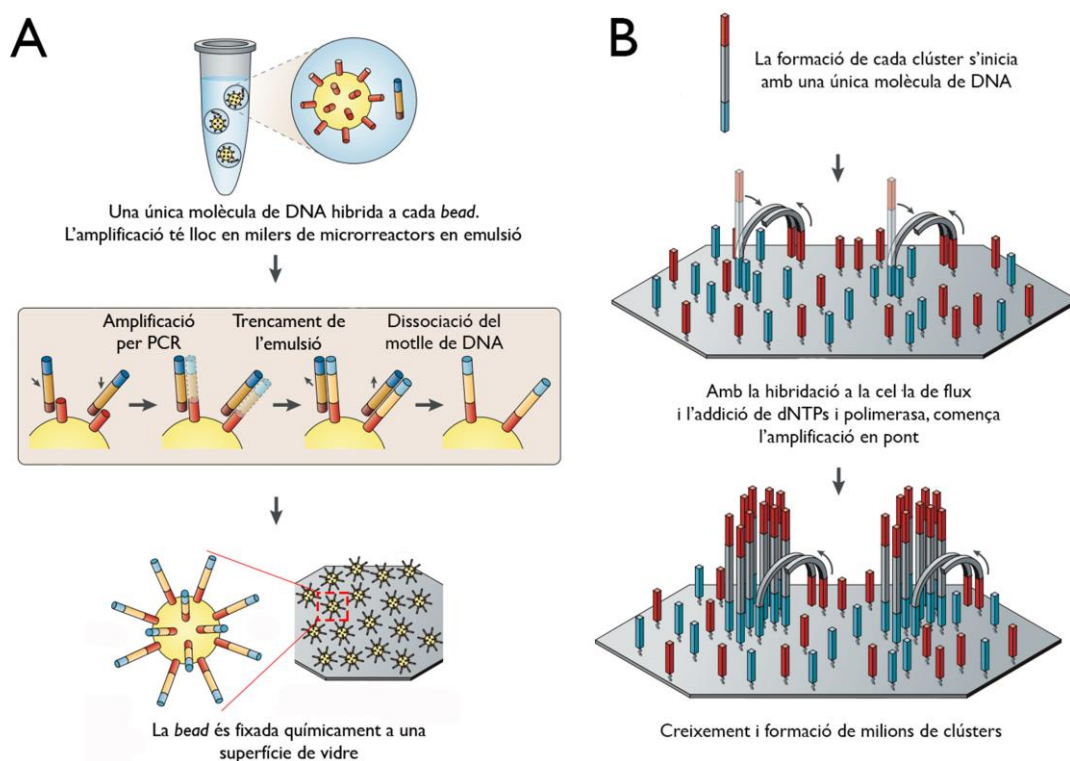


Figura 1-3 | Diferents estratègies d'immobilització de DNA motlle: **A)** emPCR; **B)** Amplificació en fase sòlida d'Illumina, mitjançant PCR en pont. Adaptada (31).

A cada cicle de seqüenciació es deixa fluir, al llarg de la placa, un únic tipus de nucleòtid trifosfat. En aquelles cadenes en les que la DNA polimerasa catalitzi l'addició d'un nucleòtid, s'alliberarà un PPI, que permetrà l'oxidació de la lluciferina per acció de l'ATP sulfurilasa i la lluciferasa, emetent llum

bioluminescent (32). La presència o l'absència de l'emissió de llum a cada pou indica la incorporació del corresponent nucleòtid i, per tant, revela la identitat de la base complementària en el motlle de DNA. El cicle de seqüenciació acaba amb la degradació dels nucleòtids no incorporats per acció de l'apirasa. La quantitat de llum produïda en la reacció catalitzada per la lluciferasa s'estima mitjançant un dispositiu sensible a la llum, amb el que delimita la placa, com un luminòmetre o una càmera fotogràfica (29). Els arxius resultants contenen els diagrames de flux de cada pou, a partir dels quals es deriven les seqüències de DNA.

El principal avantatge de la seqüenciació 454 és la velocitat i la capacitat de seqüenciar lectures de DNA més llargues que les altres plataformes, arribant fins a les 500 bases al moment de sortida, i fins a les 1000 en l'última versió del seqüenciador, al 2010. També, a diferència d'altres tecnologies de seqüenciació d'alt rendiment, no requereix d'altres processos químics per l'extensió de la cadena de DNA que no siguin els processos bioquímics usuals de la DNA polimerasa (no ha de fer rentats de nucleòtids, o desbloquejar els terminadors, com en el cas del mètode d'Illumina). Això redueix la generació d'artefactes en el procés de seqüenciació. No obstant, el processat asincrònic dels nucleòtids limita el sistema a l'haver de seqüenciar regions homopolimèriques. La llargada d'aquestes regions ha d'inferir-se a partir de la intensitat de la senyal òptica en un procés amb certa propensió a l'error. La conseqüència directa (i marca distintiva del mètode) és la inclusió d'errors a les seqüències en forma d'insercions i delecions (33).

La plataforma va deixar de rebre suport l'any 2013, moment en el que Roche va anunciar que cap al 2016 deixaria de comercialitzar-la degut als problemes inherents del mètode i a la baixa qualitat de les seqüències, en comparació amb les de la competència.

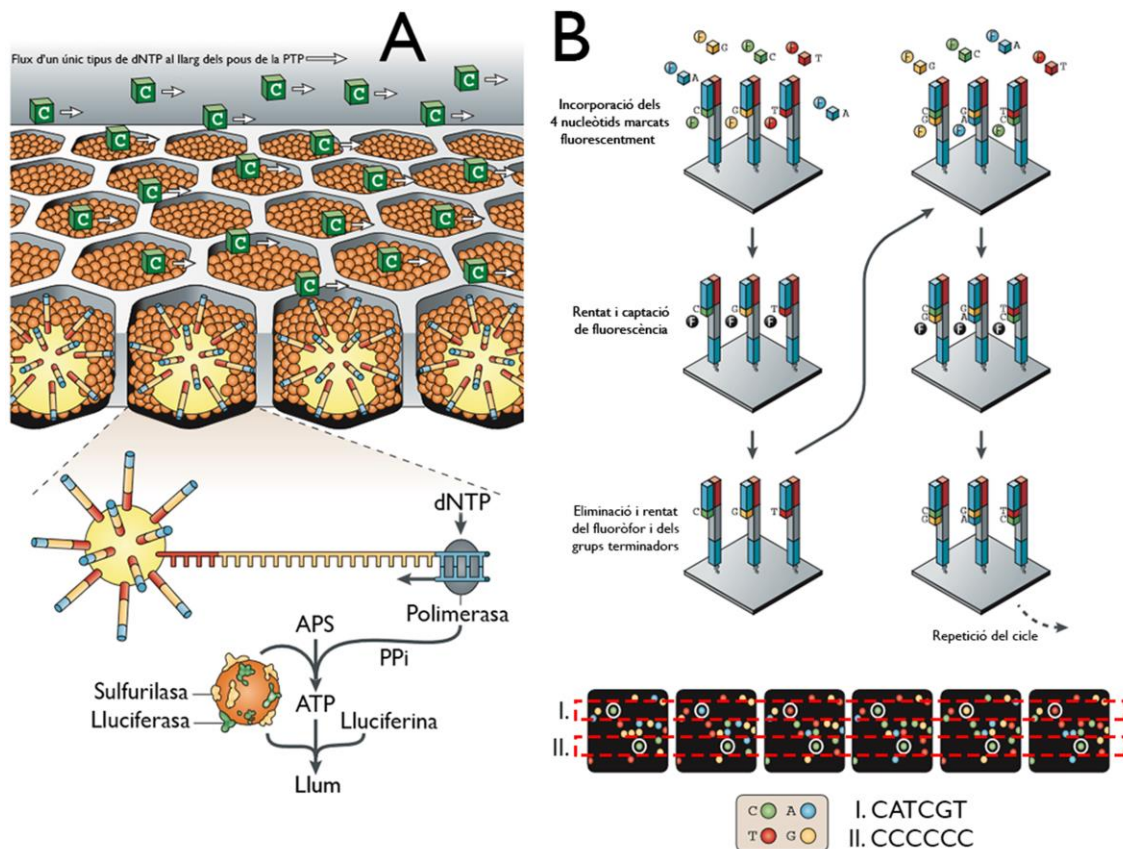


Figura 1-4 | Representació esquemàtica de dos mètodes de seqüenciació per síntesi. **A)** La reacció de piroseqüenciació, del mètode 454; **B)** La terminació reversible cíclica en 4 colors del mètode de seqüenciació d'Illumina. Adaptada (31).

II – El mètode de seqüenciació d'Illumina

La tecnologia en la que es basa el mètode de seqüenciació d'Illumina deriva de les idees de meitat de la dècada dels noranta de Shankar Balasubramanian i David Klenerman (fundadors de Solexa, més endavant adquirida per Illumina). Es van servir del concepte de terminació reversa, proposat inicialment per Canard i Sarfati (34), i acabat de desenvolupar per Turcatti et al. (35,36). Així, el 2006, Solexa va comercialitzar el *Genome Analyzer Classic*, una plataforma de seqüenciació capaç de generar entre 2 i 3 Gb de seqüències al dia, amb una precisió per base superior al 99%, però amb la llargada de les seqüències limitada a 36 bases (37).

La construcció de les llibreries de DNA genòmic és similar a l'exposada per la seqüenciació 454. La característica distintiva del mètode és l'amplificació en fase sòlida del DNA, que té lloc en una superfície de vidre, l'anomenada cel·la de flux, on es realitza la PCR en pont (Figura 1-3B) (35). Per fer-la possible, els adaptadors que flanquegen els fragments de DNA hibriden amb una capa d'oligonucleòtids que recobreix la superfície de vidre. Les polimerases creen una cadena complementària a la del fragment que ha hibridat i el motlle original és eliminat. L'extrem lliure de la cadena nova hibrida amb un altre oligonucleòtid de la superfície de la cel·la de flux, adquirint la

característica forma de pont. Aquest procés és repetit de manera simultània milers de vegades fins que, després de diversos cicles, un agent desnaturalitzador desmunta les construccions. La immobilització assegura que quedin regions amb una alta densitat d'amplicons de cadena única de DNA i adaptadors, basats en un únic fragment inicial de DNA genòmic –els clústers–.

Després de la generació de clústers arriba el moment de la seqüenciació. Els encebadors universals hibriden a l'adaptador de l'amplicó que encara queda lliure. El sistema utilitza un mètode de seqüenciació per síntesi amb nucleòtids marcats per fluoròfors i terminadors de cadena reversibles. A cada cicle d'interrogació de seqüència, els quatre tipus de nucleòtids són afegits de manera simultània als canals de la cel·la de flux juntament amb una DNA polimerasa modificada. Cada extrem 3' dels nucleòtids és bloquejat per un grup hidroxil per prevenir més addicions de bases (38). La fluorescència de tots els amplicons del clúster que hagin captat un nucleòtid és emesa en un únic flaix i captada per un sistema fotogràfic. El cicle de seqüenciació acaba amb l'escissió química del fluoròfor i del grup bloquejador, permetent la unió d'una altra base al següent cicle (Figura 1-4B). El processament de les imatges obtingudes i el filtratge de la informació de mala qualitat resulta en la generació dels arxius fastq, els contenidors de la seqüència del DNA d'interès.

Quan va sortir al mercat, la plataforma demostrava un rendiment milers de vegades més elevat que les plataformes de seqüenciació de primera generació. El seqüenciador era capaç de processar 80 milions de motlles de DNA de manera simultània, però el principal inconvenient que presentava era la llargada de les seqüències, d'únicament 36 bases. Aquesta limitació venia donada pel decaïment de la qualitat en l'emissió de la senyal òptica al llarg dels cicles de seqüenciació. El rendiment del procés no és del 100% i sempre es cometen errors, ja sigui a l'hora de l'escissió del fluoròfor, en l'eliminació del grup bloquejador o en el rentat de nucleòtids d'un cicle a un altre. Això provoca que algunes cadenes de DNA s'estenguin en asincronia amb les cadenes germanes del clúster, causant un desfasament en l'emissió de fluorescència (procés conegut com a *phasing*) (23). A més, a causa de l'ús de polimerases i nucleòtids modificats, els errors de seqüència més habitualment reportats eren les substitucions (39).

Amb el temps s'ha millorat el *hardware* dels seqüenciadors d'Illumina, amb l'addició de miralls que multipliquen la capacitat de captació de fluorescència i amb un remodelatge substancial de la cel·la de flux. També s'han millorat els reactius dels cartutxos de seqüenciació i la qualitat de les polimerases utilitzades. En l'actualitat, el seqüenciador de més capacitat d'Illumina, el NovaSeq-6000, és capaç de seqüenciar 6000 Gb per *run*, el que equival a 20 bilions de lectures aparellades de 150 bases –o a poc menys de 2000 genomes humans per *run*– (37).

El sistema de seqüenciació SOLiD™ de *Life Technologies (Sequencing by Oligo Ligation and Detection)*, va ser descrit per primera vegada l'any 2005 per Shendure et al. i comercialitzat a finals del 2007 (40,41).

Després dels passos inicials de preparació de llibreries (comuns amb els altres mètodes), el sistema SOLiD™, al igual que la seqüenciació 454, utilitza l'emPCR com a mètode d'amplificació del DNA motlle, mitjançant *beads* magnètiques (27). Després de l'amplificació, les *beads* són enriquides i fixades per unions covalents en un substrat de vidre (41).

A diferència de la resta de mètodes, la seqüenciació per síntesi no és conduïda per una polimerasa, sinó per una lligasa. En primer lloc, un encebador universal s'uneix a un dels adaptadors lliures dels amplicons de DNA (Figura 1-5). Aleshores un conjunt de sondes octamèriques amb algunes bases degenerades i marcades amb un fluoròfor a la vuitena posició són lligades en sèries successives de 7 amb les cadenes de DNA en extensió. Cada cop que es lliga la sonda, el fluoròfor s'escindeix químicament entre les posicions 5 i 6 d'aquesta, emetent fluorescència –que és captada per la plataforma– i permetent la lligació de la següent sonda. Les rondes successives de lligació permeten la interrogació de diverses combinacions de parelles de bases en primera posició al llarg de la cadena de DNA en extensió. Després de 7 rondes de lligació, la cadena en extensió és alliberada i el sistema es reinicia amb el decalatge d'una base a l'encebador. D'aquesta manera al final del procés es cobreix tota la seqüència de l'amplicó (23). Tot i que sigui un sistema combinatori complex, és un procés ben organitzat i dirigit per un sistema automatitzat al 100%. Després de l'anàlisi, les seqüències són recollides en arxius específics per aquest sistema, els *color space fasta*, que poden ser convertits a format *fastq* per diversos programes (42).

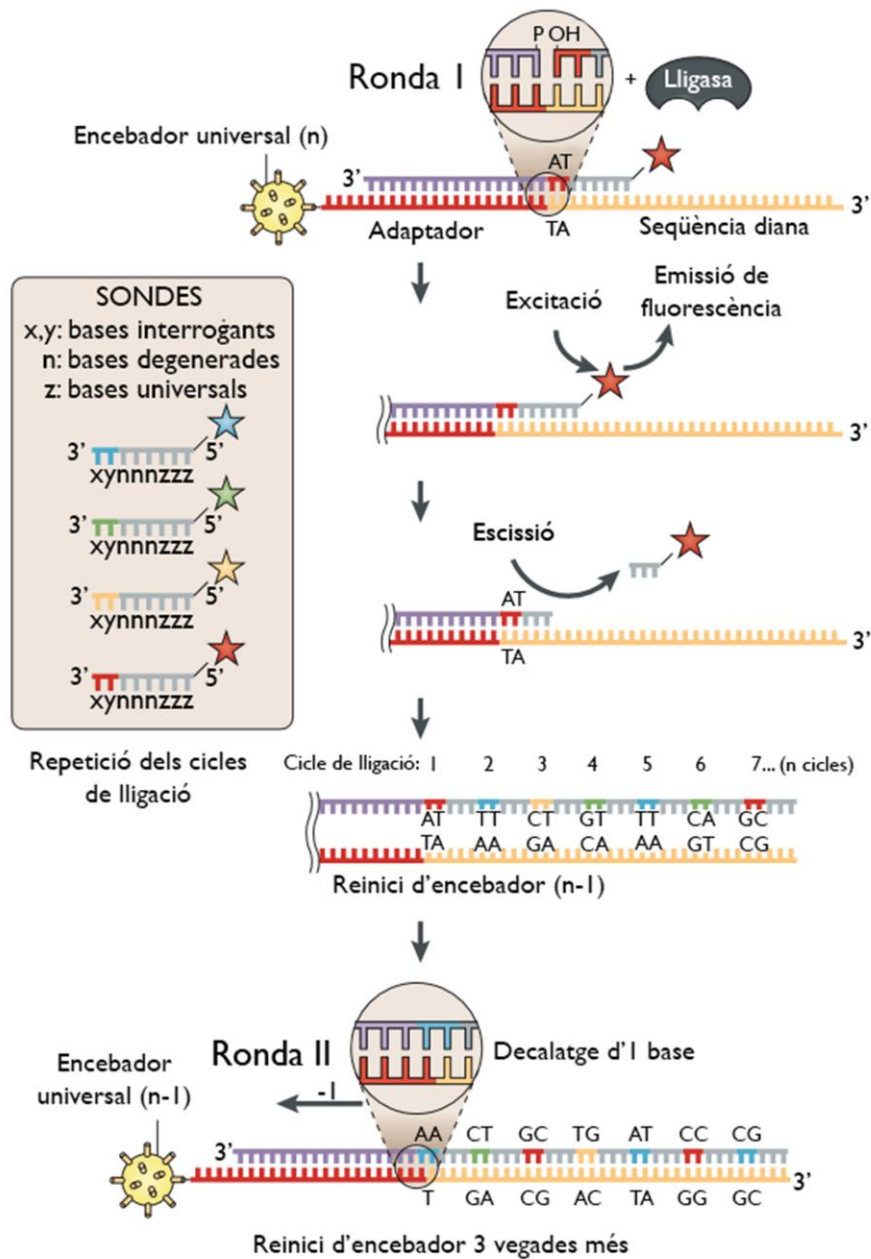


Figura I-5 | Procés de seqüenciació per lligació d'octàmers, utilitzat pel sistema SOLiD™. Adaptada (31).

Una segona característica distintiva del mètode és que les dues primeres bases de cada sonda s'associen a un fluoròfor específic (dels 4 disponibles). Aquest fet, combinat amb l'ús alternatiu d'encebadors i de sondes degenerades, fa que com a mínim cada base sigui seqüenciada un parell de vegades, permetent la correcció d'aquelles posicions en les que es cometi un error d'identificació. La limitació més important del sistema és la curta llargada de les seqüències (43). Al moment de sortir al mercat, la plataforma podia seqüenciar lectures de fins a 50 bases; en l'actualitat pot generar fins a 320 Gb de lectures de tant sols 60 bases.

IV – La seqüenciació *paired-end*

El problema que presenta la limitació de la llargada de les seqüències que generaven les plataformes de segona generació va ser parcialment superat pel desenvolupament de diversos protocols de seqüenciació *paired-end*. El concepte va ser descrit inicialment al 1981 i aplicat per primera vegada a les noves tecnologies a l'any 2009 (44,45).

La idea de fons consisteix en que, un cop fragmentat el DNA, els adaptadors que es lliguen poden hibridar amb dos encebadors diferents. Primer es seqüencia el DNA motlle utilitzant l'encebador número 1 de la parella; després de la desnaturalització del producte, la segona ronda de seqüenciació es duu a terme utilitzant el segon encebador. D'aquesta manera s'obtenen tres informacions diferents: les dues seqüències aparellades i el coneixement afegit de que les dues es troben a una mida aproximada a la mida inicial del fragment de DNA motlle, depenent els cicles d'amplificació als que s'hagi sotmès la llibreria. (38). Aquest sistema ofereix la possibilitat d'alinejar les seqüències obtingudes a la referència utilitzant la distància intermèdia com una informació afegida.

La seqüenciació *paired-end* resulta de gran ajuda per millorar l'alineament de seqüències curtes i per salvar-ne algunes que d'altra manera podrien descartar-se per presentar diversos llocs d'alineament al llarg del genoma de referència. Diversos alineadors ofereixen la possibilitat d'alinejar seqüències en regions ambigües o repetitives sempre i quan la seva parella sigui alineada de manera exacta (42). A més, el sistema resulta de gran utilitat per la detecció de variants estructurals i per la resolució dels punts de trencament d'aquestes.

V – La revolució dels seqüenciadors de sobretaula

El moment culminant de la segona generació arriba amb la comercialització de les plataformes de seqüenciació de sobretaula: seqüenciadors de mida reduïda, i de menor capacitat, però de gran importància, ja que feien abastable la seqüenciació d'alt rendiment a grups d'investigació econòmicament modestos, i obrien la porta de la recerca genètica contemporània a un ventall molt més ampli d'usuaris. En general, els protocols de preparació de llibreries van simplificar-se i els costos de seqüenciació de DNA van reduir-se dràsticament (Figura 1-2).

La primera plataforma de sobretaula en comercialitzar-se va ser el 454 *Genome Sequencer Junior System*, de Roche. Era capaç de generar 70 Mb de seqüència al dia, a 100.000 seqüències per *run*, de 400 bases de llargada. Al 2011 va sortir al mercat el primer seqüenciador personal d'Illumina, el MiSeq. En aquell moment el MiSeq generava 1.5 Gb per cada *run* de 10 hores amb seqüències *paired-end* de 150 bases. En l'actualitat s'ha perfeccionat el funcionament d'aquests seqüenciadors amb l'addició de millores en el *hardware* i en els reactius de seqüenciació. Avui en dia, un MiSeq pot generar, depenent del protocol i de les condicions de l'assaig, fins a 25 milions de seqüències *paired-end* de 300

bases (l'equivalent a 15 Gb de DNA). L'oferta de seqüenciadors de sobretaula és àmplia, fent possible l'adquisició de la plataforma adient per les necessitats de grups de recerca molt diversos.

En aquesta tesi, les mostres s'han seqüenciat amb un MiSeq. Per aquest motiu, a l'apartat de Materials i Mètodes es parla del protocol de preparació i de càrrega de llibreries de DNA genòmic per aquesta plataforma. A més s'aprofundeix en els sistemes de control de qualitat del *run* i de les seqüències generades, necessari per minimitzar la possible interferència del *phasing* en les seqüències finals.

1.1.3 – La *Next-Generation Sequencing*

En una època en la que s'han explotat intensivament les metodologies de seqüenciació de segona generació, la comunitat científica ha establert nous reptes a assolir relacionats amb la seqüenciació del DNA. Avui en dia, 14 anys després de la publicació del genoma humà, encara queden aproximadament 160 regions del genoma (algunes d'elles associades a malalties) impossibles de ser seqüenciades amb els mètodes resumits anteriorment. Ho impossibiliten els biaixos de la tecnologia – l'ús de protocols dependents d'amplificacions amb PCR, o les dificultats per amplificar de manera òptima regions amb continguts GC extrems, per exemple– i la complexitat d'aquestes regions del genoma, sovint repetitives. A més, el fet d'haver d'alinejar les seqüències a una referència haploide també pot complicar l'anàlisi d'un gran número de fragments, en funció de la variabilitat que presenti l'individu en la regió que es desitja seqüenciar. Per aquest motiu, la innovació es dirigeix cap a la producció de seqüències molt més llargues, que facilitin la qualitat de l'alineament al llarg d'un percentatge major del genoma (46), o al desenvolupament de mètodes de seqüenciació que presentin una rapidesa extraordinària i una versatilitat sense precedents a l'hora de poder seqüenciar DNA en terrenys complicats i amb protocols de preparació molt simplificats (47).

D'aconseguir aquestes fites, de manera colateral es podran assolir altres objectius importants, com l'obtenció d'un genoma humà de qualitat a un preu de 1000 dòlars, o la seqüenciació de genomes d'alta qualitat al 100% de seqüència –els anomenats *Platinum Genomes*– (48). La tecnologia desenvolupada per assolir aquests objectius és la SMRT (de l'anglès *Single Molecule Real-Time Sequencing*), amb la qual es pot seqüenciar un sol fragment de DNA de varies kilobases (Kb) de llargada sense haver d'aturar el procés entre les etapes de lectura. Aquesta és la principal característica distintiva de les plataformes de seqüenciació de tercera generació (49).

I – PacBio RS System

A mitjans del 2011 va comercialitzar-se el primer seqüenciador de tercera generació, el *PacBio RS System*, de *Pacific Biosciences* (Califòrnia, USA). El sistema aprofita al màxim les propietats

intrínseques de la polimerasa, emulant els processos interns de la cèl·lula durant la replicació del DNA. Per fer-ho es serveix principalment de dos recursos: l'ús de nucleòtids amb el fluoròfor lligat al grup fosfat terminal per visualitzar l'activitat de la polimerasa (en comptes de portar-lo unit al sucre, com els utilitzats per les plataformes de segona generació); i l'ús de cambres de visualització nanofotòniques anomenades *Zero-Mode Waveguide* (ZMW): pouets nanoscòpics de 20 zeptolitres de volum (10^{-21} litres) retroil·luminats per un làser excitador i amb un disseny que evita la fuga de la llum emesa, per tenir una longitud d'ona excessivament llarga (50). La llum atenuada del làser penetra els 20-30 nanòmetres inferiors de cada ZMW, creant així el microscopi més poderós del món (Figura 1-6).

Un complex de DNA motlle nadiu (sense haver estat amplificat) i polimerasa s'immobilitza a la base de cada ZMW, on difonen els nucleòtids modificats. Quan s'incorpora una base, es produeix l'emissió d'un pols de llum d'uns mil·lisegons de durada, provinent del fluoròfor. Aquesta llum queda retinguda a la base de la ZMW, on és captada. Com el volum del pou és tant petit, el soroll de fons és reduït unes 1000 vegades en comparació al que es genera en una cel·la de flux convencional. Després de la incorporació, la cadena del fosfat és processada i el fluoròfor queda alliberat, difonent cap a l'exterior de la ZMW. La reacció finalitza quan el complex es desencadella, un cop acabada la seqüenciació del fragment. El procés té lloc en paral·lel en els milers de ZMW distribuïts en les cel·les de SMRT del seqüenciador.

Com ja s'ha comentat, el sistema no necessita etapa d'amplificació. D'aquest punt se'n desprenen dues conclusions: la primera és que el rendiment de la seqüenciació és menor en comparació amb els mètodes de segona generació (entenent com a rendiment la quantitat total de material seqüenciat per un únic fragment de DNA); la segona és que s'eviten els biaixos propis de l'etapa d'amplificació, fent que el procés sigui extraordinàriament uniforme en regions amb continguts GC extrems. Per altra banda, la gran llargada de les seqüències generades (aproximadament un 50% del total de seqüències generades oscil·la entre les 10 i les 15 Kb (46), arribant a un màxim de 30-35 Kb) permet cobrir una porció superior del genoma amb una quantitat molt inferior de seqüències solapants; això optimitza el procés, ja que comporta un estalvi de reactius i una reducció del temps de seqüenciació important.

Com les seqüències obtingudes són tant llargues, aquestes inclouen molts errors, ja que tota polimerasa té una taxa d'error associada. Això, en seqüències de 20 Kb es tradueix en un 15% de bases amb qualitats Q12-Q15, fent que la identitat de seqüència mitjana amb el genoma de referència sigui d'un 85%. Ara bé, això tant sols suposa un problema si la cobertura total de la regió seqüenciada és inferior a 8x. Amb aquest mínim de cobertura, la identitat de seqüència consens resultant és del 99.999%, ja que és altament improbable que la polimerasa s'equivoqui sempre en la mateixa posició de la seqüència (51). Un altre punt feble de la tecnologia, potser el més important, és el problema que suposa la detecció d'insercions o delecions d'1-2 pb.

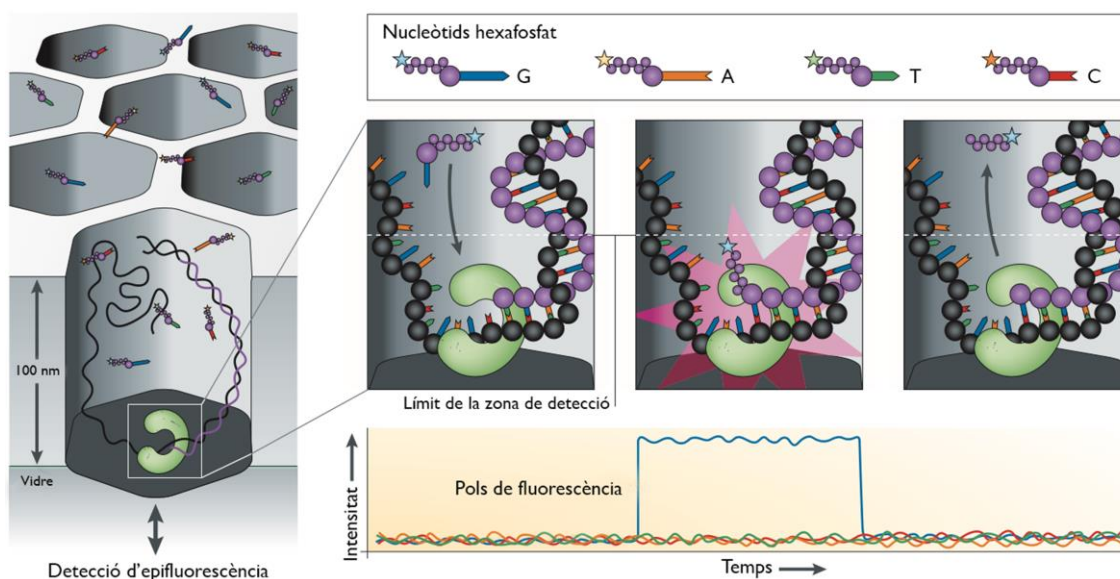


Figura I-6 | Procés de seqüenciació SMRT. Es mostra l'interior d'una ZMW. Adaptada (31).

La plataforma de seqüenciació més potent comercialitzada avui en dia per *Pacific Biosystems* és *el Sequel System*. Dins de cada cel·la de SMRT de la plataforma hi han organitzades un milió de ZMWs. La plataforma pot seqüenciar un genoma humà a una cobertura de 10x en tant sols 30 hores i amb 5 cel·les SMRT dedicades (de les vàries desenes que pot encabir).

II – MinION Nanopore Sequencing

El dispositiu USB de seqüenciació MinION, d'*Oxford Nanopore* (Oxford, UK), de tant sols 100 grams de pes, va sortir al mercat a mitjans del 2015. El sistema de seqüenciació es basa en el mesurament dels canvis en la corrent elèctrica que provoca una molècula de DNA de cadena única en passar a través d'un nanoporus proteic a una velocitat de 450 bases per segon. En passar, la cadena interromp el flux iònic, resultant en canvis detectables en la corrent, dependents de la seqüència de nucleòtids (52). La tecnologia és capaç de seqüenciar molècules extremadament llargues de DNA (de fins a 50 Kb), però de la mateixa manera que succeeix amb la tecnologia de *Pacific Biosystems*, les taxes d'error són elevades (de fins a un 20%), motiu pel que cal obtenir una seqüència consens a partir de vàries seqüències per poder sotmetre la regió a estudi amb una mínima seguretat (53).

Tot i que la tecnologia encara està en fase de perfeccionament, té característiques que la converteixen en una opció molt interessant. Els protocols de preparació de llibreries són d'una senzillesa portada al límit, i en qüestió de 10 minuts la mostra és preparada per la seva seqüenciació. Aquesta senzillesa i mida reduïda (com el palmell de la mà) fan d'aquest mètode l'opció més òptima per poder seqüenciar qualsevol organisme en terrenys inhòspits o sota condicions poc controlades (47). A més, és un instrument versàtil, ja que pot detectar modificacions de bases, com la metilació (54) i de seqüenciar el DNA i l'RNA en temps real, podent tornar a iniciar la seqüenciació des del mateix punt si

així es requereix.

El MinION disposa de 512 canals de nanoporus per cel·la de flux, i pot generar 10 Gb de seqüència al dia. Pot acoblar-se en sèrie, en plataformes que fan la funció de suport, dissenyades específicament per sumar la capacitat dels dispositius individuals. Així es pot trobar el GridION, amb capacitat per encabir 5 MinIONs, i el PromethION, amb capacitat per 48 cel·les de flux independents de 300 nanoporus cada una (amb un total de 144000 canals). Inclús s'està desenvolupant el SmidgION, el dispositiu de mida més reduïda, capaç de seqüenciar amb l'energia que captada de l'*smartphone* al que estigui connectat.

1.2 – El codi genètic humà

La seqüència del codi genètic humà, publicada a principis del segle XXI (18,19), consta de 3.234'83 milions de pb (3.2 gigabases –Gb–) condensades i empaquetades en 23 parells de cromosomes (22 parells somàtics i un de sexual). A l'esbós original de la seqüència es va predir un número aproximat de 30000 gens codificants per proteïnes (18), número que ha anat disminuint fins a dia d'avui, en el que es consideren menys de 19000 – l'equivalent a un 2% del genoma–. Aquest percentatge d'ocupació del genoma per gens codificants és molt menor a l'esperat per un organisme tant complex com l'humà (55). El motiu i la solució a aquesta paradoxa s'atribueix als complexos mecanismes de regulació del 98% restant del codi genètic.

El DNA no codificant el constitueixen un ampli ventall d'elements, i tot i que alguns no tenen una funció determinada o coneguda, altres desenvolupen funcions de gran importància, com la de protegir la informació indispensable del genoma. Els exons –els fragments de seqüències codificants dels gens– apareixen molt separats els uns dels altres per contrarestar l'efecte de possibles mutacions que podrien causar un canvi en la pauta de lectura, o per evitar que els processos de creuament genètic afectin informació sensible del genoma. En aquest sentit, els introns –les seqüències que flanquegen els exons i que són obviades per la maquinària cel·lular de transcripció del DNA a l'RNA– i les regions no traduïdes dels RNA missatgers representen el 26% del genoma (Figura 1-7). Una altra funció primordial del DNA no codificant és la regulació de l'expressió gènica. Actualment hi han catalogades centenars de milers de regions genòmiques funcionals destinades a aquesta tasca, una clara demostració de que es destina molta més logística genòmica al control que a la codificació de proteïnes (56). Existeixen seqüències reconeixibles per factors de transcripció –proteïnes d'unió a DNA que controlen la transcripció a RNA–(57). Segons la funció del lloc d'unió, aquest s'anomena activador –o *enhancer*– (58), silenciador (59) o *insulator* –un element separador– (60). També es codifiquen molècules de RNA funcional, com l'RNA ribosòmic, el de transferència, l'RNA d'interacció amb piwi i els micro RNAs. Totes aquestes espècies d'RNA controlen l'activitat del procés de traducció del 30% dels gens codificants per proteïnes en mamífers (61). La regulació també s'assoleix mitjançant la codificació d'interruptors genètics, que regulen on i quan un gen pot expressar-se (62), com les molècules d'RNA llargues i no codificants –els *long non-coding RNAs*–, que assisteixen en la prevenció del càncer de mama (63). En total s'han predit més de 3 milions de regions reguladores de DNA; aquestes poden contenir més de 15 milions de llocs d'unió de factors de transcripció, i aproximadament unes 150.000 es mostren actives en cada un dels diferents tipus cel·lulars (56).

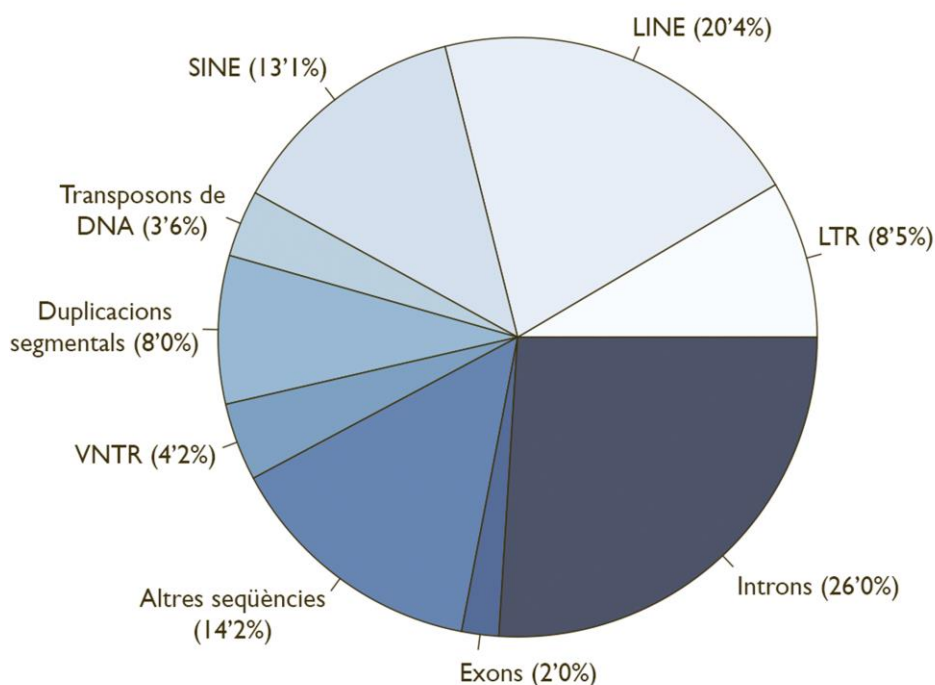


Figura I-7 | Composició del genoma humà. Adaptada (64).

Uns altres elements presents al DNA no codificant són els pseudogens, seqüències derivades de gens del propi genoma que han perdut la capacitat d'expressar-se. Són subproductes evolutius apareguts per retrotransposició o per la duplicació de gens funcionals. Han esdevingut fòssils genètics degut a les mutacions acumulades durant milers d'anys, que prevenen la transcripció del gen afectant la regió del promotor –mitjançant l'aparició de codons *stop* o amb mutacions que canviïn el marc de lectura (65)–. Hi han aproximadament uns 20.000 pseudogens coneguts, i tot i ser no codificants, molts d'ells han demostrat tenir papers importants tant en la fisiologia normal com en processos patològics (66).

El genoma humà és el resultat de processos naturals en constant evolució. A banda dels elements descrits anteriorment, el genoma actual és constituït en més d'un 50% per fragments redundants o d'alta identitat de seqüència amb altres regions genòmiques (67). Entre aquests elements repetitius, alguns desenvolupen funcions importants, com la protecció de la integritat de l'estructura del cromosoma (els telòmers), o juguen un rol crític en el metabolisme del DNA dins la cèl·lula (com els centròmers). Aquest és el cas de les repeticions en tàndem, o VNTR (de l'anglès *Variable Number Tandem Repeats*), repeticions de baixa complexitat amb centenars de milers de còpies al llarg del genoma, representant un 4.2% del DNA total de la cèl·lula. Dins d'aquest grup es troben els satèl·lits de DNA, els minisatèl·lits –entre 5 i 50 repeticions en tàndem de seqüències de 10 a 60 pb, repetides fins a 1000 vegades al llarg de tot el codi–, i els microsatèl·lits –milers de repeticions en tàndem de seqüències inferiors a 10 nucleòtids, organitzades en regions de fins a 20 Kb–. Aquests, a banda de

presentar una elevada taxa de mutació i aportar una gran diversitat genètica a la població general, constitueixen un 90% de les regions centromèriques i telomèriques dels cromosomes. Classificades dins del grup de microsatèl·lits, les repeticions de triplets de nucleòtids són de particular rellevància. En alguns casos en els que succeeixen en regions codificants poden provocar desordres genètics, com passa en la malaltia de Huntington, amb l'expansió dels triplets 'CAG' al gen *HTT* (68). Un altre grup d'elements repetitius del DNA el formen les duplicacions segmentàries –també conegudes amb el nom de repeticions de baixa complexitat (o *Low Copy Repeats*, LCRs)–. Consisteixen en seqüències d'entre 10 i 300 Kb, amb una identitat de seqüència compartida del 95%, i amb una alta homologia amb diverses regions del genoma eucariota. Tot i que són poc freqüents als genomes dels mamífers, les duplicacions segmentàries representen un 8% del genoma humà. Aquest fet es relaciona amb els processos de duplicació de certes regions genòmiques que van tenir lloc fa milers d'anys, durant l'especiació dels primats (69).

Els elements genètics mòbils (els transposons i els retrotransposons) formen un altre conjunt d'elements repetitius del DNA, amb la capacitat de moure's a través del genoma. Poden inserir-se en regions concretes i arrossegar amb ells altres elements funcionals, canviant-los de localització o inclús alterant-ne el número de còpies al genoma. A més, els llocs d'escissió acostumen a ser reparats per processos que poden provocar la pèrdua de material genètic o deixar 'cicatrus' a la seqüència, predisposant a la inestabilitat genòmica (70). Als genomes individuals es troben unes 300.000 còpies de transposons de DNA, l'equivalent a un 3.6% del genoma. A diferència dels retrotransposons, aquests poden propagar-se de manera autònoma (sense un intermediari d'RNA missatger), ja que porten codificada la seqüència d'una transposasa.

Els retrotransposons, en canvi, són romanents de retrovirus endògens acumulats durant milers d'anys d'evolució que han perdut la capacitat d'infectar la cèl·lula. Els retrotransposons de tipus LTR (*Long Terminal Repeat*) consisteixen en llargues repeticions que flanquegen diverses seqüències codificants per proteïnes enzimàtiques i estructurals. Aquestes permeten la transcripció reversa i la integració del DNA complementari dins del genoma de manera molt similar a la que duria a terme un retrovirus (71). Al seu torn, els de tipus no LTR (els LINEs i SINEs, de l'anglès *Long / Short Interspersed Nuclear Elements*) depenen o bé de les proteïnes codificades per ells mateixos, o de l'acció d'altres elements mòbils per la seva propagació, respectivament. Els LINEs tenen una llargada mitja de 7 Kb; per individu es troben unes 100.000 còpies al llarg del genoma, i representen un 20.4% d'aquest. Els SINEs són molt més petits i freqüents, amb unes 1.500.000 còpies, l'equivalent al 13.1% del total de DNA cel·lular. Les més abundants són les seqüències *Alu*, uns fragments de 300 bases de llargada que s'han relacionat amb processos de control transcripcional sobre gens codificants (72–74). En conjunt, els elements genètics mòbils representen una porció significativa del codi genètic en moltes espècies; en el cas dels humans equivalen a un 45.6% del genoma (75).

Tenint en compte la composició del genoma i els processos fisiològics i patològics en els que intervenen els elements que el formen, sorgeix el dubte sobre quin percentatge del genoma és realment funcional. A l'última actualització del consorci de l'Enciclopèdia dels Elements del DNA –ENCODE– es va reportar que més d'un 80% del genoma té una activitat bioquímica identificable. D'aquest fet es va concloure que el 80% del genoma era funcional (76). La publicació va provocar certa controvèrsia, ja que no resultava clara quina era la proporció de les regions bioquímicament anotades que servien per finalitats biològiques específiques (77). Tenint en compte el ventall de línies cel·lulars estudiades, tant pel que fa al tipus com a l'estadi de desenvolupament, i sota un punt de vista clàssic sobre quines regions considerar com funcionals (això inclou les regions codificants, empremtes de DNAsa I d'alta resolució i els llocs d'interacció DNA:proteïna) s'ha predit que un 15% del genoma és funcional. Si a aquest percentatge se li afegixen les regions on s'han identificat modificacions d'histones associades a marques de promotors o *enhancers*, el percentatge de regions funcionals puja fins al 20% (77). La tasca d'identificació d'aquestes regions funcionals a nivell biològic es planteja com un gran repte per la comunitat científica. Quan es vol provar la funcionalitat d'una regió codificant s'acostuma a generar un model cel·lular i/o animal i a observar les diferències fisiològiques que provoca el canvi. En el cas de les regions no codificants, com la gran majoria són seqüències redundants, l'observació d'un canvi en el global de l'expressió gènica pot ser molt difícil d'observar.

Per fer front a aquest repte s'han desenvolupat tècniques d'edició genòmica que permeten estudiar els possibles elements funcionals del genoma a gran velocitat, encara que a un cost elevat (78). També han aparegut eines computacionals capaces de predir els llocs d'unió dels factors de transcripció, identificant quines regions del genoma poden ser més interessants per dur a terme els estudis funcionals. Trobem, per exemple, l'estudi de Lee et al., en el que es proposa un model computacional per predir quines xarxes d'elements reguladors de gens operen en els diferents tipus cel·lulars i fins a quin punt es veuen pertorbades en pacients diagnosticats amb malalties complexes (79).

1.2.1 – Les variants genètiques i l'evolució

Les variants genètiques són canvis en la seqüència del DNA dels organismes. Aquests canvis poden ocórrer quan s'exposa un organisme a una alta quantitat d'energia (radiació ionitzant), o als agents físics o químics de l'ambient en el que aquest desenvolupa la seva activitat. Però també apareixen de manera espontània, durant els processos metabòlics del DNA, normalment per errors de les polimerases durant la replicació i la reparació d'aquest (80). És per aquest motiu que els genomes de dos individus de la mateixa espècie comparteixen una identitat de seqüència superior al 99'9%, sent les variants genètiques les úniques diferències entre els dos.

Normalment, els canvis a la seqüència del DNA no provoquen cap efecte sobre el portador i la variant és considerada neutra. Això es deu a que el canvi no queda reflectit a la proteïna, ja que el

sistema de codificació d'aminoàcids per triplets de nucleòtids és degenerat –diverses combinacions de nucleòtids codifiquen pel mateix aminoàcid–. També pot ser que hi hagi un canvi d'aminoàcid, però que aquest sigui suficientment similar com per no alterar les propietats físiques, químiques i/o estructurals de la proteïna de manera significativa (sobretot si el canvi és en una regió poc crítica per la seva funció). Altres vegades, el canvi provoca efectes negatius en detriment de l'organisme portador (81). Una tercera possibilitat és que el canvi succeeixi en una regió no codificant que, a més, no afecti als processos d'expressió o de regulació del gen en qüestió. Poc sovint, la variant pot donar al portador un avantatge selectiu sobre els seus iguals. En aquests casos la variant s'anomena adaptació (82).

Generalment, les variants genètiques es classifiquen en dos grans tipus: les variants puntuals, quan el canvi afecta un sol nucleòtid (SNV, de l'anglès *Single Nucleotide Variant*) o poques bases en forma d'insercions o delecions de més d'un nucleòtid (*indels*); i les variants estructurals, amb una llargada que oscil·la entre les 50 i diversos milions de pb –en el cas dels reordenaments a escala cromosòmica– (83).

En quant a les variants puntuals, si la substitució del nucleòtid ocorre entre purines (adenina –A– i guanina –G–) o entre pirimidines (timina –T– i citosina –C–) el canvi s'anomena transició; però si es substitueix una purina per una pirimidina o viceversa es tracta d'una transversió. Les delecions i les insercions són eliminacions o introduccions de nucleòtids en la seqüència, respectivament. Una variant és sinònima quan el canvi de nucleòtid no implica un canvi d'aminoàcid en la proteïna. Tot i que aquestes variants no acostumen a alterar la seqüència aminoacídica final de les proteïnes, s'han reportat casos en els que aconseguen desregular el correcte funcionament d'aquestes, com és el cas en diversos canals iònics operats per voltatge (84). A més, és possible que una mutació sinònima creï o destrueixi un lloc de *splicing*, alterant de manera indirecta l'estructura final de la proteïna. Les variants no sinònimes poden classificar-se en variants *missense*, quan l'aminoàcid original és canviat per un altre diferent; o *nonsense*, quan la variant transforma un codó qualsevol en un d'aturada prematur, truncant la proteïna original; o bé quan el canvi transforma un codó d'aturada provocant que la traducció s'allargui més del compte, resultant en proteïnes igualment aberrants. Pel que fa als *indels*, degut a la naturalesa ternària del codi genètic, la inserció o la deleció d'un número de nucleòtids que no sigui múltiple de tres pot modificar el marc de lectura de la seqüència codificant, resultant també en una proteïna aberrant.

Les variants estructurals poden alterar el número de còpies d'una seqüència de DNA; les delecions o duplicacions de mida intermèdia superiors a 50 pb (Figura 1-8) són les anomenades CNVs (de l'anglès, *Copy Number Variants*). Però també poden ser reordenaments genòmics balancejats, com les translocacions o les inversions. Els dos tipus de variants poden veure's involucrats en processos patogènics, a banda de contribuir de manera significativa a la inestabilitat genòmica de la regió i promoure altres alteracions del DNA. Tot i que més del 99'9% de les variants reportades del genoma humà són variants puntuals o *indels*, la quantitat de bases afectades per variants estructurals és més de

100 ordres de magnitud superior (85). Un genoma típic conté aproximadament unes 2.100-2.500 variants estructurals. Entre aquestes hi han aproximadament unes 1.000 delecions grosses, 160 CNVs, 915 insercions *Alu*, 128 LINEs, 51 insercions d'elements SVA (elements amb regions de SINEs, VNTRs i *Alu*) i una mitjana de 10 inversions. Aquest conjunt de variants afecta aproximadament uns 20 milions de bases de DNA, l'equivalent a un 0'6% del total del genoma humà (86).

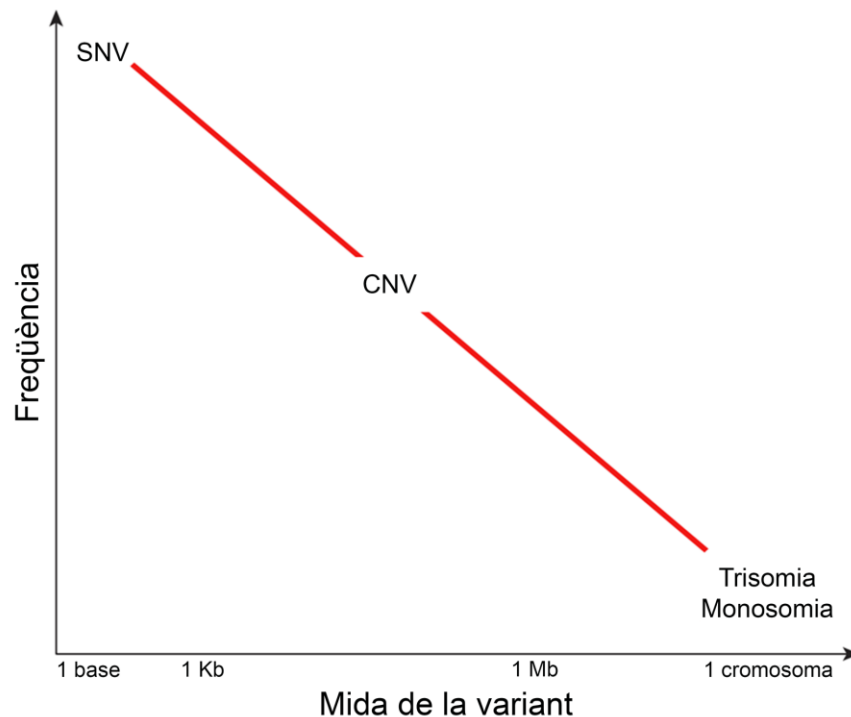


Figura I-8 | Representació de la mida i la freqüència de les principals categories de variants genètiques. Adaptada (83).

Les variants genètiques són la matèria primera de l'evolució. Les característiques genètiques de cada organisme són el resultat d'un canvi previ al genoma d'un organisme més primitiu. Quan les variants apareixen en cèl·lules constituents d'òrgans o teixits i, per tant, no transmeten la informació genètica a les següents generacions, s'anomenen variants somàtiques. Però quan la variant ocorre al codi genètic d'un gàmeta o de les seves cèl·lules progenitores, encara que no afecti fenotípicament al seu portador, la variant tindrà un impacte directe sobre la següent generació d'individus i l'oportunitat d'esdevenir més freqüent al llarg del temps. Si la variant germinal té un efecte deleteri sobre la descendència, o sigui, si la variant afecta negativament al fenotip de l'individu, fent-lo menys adaptable a l'ecosistema en el que viu o atemptant directament contra la seva salut, la variant s'associarà a un desordre genètic (87).

La selecció natural afecta als portadors de les mutacions compromentent la seva reproducció. És el que s'anomena com selecció negativa o purificadora, ja que si l'individu no arriba a reproduir-se hi ha la possibilitat de que aquella variant patogènica desaparegui per sempre. Per altra banda, si la variant

millora les capacitats adaptatives de l'individu a l'ecosistema que el rodeja, serà considerada una adaptació. En aquest cas, el portador serà sotmès a una selecció natural positiva i, amb una certa probabilitat, aquell al·lel podrà disseminar-se entre altres individus de l'espècie al llarg del temps. Per tant, les variants presents als individus d'una població, tant si són adaptatives com deletèries, són agents actius de l'evolució (87).

Durant l'última dècada, i gràcies a la immensa quantitat d'informació genòmica accessible, s'han pogut desenvolupar teories de l'origen de l'evolució del genoma humà. Així doncs, sabem que el genoma és especialment ric en duplicacions segmentàries (un 8% del total). Aquesta característica distingeix els humans dels mamífers inferiors, com els rosegadors (69). S'han reportat indicis que suggereixen que la successió de duplicacions segmentàries seriadades va orquestrar l'evolució dels primats mitjançant l'augment de la inestabilitat genòmica i la promoció de reordenaments genòmics a gran escala. Aquests processos haurien tingut un paper decisiu en la creació de nous gens (per mecanismes de fusió, fissió, i inversió genòmica) associats a una taxa inferior de recombinació i facilitant així la divergència gènica. Aquests gens nous haurien promociat l'adaptació dels humans en un entorn hostil (88,89).

Diversos estudis de genètica evolutiva han posat de relleu que les CNVs acostumen a localitzar-se allunyades de les seqüències codificants. Al seu estudi, Conrad et al. van comparar la proporció de variants puntuals presents a regions codificants delecionades amb la de regions intròniques delecionades, per concloure que les primeres apareixen significativament infrarepresentades (90). Això suggereix l'efecte de la selecció purificadora en individus amb CNVs solapant regions codificants. Aquesta hipòtesi va ser confirmada per estudis que demostraven que les CNVs es troben preferencialment apartades dels gens i dels altres elements conservats del genoma i que, en general, la proporció de delecions que solapen gens associats a malalties és significativament inferior a la de les duplicacions (91). La selecció purificadora sobre les delecions apareix com un procés transversal, intuït-se en genomes d'organismes molt diferents (92). No estranya, doncs, conèixer la implicació de les CNVs en multitud de fenotips clínics, com per exemple en les malalties neurològiques complexes: triplicacions del gen *SNCA* en pacients de Parkinson (93); duplicacions del gen *APP* en pacients d'Alzheimer (94); duplicacions a Xp11.22 en pacients amb retard mental lligat al cromosoma X (95); delecions característiques de pacients amb trastorn de l'espectre autista (96); i una gran quantitat de CNVs associades a l'esquizofrènia (96,97). També s'han relacionat CNVs amb la susceptibilitat a patir altres malalties complexes: la deleción del gen *HBA1* causa alfa Talassèmia i, al mateix temps, protegeix al portador envers la malària (98,99); l'increment significatiu de la susceptibilitat a patir SIDA per efecte del VIH en portadors d'un menor número de còpies del gen *CCL3L1* (100), al igual que succeeix amb els pacients de Lupus i glomerulonefritis amb defectes al gen *FCGR3B* (101); triplicacions del gen *PRSS1* en pacients amb pancreatitis (102) i delecions al gen *HBD2* en pacients amb malaltia de Crohn i psoriasi (103).

Tot i l'estreta relació de les CNVs amb els fenotips clínics, aquestes també han estat (puntualment) objecte de selecció positiva al llarg de l'evolució dels humans. Estudiant les amplificacions gèniques afavorides per la selecció positiva, van descobrir-se CNVs específiques del llinatge humà (104), probablement conductores del procés evolutiu que va donar peu a l'emergència de trets específics, com la cognició. Altres exemples són, per exemple, el gen *AQP7*, important per incrementar el transport de glicerol durant la mobilització de reserves d'energia i possiblement involucrat en el transport d'aigua necessari per l'excreció de la suor durant els processos aeròbics (104); o el gen codificant per l'amilasa, *AMY1*, el número de còpies del qual correlaciona positivament amb els nivells de proteïna amilasa a la saliva, necessaris per una dieta en la que el midó va anar guanyant importància progressivament (105).

1.2.2 – Mecanismes moleculars de generació de variants estructurals

Els canvis en el número de còpia en regions concretes del genoma involucren canvis en l'estructura dels cromosomes, ja que regions cromosòmiques prèviament separades poden resultar juxtaposades o encara més separades, en funció del tipus de reordenament que s'esdevingui. Els mecanismes de canvi estructural al cromosoma són els mateixos que els implicats en l'aparició de CNVs (106).

S'han descrit tres mecanismes principals de generació de variants estructurals: la recombinació homòloga, la recombinació no-homòloga i la retrotransposició (106).

I – Mecanismes mediat per recombinació homòloga

El mecanisme de recombinació homòloga al·lèlica (RHA) és la base de diversos models de reparació acurada del DNA, que es serveixen d'una seqüència idèntica (per exemple, l'al·lel de la cromàtide germana) per reparar fractures a les cadenes de DNA. Aquest mecanisme és també el responsable de la segregació ordenada dels cromosomes i de generar noves combinacions d'al·lels durant el procés de meiosi (107). El RHA requereix d'una certa identitat de seqüència, concretament de 50 pb en *Escherichia coli* i 300 en mamífers; així com també una proteïna recombinasa d'intercanvi de cadena, RecA en procariotes i Rad51, la seva ortòloga, en eucariotes. En un moment inicial del procés de recombinació homòloga, la recombinasa catalitza la invasió del dúplex de seqüència homòloga per l'extrem 3' d'una cadena simple de DNA (106). El risc de que es produeixi un reordenament genètic es minimitza amb el requeriment de fragments de seqüències amb perfecta homologia i amb la regulació de la tria correcta de molècula companya per la recombinació. La cèl·lula té fins a tres mecanismes diferents per regular aquesta tria: el primer és la reparació de variants puntuals; d'aquesta manera es dificulta la tria d'una seqüència homeòloga (que comparteix menys del 97% d'identitat), ja que és discriminada davant d'una millor opció d'aparellament; el segon mecanisme ve donat per l'acció de les

cohesines, proteïnes que mantenen properes les cromàtides germanes dels cromosomes. D'aquesta manera, la cromàtide germana sempre és la tria principal per la reparació per recombinació, per motiu de proximitat, i es restringeix l'opció d'utilitzar cadenes motlle intra o intercromosòmiques, que donarien lloc a recombinacions no homòlogues (108,109); finalment, els mecanismes de reparació de DNA eucariota són capaços de mantenir els extrems dels trencaments de doble cadena (TDC) propers, evitant la invasió de les seqüències a zones alienes amb complementarietat de bases (110,111). L'helicasa BLM, en humans, coordina la tria de motlle per fer la reparació dels extrems dels TDC (112,113).

Tot i que el mecanisme de recombinació homòloga no és considerat poc curós en la reparació, tampoc és un mecanisme infal·libre. Una recombinació defectuosa, normalment a causa de regions altament repetitives, amb LCRs o duplicacions segmentàries, pot provocar reordenaments estructurals.

a. Reparació de trencaments de doble cadena

El mecanisme de RHA més estudiat és el model de recombinació induïda per TDC, que té lloc durant la recombinació meiótica, en la sinapsi. La recombinació espontània durant la mitosi és probablement iniciada per altres tipus de lesions en cadenes simples de DNA (106).

Aquests trencaments són causats per l'efecte de la radiació ionitzant, les espècies reactives d'oxigen, o per l'acció d'agents exògens, com la quimioteràpia. També poden ser provocats pel mal funcionament (o pel reconeixement defectuós de la seqüència) dels enzims involucrats en el metabolisme del DNA. Un dels casos més documentats és el tall del DNA causat pel complex RAG (de l'anglès *recombination-activating gene*) en localitzacions diferents d'allà on hauria de dur a terme la seva funció. Són les regions de recombinació fisiològica V(D)J, involucrades en el procés de generació de la diversitat dels receptors limfocitaris B i T, integrants del sistema immunitari dels vertebrats (114,115).

Per la reparació de TDC hi han dos models principals: la formació de la doble unió de Holliday i la hibridació de cadena dependent de síntesi (Figura 1-9).

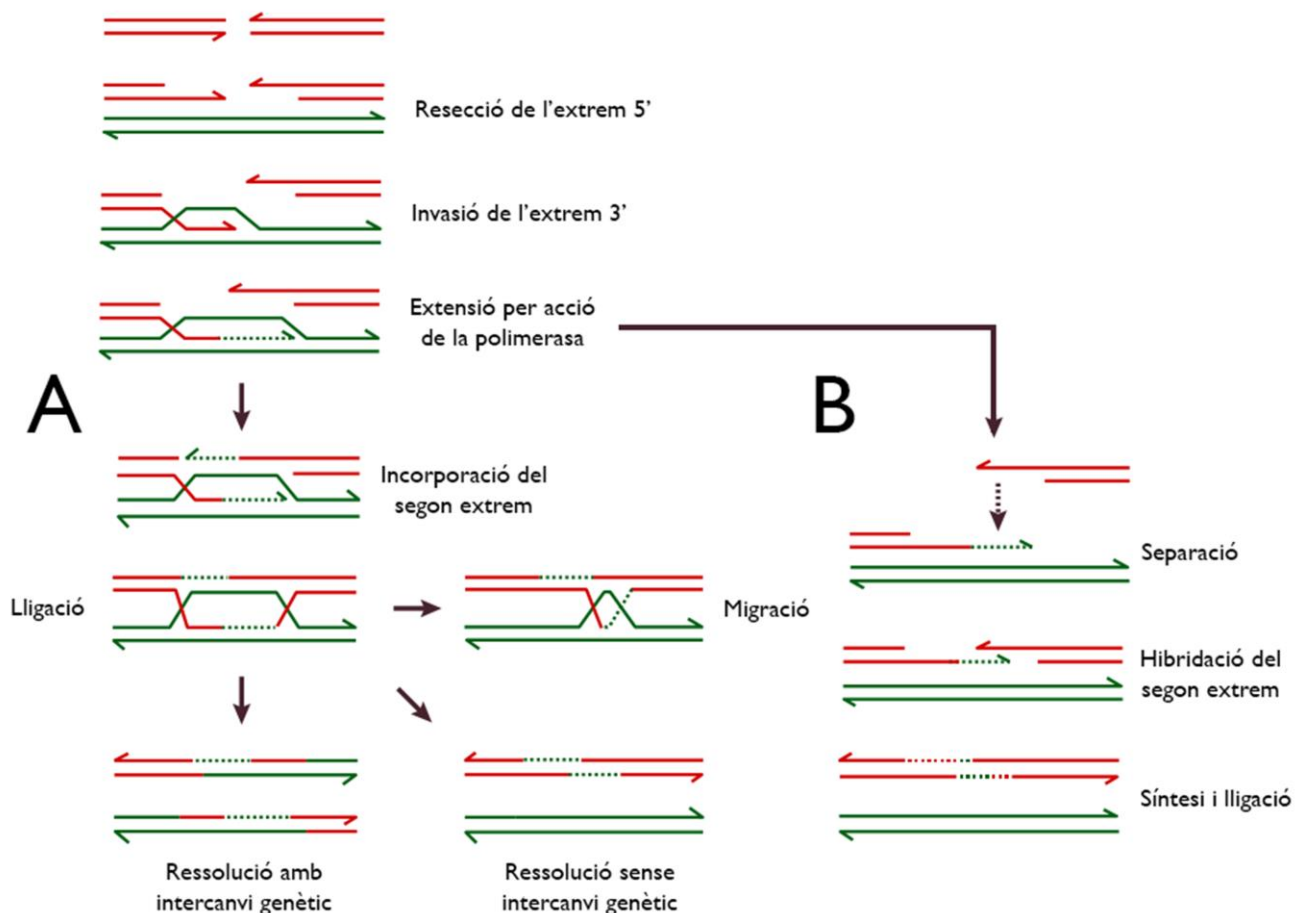


Figura 1-9 | A) Mecanisme de reparació de TDC; **B)** Hibridació de cadena dependent de síntesi. Adaptada (106).

En el primer, els extrems 5' de les cadenes on ha tingut lloc el trencament són degradats per l'acció del complex proteic MRN –que ha detectat la lesió i ha aturat el cicle cel·lular, permetent la reparació–, amb l'objectiu de deixar els extrems 3' voladissos. A aquests s'hi uneix la proteïna Rad51, catalitzant la invasió d'un dels extrems 3' a la seqüència homòloga, que hibrida i forma una estructura anomenada *D-loop*. L'extrem 3' invasor exerceix la funció d'encebador per la síntesi del DNA, que s'estén mitjançant l'acció d'una polimerasa, arribant a sobrepassar la posició del trencament original. L'extrem 3' restant és incorporat al *D-loop* per hibridació, iniciant també l'extensió del DNA. Seguidament, el pas de lligació de les dues cadenes en extensió forma una unió doble de Holliday, una estructura intermèdia entre cadenes homòlogues de DNA estabilitzada per helicases (Figura 1-10). En el pas de resolució, les unions són processades per endonucleases. En funció de la manera en que les endonucleases tallin la unió doble de Holliday, el resultat del mecanisme de reparació variarà: si les dues nucleases han resolt les unions en la mateixa direcció, no hi haurà creuament genètic; però si les direccions de tall han estat diferents, hi haurà intercanvi de fragments de DNA entre els dos

cromosomes. També és possible una resolució alternativa de la unió de Holliday en la que no hi intervinguin les nucleases, sinó mitjançant l'acció d'una helicasa i d'una topoisomerasa. De seguir-se aquesta alternativa, no hi ha risc de creuament genètic (Figura 1-9A).

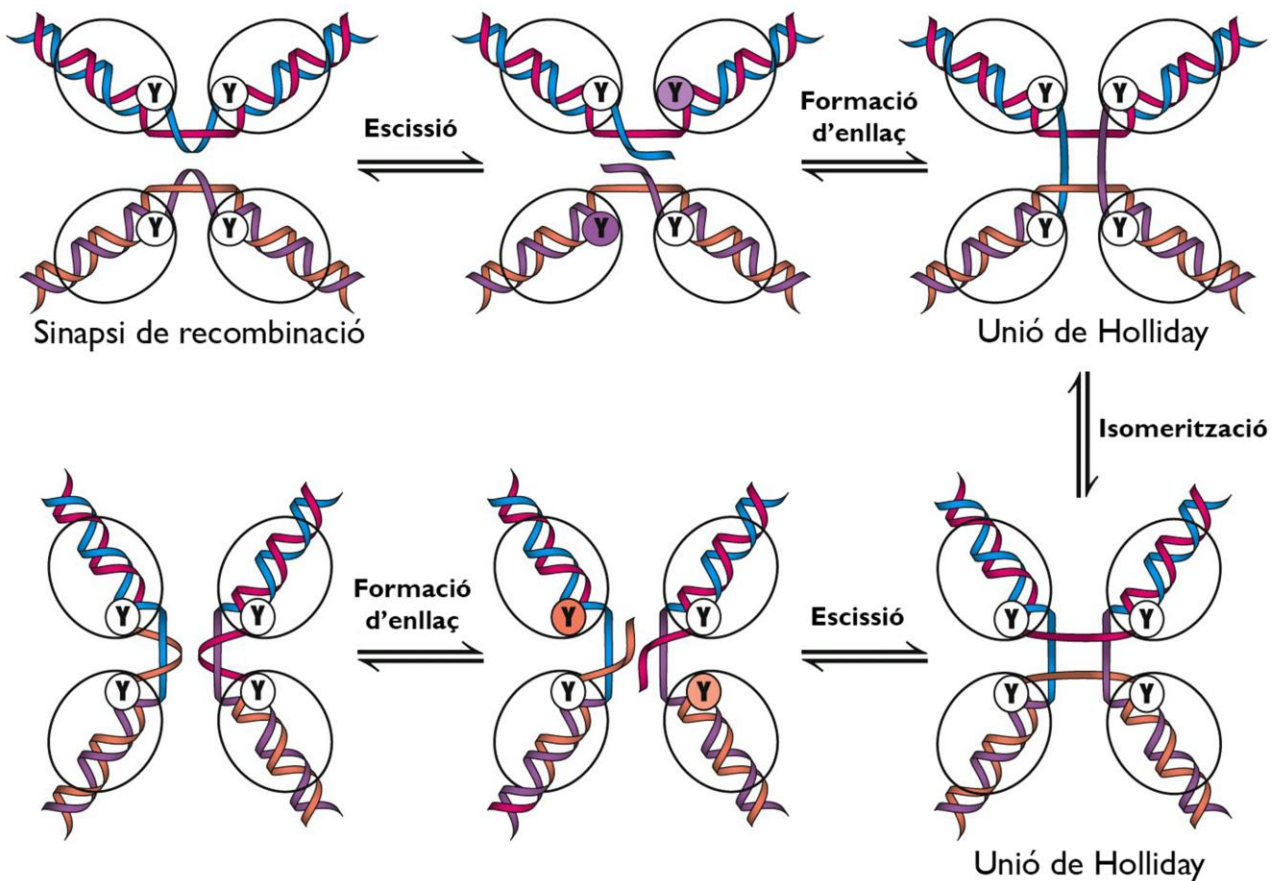


Figura I-10 | Representació esquemàtica d'una unió de Holliday; l'estructura es resol amb creuament genètic. Adaptada (116).

En la reparació de TDC mitjançant la hibridació de cadena dependent de síntesi, el model és idèntic a l'anterior fins al pas en que l'extrem 3' invasor estén la cadena hibridada per l'acció de la polimerasa. Un cop estesa, la cadena és separada de la cadena motlle homòloga per una helicasa. Si la cadena alliberada es troba amb la seva cadena germana s'hi podrà tornar a hibridar gràcies a l'extensa complementarietat de bases. En tal cas, les dues cadenes acabarien d'estendre's per acció de les polimerases i posteriorment quedarien lligades, completant la reparació sense creuament genètic (Figura 1-9B).

b. Replicació induïda per trencament

La replicació induïda per trencament té lloc quan l'helicasa replicativa es troba un tall a la cadena

de DNA i la forca de replicació sencera col·lapsa. El mecanisme es considera com una variant del model de reparació per hibridació de cadena dependent de síntesi (Figura 1-11).

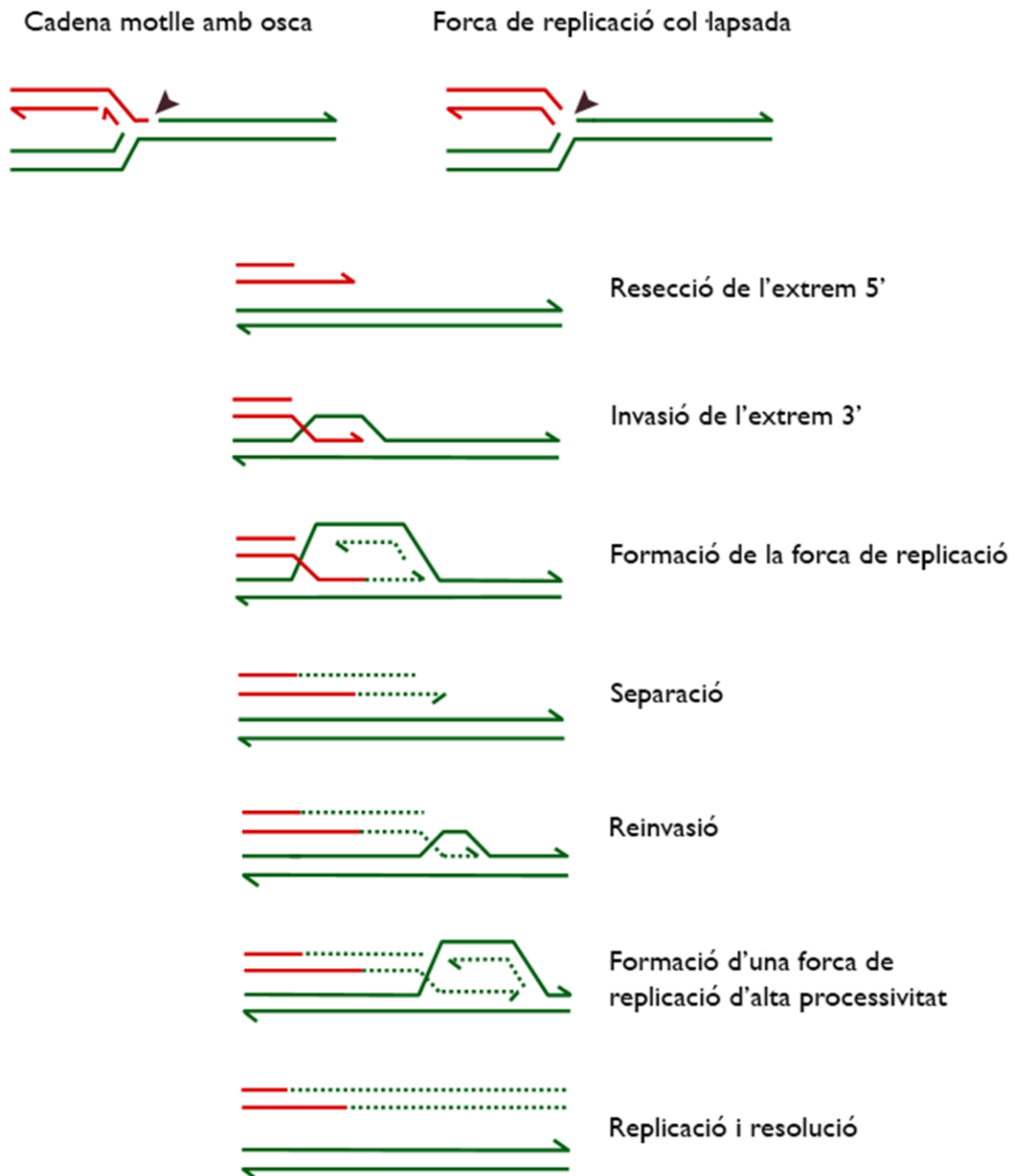


Figura 1-11 | Mecanisme de replicació induïda per trencament. Adaptada (106).

De la mateixa manera que als mecanismes anteriors, quan col·lapsa la forca es processa l'extrem 5' de les cadenes on hi ha el tall, i l'extrem 3' envaeix una regió homòloga –teòricament la seqüència de la cromàtide germana–. L'extrem invasor fa la funció d'encebador i estén la cadena per acció d'una polimerasa de baixa processivitat –capaç d'addicionar pocs nucleòtids de manera seqüencial–, incloent tant la cadena acceptora com la donadora (117). La processivitat va lligada a la velocitat de

polimerització; si un sistema és lent acaba desestabilitzant-se, ja que fa falta molta energia per mantenir-lo. Quan l'extrem 3' s'estén i es separa del motlle, si no és capaç de trobar la seva cadena complementària amb la que hibridar, aquest pot tornar a envair una regió amb homologia i continuar estenent-se en forques de baixa processivitat. Aquest procés d'invasió, extensió i separació pot repetir-se diverses vegades fins que s'acabi formant una forca de major processivitat, moment en el que es completarà la replicació fins al final de la molècula (118).

c. Recombinació homòloga no al·lèlica

La recombinació homòloga no al·lèlica (RHNA) és causada per l'alineament de dues seqüències de DNA paràlogues –que comparteixen una alta similitud de seqüència, però es troben en punts del genoma– amb resultat de creuament genètic (119).

Si durant el procés de reparació per recombinació s'utilitza com a regió homòloga una repetició orientada en el mateix sentit de la cadena, el creuament genètic resulta en productes recíprocament duplicats o deletats del fragment de DNA que es trobi flanquejat per les repeticions (Figura 1-12/A(i)). Al seu torn, les repeticions invertides generen inversions de l'interval genòmic flanquejat per les repeticions. Si les repeticions són a cromosomes diferents, el resultat són translocacions cromosòmiques (119,120). Un cop es resol el reordenament, les cadenes segreguen en la pròxima divisió cel·lular, alterant així l'organització genètica de les cèl·lules filles.

La RHNA també pot ocórrer durant el procés de replicació induïda per trencament, quan la cadena de DNA trencada utilitza una homologia no al·lèlica (o ectòpica) per restaurar la forca de replicació. En funció de la tria de fragment homòleg es generen duplicacions o deletions (Figura 1-12/A(ii)). Un altre mecanisme en el que es pot donar la RHNA és durant el procés d'hibridació de cadena simple, utilitzat per la reparació de TDC. Si durant la resecció dels extrems 5' de les dues cadenes no s'inicia la invasió d'un extrem 3' a una altra cadena que presenti homologia, les exonucleases continuaran el processament al llarg de la cadena. És possible que durant el procés es reveli una seqüència complementària, fent que les cadenes puguin hibridar. L'eliminació dels extrems sobrants, l'extensió del DNA i la lligació completaran la reparació del TDC, però amb la deleció de la seqüència de DNA entre les seqüències repetides (Figura 1-12/B).

Si la RHNA succeeix durant la meiosi, amb un creuament genètic desigual, resulta en reorganitzacions genòmiques constitutives que poden ser polimorfismes benignes o causar desordres genòmics tant esporàdics (si són *de novo*) com familiars –heretables–(85,121,122). Per altra banda, si té lloc durant la mitosi, el resultat són poblacions de cèl·lules somàtiques portadores de variants estructurals entre cèl·lules normals –els anomenats mosaics–(123,124).

Tot i les mesures de seguretat cel·lulars per evitar la RHNA explicades anteriorment, al llarg del genoma hi han regions enriquides amb LCRs o duplicacions segmentàries adjacents, amb presència de subseqüències en tàndem i altres en orientació reversa (119,125). Aquestes regions, estructuralment complexes, amb llargades que poden superar les 10 Kb i amb identitats de seqüència properes al 95-97%, actuen com a substrat típic per events de RHA (126,127) i RHNA (119,128). Algunes inclús es solapen (129-131), i fins i tot tenen una seqüència signatura (o *motif*) d'actuació en *cis* (per exemple, la seqüència 'CCNCCNTNNCCNC') que hi apareix enriquida (132). L'alteració en el número de còpia de certes regions del codi genètic genera una inestabilitat genòmica que predisposa a l'aparició de reordenaments genòmics a gran escala (133-136).

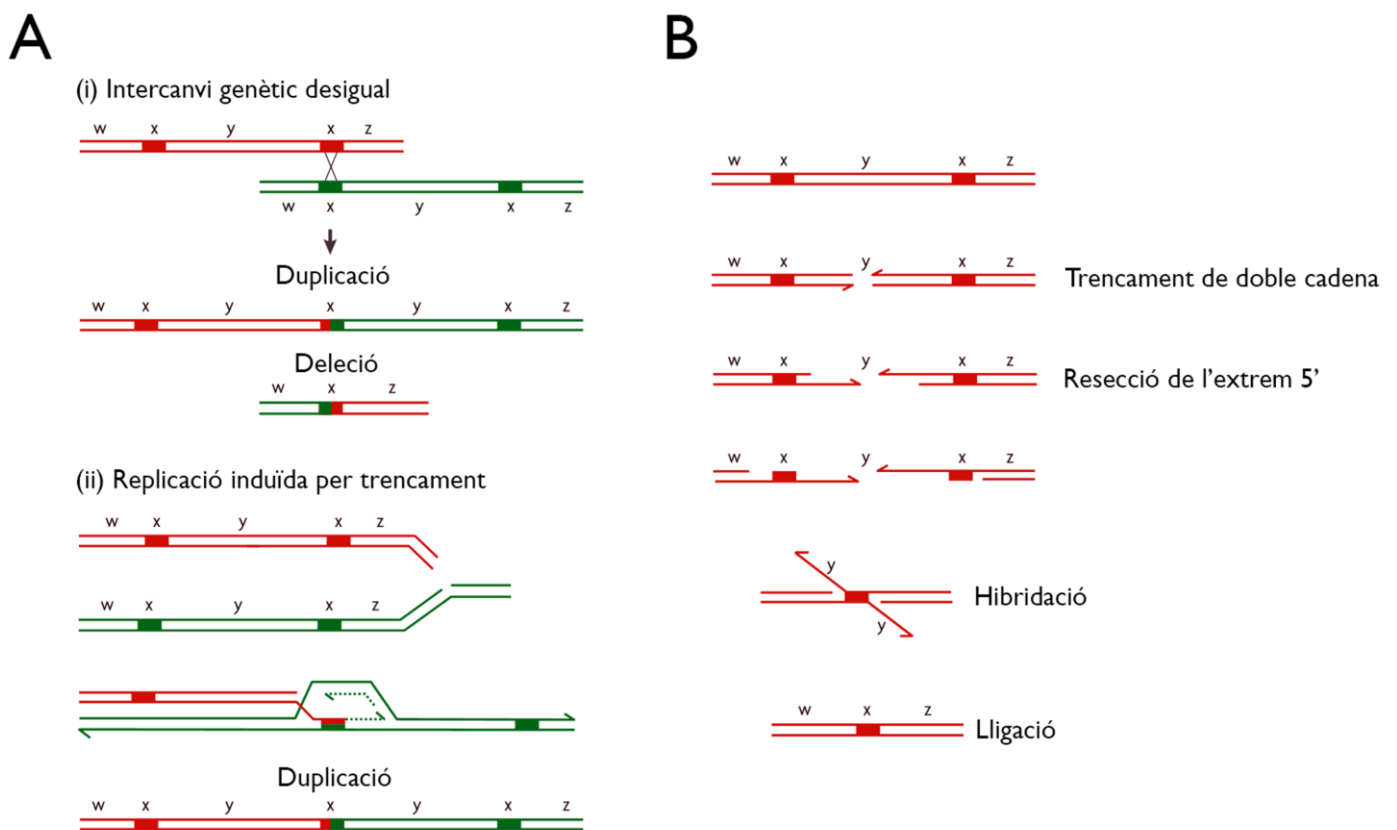


Figura I-12 | Mecanismes de formació de creuaments genètics: **A)** RHNA; **B)** Hibridació de cadena única. Adaptada (106).

Les regions d'elevada complexitat, els elements genètics mòbils, com els retrotransposons LINE (137), les seqüències *Alu* (138-140) o els pseudogens coincidents (141,142) també actuen com a substrat per la RHNA sempre i quan presentin una identitat de seqüència suficientment alta com per facilitar la recombinació homòloga. Aquests fragments de DNA són considerablement més curts (137,143) i sovint

són subestimats pels assaigs de detecció de CNVs (144,145).

II – Mecanismes de reparació no homòloga

A banda dels mecanismes de reparació basats en la recombinació homòloga, també n'hi han d'altres que o bé no requereixen d'homologia, o bé en necessiten una de molt modesta. Quan l'homologia no és requerida per assegurar que les cadenes de DNA són reparades en la posició correcta, sempre existeix la probabilitat de que succeeixi un reordenament genòmic. Els mecanismes de reparació de TDC per recombinació no homòloga són la unió d'extrems no homòlegs (UENH) i el cicle de trencament-pont-fusió (70,146–148).

També hi han evidències que suggereixen l'existència de mecanismes de formació de mutacions estructurals que no necessiten homologia de seqüència i que actuen durant el procés de replicació del DNA, com el mecanisme de lliscament durant la replicació, el mecanisme d'estabilització de forca i canvi de motlle –FoSTeS, de l'anglès *Fork stalling and template switching*– (149) i el de replicació induïda per trencament mediat per microhomologia (150).

a. Unió d'extrems no homòlegs

La UENH és el mecanisme de reparació de TDC predominant i àmpliament conservat en la gran majoria d'organismes. Té dos trets característics que el defineixen: el primer és la independència d'un substrat amb homologia extensa; el segon és que a diferència de la majoria de mecanismes de reparació, deixa imperfeccions als llocs de reparació, a mode de 'cicatris' informatives. L'aspecte positiu del mecanisme, característica que el fa el predominant, és la rapidesa amb la que actua, deguda a la gran densitat de proteïnes Ku, que inicien el procés de reparació. Així, la integritat estructural del cromosoma és restaurada de manera molt eficient. Si el sistema fos més lent, la lesió al DNA podria traduir-se en la pèrdua de centenars de gens d'un mateix segment cromosòmic (151).

El mecanisme s'inicia pel reconeixement de la lesió i la unió de la proteïna Ku als extrems del DNA afectat. Ku és un heterodímer de Ku70 i Ku80, que formen el component d'unió a DNA de la proteïna cinasa dependent de DNA (DNA-PK). L'heterodímer forma un anell que envolta el dúplex de DNA, actuant com a suport estructural i d'alineament pels extrems del DNA i protegint-los de la degradació i de la unió promiscua a regions irrellevants. A més, la proteïna permet l'accés de les polimerases, nucleases i lligases als extrems trencats per promoure la unió dels extrems. Quan Ku és al seu lloc d'acció, recluta la subunitat catalítica de DNA-PK, l'anomenada DNA-PKcs, una proteïna serina/treonina cinasa amb funcions de senyalització en resposta a l'estrès cel·lular. La DNA-PKcs fosforila diverses proteïnes nuclears, incloent el complex de reparació XRCC4/DNA lligasa IV (Figura 1-13A). La UENH requereix que els extrems del DNA siguin roms per efectuar la lligació, per tant, a la

segona etapa del procés de reparació es requereix l'acció de nucleases i de polimerases per tallar o afegir nucleòtids als extrems. La funció de les nucleases en la UENH és poc coneguda; la DNA-PKcs es fosforila i s'uneix a la proteïna Artemisa, una metal·lo-beta-lactamasa. Es creu que el complex resultant (Artemisa/DNA-PKcs) té activitat nucleasa i que, ocasionalment, talla les solapes dels extrems (152). Per altra banda, són les polimerases de la família X, Pol λ i Pol μ , les encarregades d'omplir els forats restants als extrems (153). És en aquest punt del procés en el que es genera el rastre característic de nucleòtids. El tercer i últim pas del mecanisme és la lligació dels extrems. Quan aquests s'han processat, el complex Ku70/80 recluta XRCC4, que forma un complex molt fort amb el factor Cernunnos tipus XRCC4 (Cer-XLF). Aleshores la DNA lligasa IV s'acobla al complex, s'uneix als extrems de DNA i catalitza la formació de l'enllaç fosfodièster, gràcies a la funció estabilitzadora i estimuladora de l'activitat lligasa de XRCC4 (154).

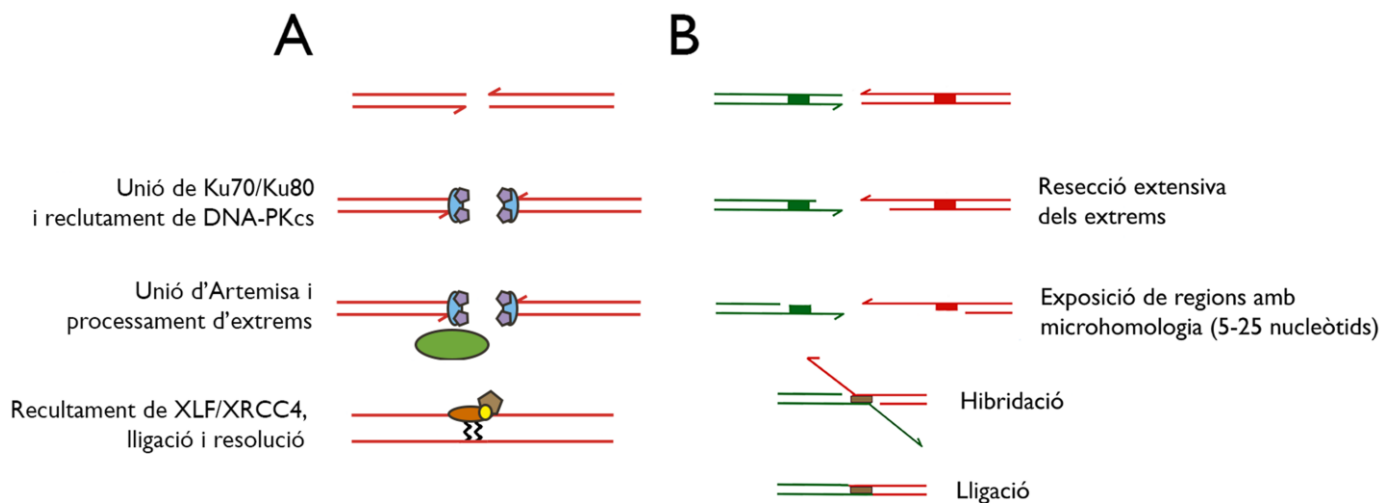


Figura I-13 | A) Mecanisme d'UENH; **B)** UENH mediada per homologia. Adaptada (155).

La tria entre la UENH i la recombinació homòloga per la reparació de TDC és regulada pel primer pas del mecanisme recombinatori. En aquest és necessària la resecció de l'extrem 5' per disposar d'una cua llarga de cadena única que pugui unir-se a una regió homòloga. Els TDC que no hagin estat processats podran ajuntar-se per UENH, però la resecció compromet la regió a ser reparada pel mètode recombinatori (156). La UENH és activa a través de tot el cicle cel·lular, però és molt més rellevant durant la fase G1, quan no hi han motlles de DNA homòlegs disponibles per recombinar. Aquesta regulació s'aconsegueix per la cinasa dependent de ciclina Cdk1, que és desactivada en G1 i expressada en S i G2. Cdk1 fosforila la nucleasa Sae2, permetent l'inici de la resecció (157).

Els punts de trencament dels reordenaments generats per la UENH sovint són localitzats entre elements repetitius del DNA com VNTRs, LINEs o seqüències *Alu* (158). A més, a les proximitats de moltes d'aquestes unions es troben motius de seqüència (per exemple 'TTTAAA') coneguts per la capacitat de causar TDC i curvatures al DNA (158,159). Això suggereix que, encara que no hi hagi un

requeriment d'homologia de substrat, la UENH és promoguda per una certa arquitectura genòmica (160).

S'han reportat i relacionat amb aquest sistema de reparació algunes reorganitzacions genòmiques aberrants, com translocacions importants i fusions telomèriques –segell distintiu de les cèl·lules tumorals– (161).

La unió d'extrems intervinguda per microhomologia és un mecanisme de reparació de TDC alternatiu a la UENH (Figura 1-13/B). La característica més destacable del mecanisme és l'ús de seqüències de microhomologia –de 5 a 25 pb– per l'alineament dels extrems, resultant sempre en petites delecions al DNA que flanqueja el trencament original (147). És considerat un mecanisme amb una alta tendència a cometre errors; sovint resulta en la generació d'anomalies cromosòmiques, com delecions, translocacions, inversions i reordenaments complexes que propicien la carcinogènesi mitjançant la creació d'oncogens. Degut al desavantatge que comporta la introducció de delecions dins del codi genètic, la cèl·lula en regula el seu ús, i tant sols l'executa a la fase S del cicle cel·lular, quan no queda més remei (ja que no pot utilitzar la UENH –fase G1– o la recombinació homòloga –fase S tardana, cap a G2– (147).

b. Cicle de trencament – pont – fusió

Després d'una replicació incompleta, per exemple a causa de l'estrès cel·lular provocat per la inhibició de la replicació del DNA (162,163), un cromosoma pot perdre un dels seus telòmers per un TDC. En tal cas, després de la replicació hi hauran dues cromàtides germanes a les quals els hi faltará un telòmer. El descobriment de formacions cromosòmiques aberrants observables al microscopi va permetre proposar un model de formació de cromosomes dicèntrics (Figura 1-14) a partir de la fusió dels extrems sense telòmer de dues cromàtides germanes (164). Segons aquest model, després de la fusió i durant l'anafase, els dos centròmers són empesos cap a nuclèols separats, adquirint una forma de pont característica. En algun moment, degut a la tracció de la maquinària de replicació cel·lular, el pont es trenca aleatòriament, generant grans duplicacions invertides en algunes cromàtides. Després del trencament, als dos cromosomes resultants els hi torna a faltar un telòmer, podent tornar-se a fusionar altre cop i establint així el cicle, que es repetirà fins que el cromosoma adquireixi un telòmer provinent d'una altra font. Aquest mecanisme juga un paper important en l'amplificació de segments genòmics durant l'aparició i desenvolupament del càncer (148,165).

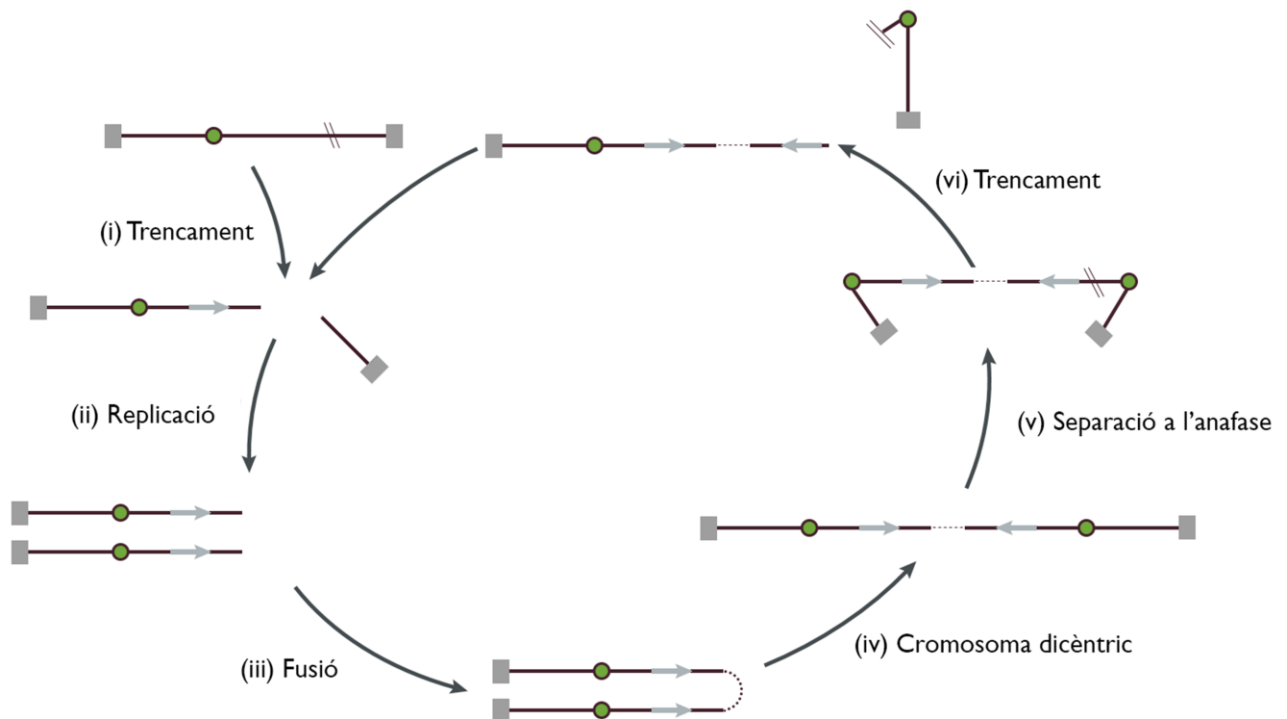


Figura 1-14 | Cicle de trencament – fusió – pont. Adaptada (106).

c. Lliscament durant la replicació

Aquest mecanisme mutagènic acostuma a donar-se en regions complexes del genoma, amb presència de repeticions en tàndem o de microsatèl·lits. En funció de la cadena afectada en la que hi hagi el lliscament, el resultat del mecanisme serà una duplicació o una deleció. Si la cadena de nova síntesi llisca cap a 5' formant una estructura secundària en forma de bucle, al tractar-se d'una regió repetitiva, els nucleòtids de la seqüència flanquejant romandran ben aparellats, mantenint l'estabilitat. Si l'estructura secundària no es resol mitjançant l'acció de nucleases, la DNA polimerasa tornarà a sintetitzar un fragment de DNA que ja havia sintetitzat prèviament. Quan la cèl·lula continuï amb la replicació i les dues cadenes inicials es separin, la nova cadena resultarà amb una duplicació de la seqüència inclosa al bucle. Al contrari, si és la cadena motlle la que forma el bucle, la DNA polimerasa s'alliberarà temporalment d'aquesta, podent tornar a acoblar-se al motlle i obviant la seqüència inclosa al bucle, generant una deleció (166) (Figura 1-15/A). Degut a la limitació en la mida dels bucles, es creu que el mecanisme opera dins de forques de replicació d'una llargada aproximada a 2 Kb. Per aquest motiu no es relaciona amb la majoria d'esdeveniments que provoquen canvis en el número de còpies en humans, ja que les llargades de les seqüències implicades acostumen a ser de diversos milers fins a milions de bases (106).

Algunes malalties humanes han estat associades a l'expansió de repeticions del genoma, provocades pel mecanisme de lliscament durant la replicació. La malaltia de Huntington és un cas paradigmàtic: els errors durant la replicació causen l'expansió de triplets de nucleòtids al gen *HTT*,

codificant per la proteïna Huntingtina. El resultat és una proteïna disfuncional, causant de la malaltia. Mentre que un individu normal presenta de 6 a 35 repeticions, en un afectat se'n poden identificar entre 36 i 121 (167).

d. FoSTeS

A partir dels estudis sobre els errors en l'amplificació dels gens de l'operó *lac* induïts per estrès a *Escherichia coli* (168,169) i mitjançant tècniques d'anàlisi genòmica d'alta resolució, Slack et al. van proposar un model de mecanisme mutagènic pel qual, durant la replicació, la forca pot establir-se permetent a la cadena en síntesi diversos canvis de motlle on continuar la replicació (170). La hipòtesi que va portar a la proposta del model va plantejar-se degut a que la llargada mitja de les seqüències amplificades era de 20 Kb, massa elevada com per tenir lloc en una única forca de replicació (149).

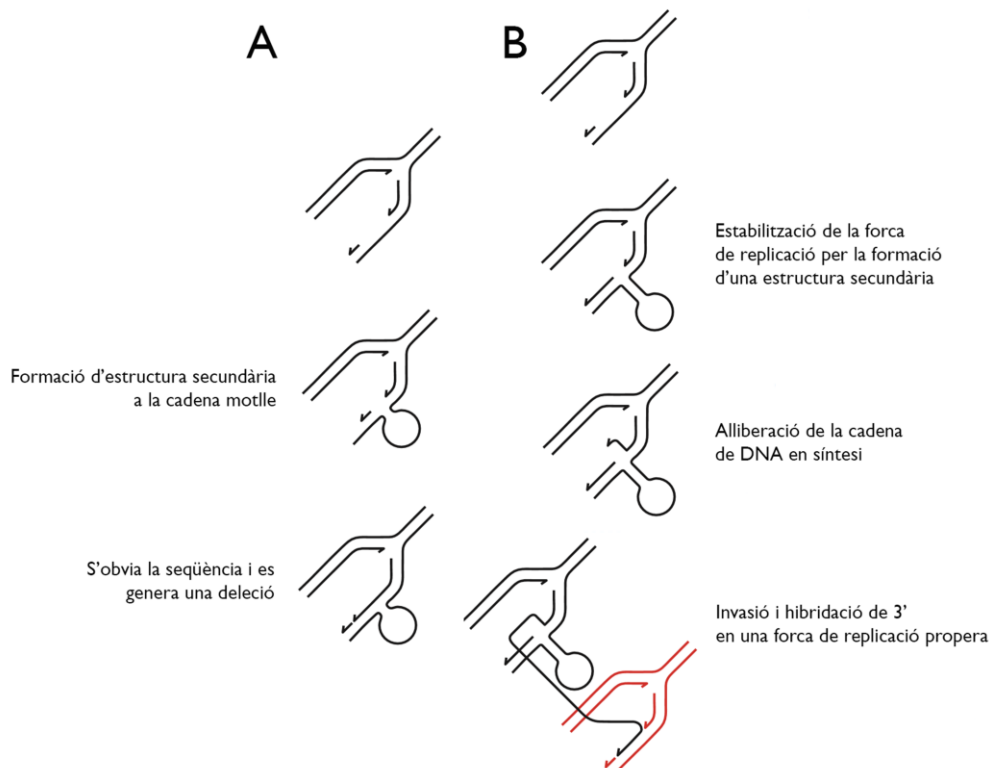


Figura 1-15 | Esquema del mecanisme de **A)** Lliscament durant la replicació i **B)** FoSTeS. Adaptada (106,171).

D'acord amb aquest model (Figura 1-15/B), una forca de replicació de DNA pot establir-se per l'adquisició d'una estructura secundària que bloquegi el seu progrés, o col·lapsar-se per la presència de fractures a la cadena motlle. L'extrem 3' de la cadena en síntesi s'allibera del motlle original i hibrida en una altra cadena motlle exposada per una forca de replicació on podrà continuar la síntesi del DNA.

Aquesta forca no ha de ser necessàriament adjacent a l'original, però sí que ha de ser propera en un context espacial –en tres dimensions– i presentar microhomologia (menys de 20 nucleòtids). Depenent de la direcció de la progressió de la forca i de quina sigui la cadena que s'utilitzi com a motlle en la nova localització, el fragment incorporat de manera errònia pot acabar en orientació inversa a la seva posició original. A més, en funció d'on sigui la nova forca (a *upstream* o *downstream* de l'original), es generarà una duplicació o una deleció, respectivament.

Aquest procés de desacoblament, invasió i síntesi pot ocórrer múltiples vegades en sèrie, depenent de la baixa processivitat de la polimerasa de DNA involucrada, i resultant en reorganitzacions cromosòmiques realment complexes (170).

e. Replicació induïda per trencament intervinguda per microhomologia

En base a les observacions prèvies sobre el mecanisme FoSTeS (170), Hastings et al. van proposar un model replicatiu generalitzat de canvi de motlle de DNA durant la replicació en cèl·lules sotmeses a estrès replicatiu i amb característiques del model de replicació induïda per trencament (exposat anteriorment). Aquest podria ser el causant de la formació d'un percentatge generós de les variants estructurals als genomes de tots els dominis de la vida (150). La replicació induïda per trencament intervinguda per microhomologia no tant sols pot induir la formació de CNVs, sinó que també crea regions de LCRs proveïdores de l'homologia necessària per la RHNA, predisposant a l'aparició de reordenaments genòmics en les generacions futures. El mecanisme pot causar mutacions estructurals somàtiques associades a processos cancerígens i és darrere dels reordenaments genòmics associats a l'emergència dels trets específics dels primats (150).

Segons el model, la cadena d'una forca de replicació col·lapsada –en absència de suficient proteïna recombinasa per revertir la situació– pot formar una nova forca de replicació per microhomologia, requerint tant sols entre 2 i 5 nucleòtids. L'extrem estès pot dissociar-se repetidament i tornar a formar forques de replicació de baixa processivitat en una àmplia varietat de motlles. Eventualment, el canvi portarà la cadena a la cromàtide germana original i formarà una forca de replicació d'alta processivitat, on podrà completar la replicació. El producte final contindrà la seqüència de les diferents regions genòmiques on hagi hibridat la cadena en síntesi. Si el retorn a la cromàtide germana succeeix *upstream* o *downstream* de la posició de col·lapse original determinarà si hi ha una deleció o una duplicació (Figura 1-16).

III – Retrotransposició

El tercer mecanisme principal de formació de variants estructurals és la retrotransposició. Els elements retrotransposables LINE comprenen aproximadament el 20.4% del genoma humà, i són els

únics retrotransposons inserits al genoma que continuen actius (172). De les aproximadament 510.000 còpies de LINE1 (un tipus de LINE, també anomenat L1) presents al nostre genoma, tant sols uns 128 L1 són de llargada complerta (al voltant de les 6 Kb) i tenen dos marcs oberts de lectura (ORFs, de l'anglès *Open Reading Frames*) intactes: ORF1 codifica per una proteïna d'unió a RNA i ORF2 codifica per una proteïna amb activitat endonucleasa i transcriptasa reversa (86,173,174).

La transposició té lloc via un intermediari de DNA, transcrit per l'RNA polimerasa II (173). La transcripció reversa i la integració es creu que tenen lloc en un procés doble anomenat 'transcripció reversa dirigida per encebador' (175). La inserció resultant és flanquejada per dues seqüències anomenades 'llocs diana duplicats', característiques del procés.

Certes característiques dels elements retrotransposables, com la gran abundància al genoma, l'alta identitat de seqüència amb altres regions i l'habilitat de moviment, els fan els majors contribuïdors a la inestabilitat genòmica (176,177). Els L1 són responsables de la mobilització d'elements *Alu*, *SVA* i de retrogens (172,173,178); sovint s'han identificat com la causa d'una àmplia varietat de desordres genètics, com l'hemofília, la hipercolesterolèmia familiar o el càncer de còlon i de mama (179).

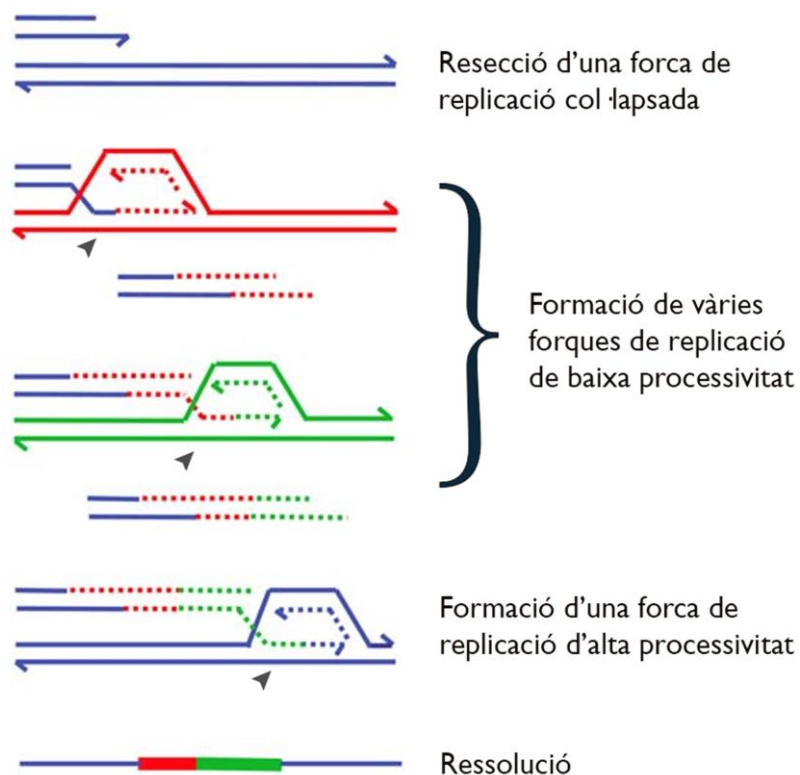


Figura I-16 | Mecanisme de replicació induïda per trencament intervinguda per microhomologia. Adaptada (150).

1.2.3 – Mecanismes moleculars de transmissió de fenotips clínics mitjançant CNVs

Els mecanismes moleculars pels quals les CNVs poden donar lloc a fenotips clínics són sis (69,121). El primer és el dosatge gènic (Figura 1-17/A). Si la CNV afecta un gen sensible a dosi, l'alteració de la quantitat total de proteïna expressada pot causar efectes patològics per haploinsuficiència (180) o per triplosensibilitat (119). Haghghi et al. van publicar un cas de miocardiopatia hipertròfica causada per l'haploinsuficiència del gen *PLN*, codificant per una proteïna relacionada amb el metabolisme del calci al reticle sarcoplasmàtic del múscul cardíac (181). La interrupció gènica (Figura 1-17/B) esdevé quan el punt de trencament d'una deleció, inserció o duplicació en tàndem es localitza enmig de la seqüència codificant d'un gen i causa una pèrdua de funció per inactivació. Aquest és el mecanisme fisiopatològic que causa, per exemple, la deuteranòpsia –la manca de sensibilitat en la percepció del color verd (182)–. Al seu torn, la fusió gènica (Figura 1-17/C) causada pel reordenament de la seqüència entre dos gens diferents pot generar una mutació de guany de funció. Aquest és el mecanisme predominant entre els càncers associats amb translocacions cromosòmiques somàtiques. Un exemple de fusió gènica associada a un fenotip clínic és el de la hipertensió i l'aldosteronisme remeiable per glucocorticoides; els gens codificants per l'aldosterona sintasa (*CYP11B2*) i l'esteroid 11 beta hidroxilasa (*CYP11B1*) tenen una identitat de seqüència del 95% i s'ha comprovat que la fusió dels dos gens (causada per RHNA) segrega entre els pacients (183). També pot ser que la CNV aparegui en una localització específica del genoma en la que hi hagi algun element de regulació important (Figura 1-17/D), com llocs de *splicing*, o que alteri la regulació de la regió i dels gens pròxims (179). A l'estudi de Velagaleti et al. es reporten dues translocacions amb punts de trencament a aproximadament 900 Kb *upstream* i 1.3 Mb *downstream* que causen displàsia campomèlica a causa de la desregulació del gen *SOX9* (184). Altres estudis, com el de Lupiáñez et al. reporten la implicació dels reordenaments genòmics en la desregulació dels dominis topològics (regions de regulació local del genoma acotades i separades per proteïnes) causants de malformacions a les extremitats de diversos organismes (185). Un altre mecanisme és la promoció d'al·lels recessius o variants polimòrfiques funcionals per deleccions i duplicacions al·lèliques (Figura 1-17/E). Per exemple, l'activitat del factor 12 (FXII) present al plasma sanguini en pacients amb la deleció típica de la síndrome de Sotos és predominantment determinada pel polimorfisme funcional de l'al·lel hemizigot romanent (186). Finalment, l'últim mecanisme descrit fins a data d'avui és l'efecte potencial de les CNVs en els processos de transvecció (Figura 1-17/F). La transvecció és un fenomen epigenètic de regulació a distància, per aparellament d'al·lels de cromosomes homòlegs, mitjançant el qual es pot activar o reprimir un gen (187). Si un dels al·lels implicats resulta deleccionat, la regulació de la regió pot veure's afectada. Estudiant models de ratolí, Yan et al. van relacionar la penetrància de les anomalies craniofacials amb una seqüència de 590 Kb de seqüència genòmica propera al gen *RAI1*, en la qual podien tenir lloc transveccions o altres factors reguladors a distància (188).

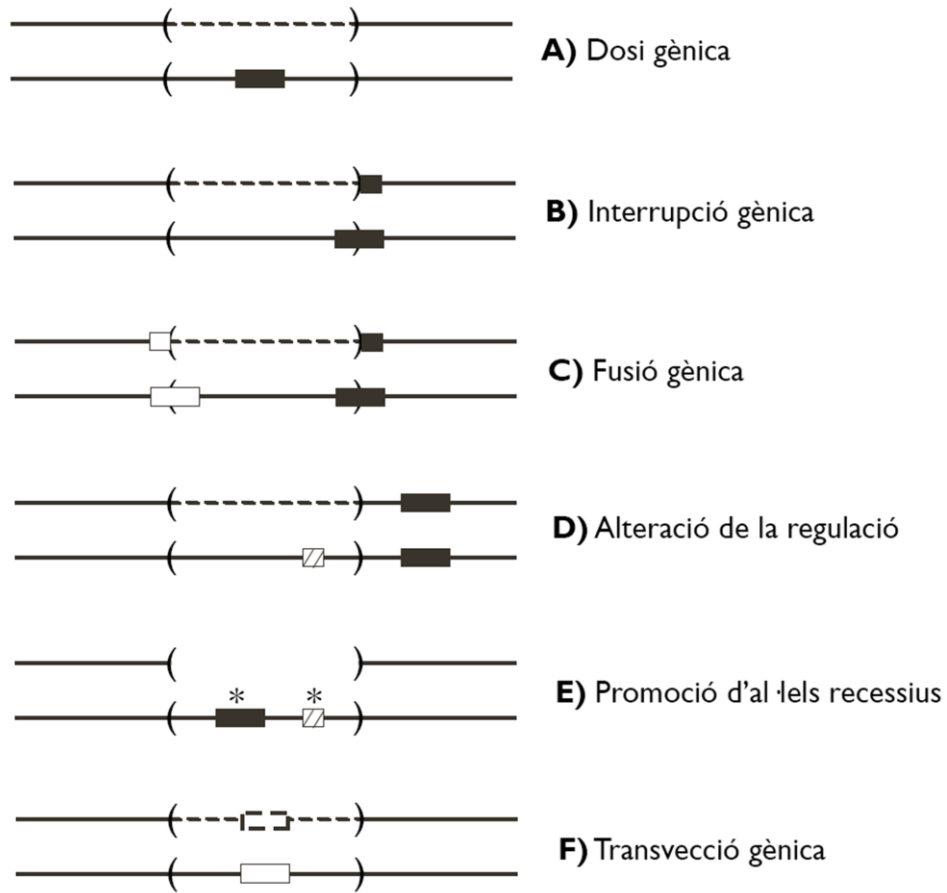


Figura I-17 | Transmissió de fenotips clínics mitjançant CNVs. Adaptada (121).

1.3 - La biologia computacional i la bioinformàtica

La bioinformàtica és la disciplina que persegueix la solució de problemes biològics mitjançant l'ús de mètodes estadístics, eines computacionals, el maneig i l'explotació de bases de dades i el disseny de *pipelines* d'anàlisi –seqüències d'instruccions informàtiques que s'executen de manera simultània–. Tot i que diferenciada, la bioinformàtica és una disciplina germana de la biologia computacional; es pot considerar que és la implementació a la pràctica d'aquesta. Les dues disciplines han evolucionat en paral·lel, l'una gràcies a l'altra. Sense la biologia computacional, i sense la publicació de nous algorismes i avenços tècnics, la bioinformàtica s'estancaria; per altra banda, sense la bioinformàtica, la biologia computacional seria irrellevant (en un sentit estrictament pragmàtic). Tot i que no hi ha cap norma escrita, el perfil típic d'un bioinformàtic acostuma a ser el d'algú format essencialment en biologia i estadística, i que ha desenvolupat les habilitats necessàries de programació per escriure codi i *pipelines* d'anàlisi que incloguin l'ús de *softwares* que, sovint, hauran desenvolupat els biòlegs computacionals. Aquests, al seu torn, s'han format principalment en enginyeria, matemàtiques, algorítmica o disciplines similars. Probablement, a mesura que apareguin noves tècniques computacionals i experimentals, les connotacions associades als dos perfils aniran canviant i les diferències entre les dues disciplines siguin cada cop menys profundes. Això s'està veient ja actualment, en un moment en el que la bioinformàtica s'interessa cada cop més per models probabilístics que requereixen de planificació i arquitectura informàtica. Un bon exemple són els models de *Deep Learning* basats en xarxes neuronals (189,190).

La biologia computacional emergeix a l'inici dels anys seixanta gràcies a l'accés de la comunitat científica a ordinadors digitals d'alta velocitat (desenvolupats pels programes armamentistes de la segona guerra mundial) i l'expansió i l'interès creixent en l'obtenció de col·leccions de seqüències d'aminoàcids (191). Des dels seus inicis, els interessos de la biologia computacional han estat molt diversos, englobant camps tant distants com la informàtica estructural (cristal·lografia de raigs X, microscòpia electrònica o ressonància magnètica nuclear) el modelatge i la dinàmica macromolecular, la teoria computacional, la creació i definició de gramàtiques de llenguatges de programació i un immens etcètera. Per aquesta tesi són d'especial rellevància, sobretot, els estudis relatius a l'anàlisi i l'alineament de seqüències de DNA i RNA, el desenvolupament de bases de dades moleculars, la predicció de l'estructura de macromolècules (amb un èmfasi especial en l'estructura de les proteïnes) i els estudis d'evolució filogenètica i macromolecular (192).

La bioinformàtica d'avui en dia es sustenta en els fonaments intel·lectuals i tècnics establerts pels científics dels inicis de l'era de la computació (191). A continuació es presenta una selecció d'alguns dels estudis pioners més rellevants, àmpliament revisats a Ouzounis & Valencia (192): en el camp de l'evolució es publica la teoria sobre l'origen de la vida i l'evolució humana a partir dels events de duplicacions gèniques (193,194); les primeres anàlisis filogenètiques de famílies de macromolècules (195–197) i l'evolució dels procarïotes, amb la identificació dels *Archaea* com a nou domini independent de la vida (198). Es publica el dogma central de la biologia, després dels estudis seminals dels processos

de transcripció i traducció del RNA (199); apareixen els primers mètodes d'alineament i de comparació de seqüències (200–202) i es publica la hipòtesi d'ús preferencial de codó, formulada en base a estudis computacionals (203). En el camp de les proteïnes, i amb l'objectiu de desxifrar el que es considerava el 'segon codi genètic' –codificat per les seqüències d'aminoàcids–, apareixen els primers estudis de descripció, visualització, anàlisi i predicció d'estructures secundàries. En aquests es deriven les preferències dels residus d'aminoàcids per la formació d'estructures secundàries (204,205); es duen a terme càlculs d'accessibilitat dels solvents a les estructures proteiques (206) i es representen per primera vegada les hèlix alfa (207).

A partir d'aquests estudis fundacionals, el ritme de creixement, de publicacions i d'innovació en el camp és frenètic. Durant la dècada dels vuitanta es desenvolupen algoritmes clau, com el de Smith-Waterman (208,209) i els algoritmes de la família FASTA per cerques a bases de dades (210). També s'identifiquen els primers motius de seqüències en diverses proteïnes funcionals, com els motius d'unió a ATP (211) i els motius *zinc-finger* (212); es publiquen les seqüències consens de les regions ORF (213,214) i es fan les primeres anàlisis de plegament del RNA (215,216). Es deslloriguen les bases del plegament proteic i es publiquen una gran quantitat d'estudis descriptius amb els principis estructurals de les proteïnes. Alguns estudis paradigmàtics són els dels ponts disulfur (217), les làmines beta (218) i els patrons d'empaquetament en estructures helicoidals (219). El RNA ribosòmic s'utilitza com a marcador filogenètic per primera vegada (220,221), s'aprofundeix en els mecanismes i les dinàmiques metabòliques del DNA (222) i es publiquen estudis sobre l'evolució de l'*splicing* (223), dels exons (224), dels introns (225) i dels retrovirus (226). Durant aquest període té lloc la primera fase de desenvolupament tècnic i logístic per la construcció de grans bases de dades amb control de qualitat de la informació rebuda (227,228); apareixen dos grans recursos per l'enviament de dades nucleotídiques que, com ja s'ha comentat, resultarien vitals pel Projecte Genoma Humà, *GenBank* (229) i *EMBL Data Library* (230). Durant la dècada dels noranta, amb l'automatització de la seqüenciació del DNA i l'accés lliure a internet, la biologia computacional i la bioinformàtica viuen un període d'intens desenvolupament. Tot i la gran heterogeneïtat de sistemes operatius, en el món acadèmic hi abunden les terminals Unix i Macintosh. Apareixen llenguatges inspirats en la utilitat *awk* d'Unix, com *Python* o *Perl* (192) –el llenguatge utilitzat en el desenvolupament de l'algoritme de detecció de CNVs presentat en aquesta tesi–. Es desenvolupa l'algoritme BLAST d'alineament de seqüències (231) i els primers programes sofisticats per la predicció de gens (232,233).

Durant l'última dècada, amb la finalització dels projectes genòmics de finals del segle XX i l'arribada de les plataformes de seqüenciació d'alt rendiment, el desenvolupament d'eines i mètodes d'anàlisi de grans conjunts de dades biològiques ha viscut un creixement exponencial. Aquest ha estat impulsat per la connectivitat moderna, els centres de supercomputació i la capacitat millorada dels ordinadors personals. Tot i la notorietat i la rellevància científica actual de la biologia computacional i de la bioinformàtica, encara queden grans reptes per assolir. És necessària la unificació de formats i de bases de dades, per facilitar i millorar el coneixement comú, fent-lo accessible per tothom. Amb la

millora de les bases de dades, gràcies a iniciatives com el projecte *Platinum Genomes* (234), i amb el creuament de grans sets de dades complementàries (de projectes genòmics, transcriptòmics i epigenètics) s'avançarà en el camí cap a donar resposta a la demanda actual d'informació sobre la rellevància del gran número de Variants de Significat Incert –VSI–, identificades en multitud de pacients. Es podrà aprofundir en els complexos entramats de la regulació gènica, i s'obrirà camí cap a una millora de la diagnosi genètica, que tingui en compte el *background* genòmic complet de l'individu. Altres problemes importants de caire logístic, com l'emmagatzematge i la seguretat de les enormes quantitats d'informació generades diàriament per projectes individuals de grups de recerca modestos, queden relaxats amb l'aparició de solucions d'emmagatzematge al núvol, sempre i quan es disposi d'una bona connectivitat al lloc de treball.

1.3.1 – La bioinformàtica en l'anàlisi de dades de seqüenciació d'alt rendiment

En l'actualitat, la bioinformàtica es centra principalment en l'anàlisi de conjunts de dades biològiques provinents d'experiments de seqüenciació d'alt rendiment. Aquestes dades són la matèria primera de les anomenades tecnologies “-òmiques”, sufix que significa el mesurament de la col·lecció completa de les molècules d'informació biològica d'un organisme. Els principals camps d'investigació són els que involucren els tres tipus de molècules del dogma central de la biologia (199). Així doncs, existeix la genòmica, o l'estudi quantitatiu i qualitatiu del DNA i dels genomes –o subconjunts de gens del mateix–; la transcriptòmica, l'estudi dels trànscrips, com l'RNA missatger, els micro RNAs o l'RNA ribosòmic; i la proteòmica, l'estudi de l'abundància de les proteïnes i les seves propietats físico-químiques. Aquests camps d'investigació han propiciat l'aparició d'altres relacionats. A la xarxa pot trobar-se una llarguíssima llista (235), però citant alguns exemples tenim l'epigenòmica, l'estudi del conjunt de modificacions epigenètiques (com les modificacions d'histones i el seu efecte sobre la seqüència del DNA (236)); la farmacogenòmica, l'estudi de l'efecte de les variacions genòmiques en la resposta als fàrmacs (237); i la nutrigenòmica, l'estudi de la interacció dels diferents aliments amb el genoma, incrementant el risc a patir malalties comunes cròniques (238).

L'aparició de les tecnologies “-òmiques” representa un canvi en el paradigma de la biologia molecular. Abans de l'entrada en escena de les plataformes d'alt rendiment, les dades provinents d'assajos de biologia molecular eren escasses, ja que la manipulació i el tractament de les mostres era molt costós. Per aquest motiu, l'anàlisi i la validació de les dades era relativament senzill. En contrast, aquestes han permès incrementar diversos ordres de magnitud la relació cost-eficiència en la producció de dades, resultant en quantitats ingents d'informació que sovint són sorolloses –necessiten passar per controls de qualitat i processos de normalització–. Aquestes dades no poden ser organitzades, emmagatzemades i analitzades sense un marc d'anàlisi bioinformàtic dissenyat acuradament per assegurar-ne la interpretació biològica adequada. S'ha de tenir en compte que algunes de les aplicacions de l'anàlisi bioinformàtica són de naturalesa sensible (com ho és el tema que ocupa aquesta

tesi). Per exemple, la detecció de mutacions patogèniques en desordres genètics heretables, el cribratge de familiars i la diagnosi de malalties complexes, tot encarat al desenvolupament de mètodes diagnòstics i/o terapèutics que facin possible i millorin la medicina genòmica personalitzada.

Donat que les dades encara sense processar (les seqüències provinents de les plataformes de seqüenciació) sovint es presenten en format fastq, els primers passos de control i processament de les diferents anàlisis acostumen a ser comuns. Els processos de control de qualitat de les seqüències i l'etapa d'alineament són bons exemples. En el primer s'eliminen aquelles seqüències (o fragments de seqüència) que hagin acumulat errors, o que s'hagin seqüenciat amb una qualitat inferior als estàndards esperats; de no fer-ho, els resultats posteriors podrien veure's alterats i perdre fiabilitat. Aquesta tasca, com de fet la gran majoria de totes les altres, pot realitzar-se mitjançant el disseny i l'execució de *scripts* propis, fets a mida per les característiques de cada experiment, o mitjançant les eines bioinformàtiques publicades per la comunitat científica a tal efecte (239). Al seu torn, a l'etapa d'alineament es comparen les seqüències a un genoma de referència. Si aquesta referència no existeix prèviament, s'haurà de realitzar un alineament *de novo* de les seqüències obtingudes mitjançant un *software* dissenyat específicament per aquest objectiu. De manera intuïtiva, resulta evident que la tria del *software* d'alineament no és una decisió trivial, sinó que s'ha de decidir en funció de certs factors, com per exemple: la naturalesa de la seqüència en sí, si ha estat sintetitzada amb el mètode del bisulfit (mitjançant la tecnologia SOLiDTM); la llargada de les seqüències (que pot oscil·lar entre les 16 bases – en el cas dels micro RNAs processats– fins a diversos milers de bases si són generades amb plataformes de NGS) o si les seqüències són úniques o aparellades (*paired-end*). També serà d'utilitat conèixer si l'algoritme utilitza els paràmetres de qualitat (de base i d'alineament), o com gestiona les seqüències *multimap* –aquelles que alineen a més d'una regió genòmica– i les variants o els errors presents a les seqüències.

En el procés de tria és vital no perdre de vista quin és l'objectiu últim de l'estudi, ja que per una anàlisi epigenòmica o transcriptòmica les seqüències amb bases no coincidents a la referència han de ser descartades; però per un estudi de genotipat aquesta és una informació absolutament essencial sense la qual l'anàlisi no té cap mena de recorregut. De manera més secundària, els requeriments de velocitat o de *hardware* de l'algoritme poden acabar d'ajudar a prendre la decisió. Aquests factors són rellevants, per exemple, en el disseny de *pipelines* per anàlisis rutinàries de laboratoris clínics, on pot haver un gran volum de feina.

Per les següents etapes d'anàlisi, en funció de quin experiment s'estigui duent a terme, seran necessaris *softwares* específics per les diferents tasques. Alguns dels exemples més representatius són: els estudis genòmics de genotipat de mostres problema, en els que tant la identificació com l'anotació de variants s'haurà de dur a terme amb els anomenats *variant callers*, i altres *softwares* específics per l'explotació dels repositoris públics de dades genètiques poblacionals, o de predictors de patogenicitat –els predictors *in silico*–; els estudis de transcriptòmica (RNA-seq), en els que s'hauran d'identificar i

quantificar els trànscrips seqüenciats, descobrir-ne de nous i analitzar l'expressió diferencial dels gens de l'organisme a estudi (240); els estudis epigenòmics (ChIP-seq), en els que els passos d'anàlisi més habituals són la identificació dels pics d'acumulació de seqüències complementàries al DNA –amb els anomenats *peak callers*–, i l'anàlisi de motius de seqüències d'unió a DNA (241).

Entre les pràctiques habituals de la biologia computacional i de la bioinformàtica s'inclou la de posar a disposició dels usuaris en repositoris *on-line* les eines desenvolupades per tasques d'anàlisi concretes. Aquestes eines acostumen a alliberar-se a la xarxa de manera gratuïta i l'usuari, sota la seva responsabilitat, és lliure de modificar-ne el codi de la manera que més li convingui pels seus objectius. La comunitat també demostra una alta activitat en fòrums web, on s'acostuma a donar suport en la resolució de problemes als usuaris que puguin necessitar-ho (242). La velocitat en el desenvolupament d'eines noves és molt elevada i les llistes de *software* que es proposin en paper no trigaran gaire temps a quedar obsoletes. Tot i així, a la xarxa poden trobar-se revisions completes i en constant actualització de tots els algorismes d'anàlisi publicats fins la data, en les que s'especifica per a quins processos han estat dissenyades i les seves característiques tècniques (243).

1.3.2 – Detecció de variants estructurals

I – Mètodes tradicionals

Els primers mètodes de detecció de variants estructurals van aparèixer amb l'emergència de la citogenètica humana. Al 1959, Lejeune et al. van associar, a partir de l'anàlisi visual del cariotip humà, la trisomia del cromosoma 21 amb la Síndrome de Down (244). El descobriment d'aquesta anomalia cromosòmica va precedir-ne altres també associades amb fenotips clínics, com les síndromes de Turner (45, X), de Klinefelter (47, XXY), de Patau (trisomia del cromosoma 13) i d'Edwards (trisomia del cromosoma 18). A l'any 1963, Lejeune et al. descobreixen la primera síndrome delecional, la de *Cri-du-Chat*, caracteritzada per la deleció del braç curt del cromosoma 5 (245–249).

Amb el desenvolupament de les tècniques de bandejat cromosòmic de principis dels anys setanta (Figura 1-18/A) es van poder detectar un gran nombre d'anomalies cromosòmiques, de manera que en poc temps el cariotip va utilitzar-se en la diagnosi clínica rutinària d'individus amb retard mental i/o malformacions congènites, alteracions del desenvolupament sexual i altres (250,251). El poder de resolució del cariotip convencional va millorar amb el pas del temps, fins arribar a l'actual, d'entre 5-10 Mb.

La citogenètica tradicional, tot i resultar útil per la diagnosi d'anomalies cromosòmiques, presentava certes limitacions com per exemple la impossibilitat de detectar anomalies cromosòmiques submicroscòpiques (o críptiques) inferiors a 5 Mb. A més, els mosaics inferiors al 14% podien passar fàcilment desapercibuts. La tècnica és molt laboriosa i requereix cultiu cel·lular (252,253). Aquestes

limitacions van posar de manifest la necessitat de mètodes més resolutius i implementables en la rutina de diagnosi clínica.

Per intentar donar solució a aquestes necessitats, durant la dècada dels vuitanta es van desenvolupar les noves tècniques de citogenètica molecular. La primera en veure la llum va ser la Hibridació *In Situ* Fluorescent, o FISH (de l'anglès *Fluorescent In Situ Hybridization* –Figura 1-18/B–). Aquesta es basa en la utilització de sondes de DNA marcades i dissenyades per hibridar a regions cromosòmiques en metafase o nuclis en interfase (254). La tècnica facilitava la caracterització de les reorganitzacions gràcies a la vistositat del marcatge, però la millora en la resolució no era significativa. A més, la tècnica tant sols permet estudiar regions genòmiques prèviament seleccionades, la quantitat de loci sotmesos a estudi en cada assaig és molt baixa (de 2 a 3, generalment) i la identificació de duplicacions en tàndem no és possible.

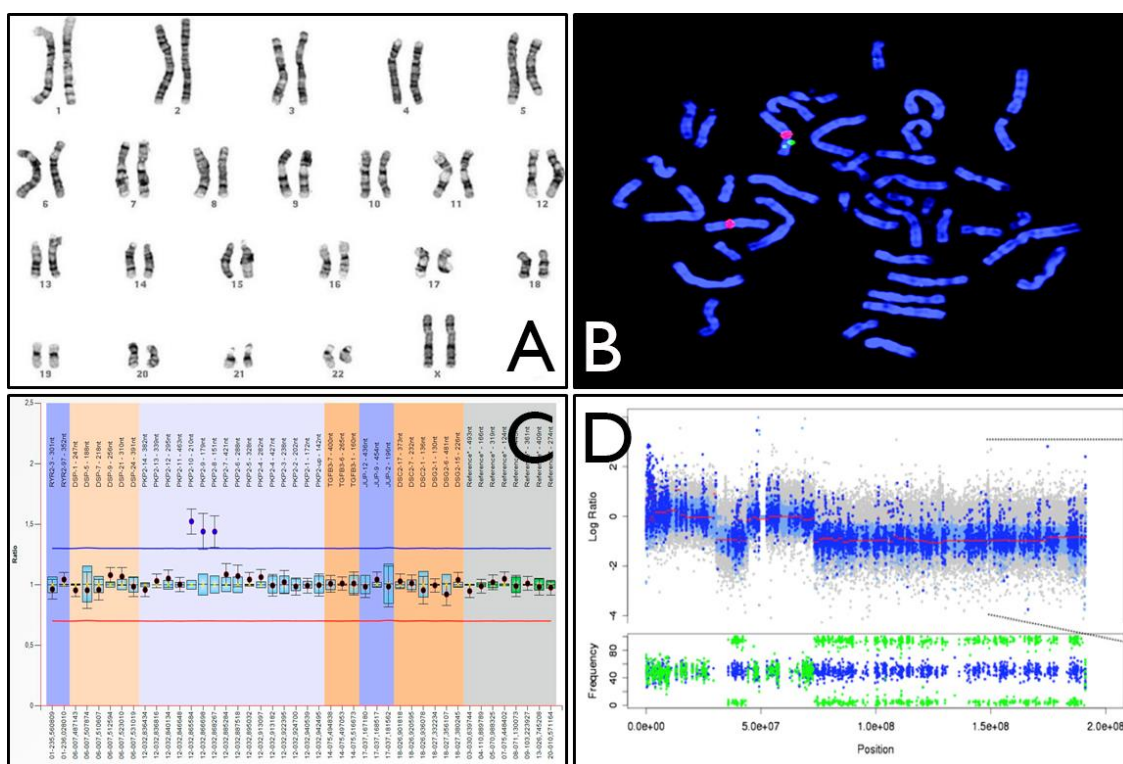


Figura I-18 | A) Cariotip tradicional; **B)** Cariotip marcat amb FISH; **C)** Duplicació de 3 exons a *PKP2* per MLPA; **D)** Detecció de CNVs mitjançant un *array* de SNPs. Adaptada (255,256).

Amb l'objectiu d'analitzar tot el genoma sense restriccions, als anys noranta es van desenvolupar les noves tècniques basades en FISH, com l'*Spectral Karyotyping*, la Multiplex-FISH i la Hibridació Genòmica Comparada (de l'anglès *Comparative Genomic Hybridization* –CGH–). Les dues primeres permeten la visualització simultània de tots els cromosomes en colors diferents, facilitant la visualització de les reorganitzacions. La CGH, per altra banda, es basa en la cohibridació d'un DNA problema i un DNA control (cada un marcat amb un fluoròfor diferent, usualment verd i vermell) sobre cromosomes

normals en metafase. Segons la diferència en la mesura de les intensitats de les fluorescències s'infereixen els desequilibris cromosòmics al llarg de tot el genoma (257–259). Aquestes tècniques, tot i millorar en certs aspectes la citogenètica convencional (sobretot en la visualització dels resultats), no van arribar a implantar-se mai en el camp diagnòstic perquè continuaven sent massa laborioses i per presentar resolucions molt similars a les de la citogenètica convencionals.

En canvi, el desenvolupament de les tècniques de PCR quantitativa (qPCR, també anomenada PCR en temps real) i de MLPA (de l'anglès *Multiplex Ligation-dependent Probe Amplification*) sí que va proveir a la comunitat medicocientífica amb dues eines d'elevada resolució, automatitzables i econòmiques, capaces de detectar CNVs a escala exòmica (Figura 1-18/C). La primera es basa en el monitoratge de l'amplificació d'una diana de DNA mitjançant l'ús d'una sonda marcada amb un fluoròfor reporter que emet la fluorescència després de la hibridació amb la cadena complementària (260); la segona es basa en l'amplificació dels loci diana mitjançant una PCR múltiple, la seqüenciació dels resultats per electroforesi capil·lar i la comparació del patró de pics obtingut per la mostra problema amb les mostres de referència, permetent el cribatge de fins a 50 regions genòmiques per assaig (261). Les dues tècniques continuen presentant certes limitacions: tant sols poden estudiar-se regions prèviament seleccionades, no detecten reorganitzacions cromosòmiques equilibrades i requereixen un DNA de bona qualitat, ja que, al cap i a la fi, ha d'estar en condicions per poder hibridar de manera adequada (262,263). En l'actualitat, tant la qPCR com la MLPA continuen sent tècniques *gold standard* per la validació de variants estructurals.

II – Mètodes bioinformàtics

La tècnica de citogenètica molecular més popular durant la dècada passada va ser l'anàlisi cromosòmica basada en *microarrays* (*Chromosomal Microarray-based Analysis*, CMA), ja que permetia la detecció de CNVs al llarg de tot el genoma a una resolució molt superior a la del cariotip –1 Kb–, es tractava d'una tècnica automàtica i no requeria de cultiu cel·lular previ. La utilització de la CMA va propiciar una acceleració sense precedents en el descobriment de les causes genètiques de desordres esporàdics, la descripció de noves síndromes microdelecionals i microduplicacionals, la caracterització millorada de les ja conegudes, i el descobriment de gens associats a anomalies congènites aïllades (264).

Inicialment, la CMA va ser dissenyada com un *array* de baixa resolució per la detecció de CNVs en tumors (265) i utilitzava o bé sondes de tipus BAC (de l'anglès *Bacterial Artificial Chromosomes*) o bé sets de menys de 100.000 oligonucleòtids. La tecnologia no va trigar a optimitzar-se per la detecció de reorganitzacions cromosòmiques en desequilibri en individus amb anomalies congènites, una informació crucial per l'establiment de correlacions genotip-fenotip, la determinació del pronòstic d'anomalies cromosòmiques i pel procés d'assessorament genètic (266). Utilitzant sets de més d'un

milió d'oligonucleòtids (267), es poden detectar aberracions genètiques acotables a un únic gen, incrementant dràsticament la sensibilitat d'aquesta tècnica (268) i fent que ofereixi un potencial diagnòstic molt superior al del cariotip convencional en pacients amb retard en el desenvolupament, desordres de l'espectre autista i/o anomalies congènites múltiples d'origen desconegut (15-20% contra un 3% si s'exclouen la Síndrome de Down i altres síndromes cromosòmiques similars). Per aquests motius, la CMA va passar a ser tècnica de primera elecció per l'estudi d'aquests pacients (269).

Existeixen dos tipus de plataformes de *microarrays*: les d'hibridació genòmica comparada –*array* CGH, o aCGH– i les basades en l'estudi de polimorfismes genòmics –*array* de SNPs–. En l'aCGH, es marca el DNA del pacient i un DNA de referència amb diferents fluorocroms (Cy5 i Cy3) i es cohibriden sobre una matriu de seqüències conegudes d'oligonucleòtids de DNA. Es pot escollir entre tres dissenys de plataformes d'aCGH, les de genoma complet, que com el seu nom indica permeten la detecció de CNVs al llarg de tot el genoma a una elevada resolució; les dirigides, que es centren en l'estudi de regions amb repercussió fenotípica coneguda (270) i les que combinen les dues possibilitats. Al seu torn, les plataformes d'*arrays* de SNPs (Figura 1-18/D) permeten la detecció de CNVs al llarg de tot el genoma mitjançant l'anàlisi de la pèrdua d'heterozigotitat dels SNPs inclosos. Únicament requereixen el DNA del pacient per hibridar-lo sobre el *microarray*, i els resultats obtinguts es comparen amb els d'un control. A diferència de l'aCGH, els *arrays* de SNPs presenten una major sensibilitat en la detecció de mosaics de baixa proporció (271).

Degut a la gran quantitat i a la naturalesa de les dades obtingudes en aquests assajos, la utilització de *software* de processament dels resultats és essencial. Per aquestes tècniques s'acostuma a fer servir el *software* proporcionat per la mateixa casa comercial a la que es compra el kit de reactius (272).

Amb l'arribada de les tecnologies de seqüenciació d'alt rendiment, la possibilitat de detectar variants estructurals de manera indirecta amb els resultats de la seqüenciació va sembrar el terreny per les primeres publicacions d'eines de detecció de CNVs. Per algunes patologies concretes, amb una alta incidència de variants estructurals en regions restringides del genoma, aquestes són detectades de manera rutinària mitjançant kits de MLPA o assaigs de qPCR. Però amb el temps, la tendència a utilitzar les dades de seqüenciació com a primera tècnica de detecció de CNVs i servir-se de la CMA com a mètode de validació ha anat incrementant, sent el procés normal avui en dia (273).

Els *softwares* de detecció de CNVs a partir de dades provinents de seqüenciació d'alt rendiment –llistat exhaustiu a (243)– cobreixen principalment tres aproximacions analítiques: l'estimació del número de còpia dels segments genòmics a partir de dades de cobertura (274–276), la incorporació de la informació extra que aporten les seqüències *paired-end* per millorar la precisió de la detecció (277) i la utilització de la informació de seqüències *split-read* per la localització de punts de trencament amb una resolució exacta (278,279).

a. Detecció basada en la comparació de cobertures

Aquesta aproximació resulta particularment útil per la detecció de CNVs en exomes o en experiments de seqüenciació dirigida a regions d'interès, ja que no es basa en l'anàlisi de les seqüències solapants als punts de trencament, elements amb una probabilitat molt baixa de ser capturats per les sondes del disseny. Generalitzant, la idea es basa en la comparació del número de seqüències alineades en finestres genòmiques amb el número esperat sota un model estadístic determinat. Els valors desviats informen de la presència de variants estructurals desequilibrades, però les equilibrades passen desapercibudes.

De manera similar a la metodologia utilitzada en els assajos d'aCGH, la ràtio del comptatge de lectures entre la mostra problema i una referència es prefereix a l'anàlisi d'una única mostra, per poder controlar així l'extensa variabilitat típica observable en l'eficiència d'hibridació de les sondes al llarg de les regions (274–276). La naturalesa disseminada i la mida reduïda de les regions d'interès suposa un repte per la detecció de CNVs en tot un genoma sencer, inclús si la cobertura en aquestes regions és elevada (cosa poc probable). La poca cobertura que s'acostuma a assolir fa que la resolució de les dades sigui massa baixa per poder aplicar algorismes de segmentació o normalitzacions dels biaixos intrínsecs de la seqüenciació de manera raonable (280).

A l'inici d'aquesta tesi hi havien publicats diversos *softwares* de detecció de CNVs. No obstant, cap d'ells resultava una eina òptima pel cribratge de pacients en un context clínic (a l'apartat 4.2 de Resultats i Discussió s'aprofundeix en aquest punt). La majoria estaven específicament dissenyats per l'anàlisi d'exomes (274,275,281,282) i no atacaven certs problemes importants de les dades, com l'existència de covariables que afecten la comparació entre mostres. Per exemple, l'eficiència d'hibridació de les sondes, que disminueix a causa de les característiques locals de certes regions, com un contingut GC extrem, la discrepància en el comptatge de seqüències totals entre les mostres problema i les control, i les diferències en el percentatge de seqüències *on-target* (aquelles que alineen exclusivament a les regions d'interès). Tampoc permetien la construcció d'un pseudo-control amb el que comparar les mostres problema en absència de controls, punt que pot arribar a ser limitant en certs anàlisis o per segons quins usuaris. La versió inicial d'ExomeDepth (281) aplicava un model ocult de Markov per inferir les regions afectades per les variants estructurals en base a la probabilitat d'observació d'aquests events. Aquest tipus de models són útils per experiments amb una gran quantitat de valors, com és el cas d'un exoma, però resulten menys eficaços a mesura que hi han menys regions involucrades, com en el cas dels panells de captura. L'única eina per l'anàlisi de dades provinents de panells de captura, el CONTRA (280), en el moment de la seva publicació no tenia en compte els artefactes que podien generar-se a partir d'un disseny de sondes poc optimitzat, ni la valuosa informació que aporten les seqüències aparellades per la discriminació de falsos positius.

b. Detecció basada en les seqüències aparellades

Els algoritmes de detecció basats en la informació aportada per l'alineament de les seqüències estan orientats a experiments de seqüenciació de genoma sencer, ja que la probabilitat de capturar una regió solapant a un punt de trencament és molt més elevada, al seqüenciar el genoma com un continu. Aquests algoritmes permeten la detecció de variants estructurals desequilibrades (delecions i duplicacions) i equilibrades (insercions, inversions i translocacions).

La identificació de CNVs a partir de seqüències aparellades es basa en la detecció d'agrupacions de parelles de seqüències discordants, en les que la mida de l'insert que les separa sigui anormal – massa llarga o curta– o que ho siguin les orientacions de les dues seqüències (279) (Figura 1-19/A).

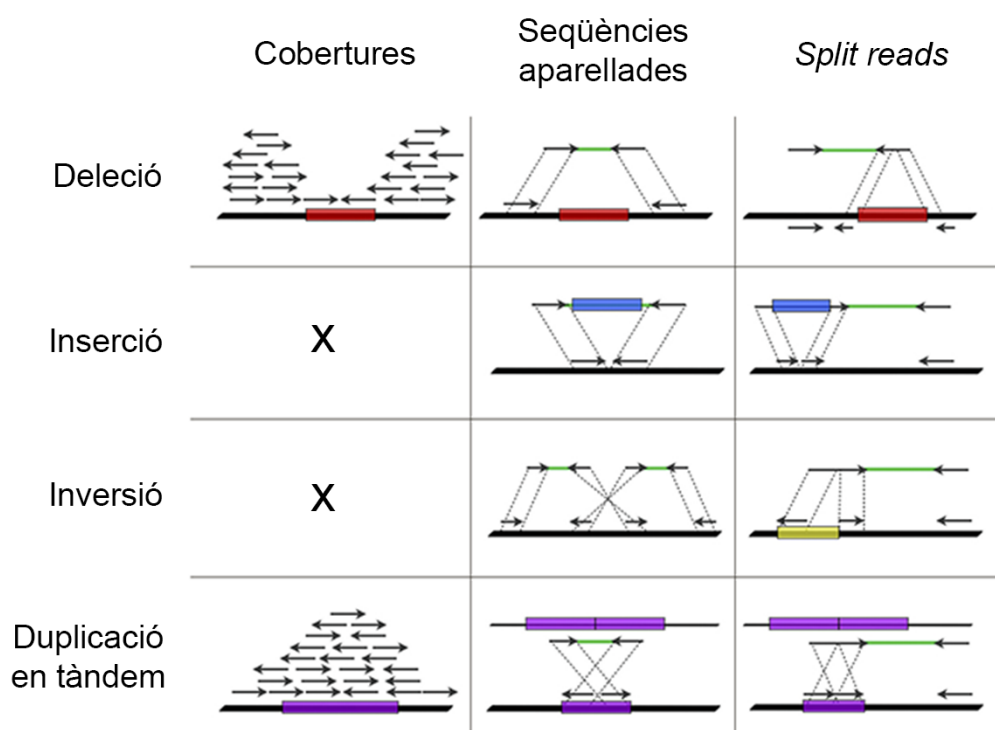


Figura 1-19 | Diferents aproximacions de detecció de CNVs mitjançant mètodes bioinformàtics: **A)** Comparació de cobertures; **B)** Informació de seqüències aparellades; **C)** Mitjançant *split-reads*. Adaptada (283).

c. Detecció basada en els alineaments parcials

Per altra banda hi han els algoritmes que detecten i analitzen les seqüències *split-read*, aquelles que solapen els punts de trencament i que apareixen parcialment alineades. En els extrems d'aquestes s'identifiquen subseqüències marcades amb senyals de no coincidència amb el genoma de referència, l'anomenada signatura de *soft-clipping* (Figura 1-19/C). Processant aquestes seqüències i realineant exclusivament la porció de *soft-clipping* es pot determinar la localització del punt de trencament amb una resolució exacta a nivell de nucleòtid (279,284).

1.4 - La Mort Sobtada Cardíaca

1.4.1 – Definició i epidemiologia

La Mort Sobtada (MS) és un episodi letal i inesperat, que té lloc en el marge d'una hora des de l'aparició dels primers símptomes. Pot manifestar-se com el primer símptoma d'una patologia cardíaca en un individu aparentment sa, asimptomàtic fins al moment de la MS i, per tant, sense un diagnòstic previ; però també pot esdevenir en un individu malalt, independentment de la història prèvia de malaltia cardiovascular (285,286).

Quan la mort és extrahospitalària, en la majoria de casos es requereix una autòpsia medico-legal que n'aclareixi les causes, ja que es considera sospitosa. Quan s'exclou la mort d'origen violent, aquesta es classifica com a mort natural i, en absència d'una diagnosi o causa un cop realitzada l'autòpsia és classificada com a Morts Sobtada Inexplicada (MSI). Quan l'autòpsia descarta les causes extracardíaques (les hemorràgies cerebrals internes o subaracnoïdes, l'anafilaxi, les hemorràgies agudes, el xoc sèptic...), la causa més plausible de la mort és la Mort Sobtada Cardíaca –MSC–. Si l'autòpsia descarta l'infart aquesta passa a ser catalogada com blanca i probablement es tracti d'un cas de MSC d'origen arritmogènic (287).

Pel desenvolupament de la MSC és necessari un substrat, la malaltia, que ja sigui per causa mecànica o elèctrica contribueix a un entorn arritmogènic; i un desencadenant, com l'exercici físic o l'estrès emocional, que provoqui una inestabilitat cardíaca que desencadeni l'arítmia fatal, dificultant el bombeig de sang oxigenada als òrgans. En alguns casos, però, pot mancar el desencadenant extern. És el cas de la MSC en repòs, durant la nit (288,289).

En l'actualitat, la MSC presenta una incidència d'entre 50 i 100 individus per cada 100.000 en població general, i és la responsable del 80-90% de totes les MS en adults. Fins a quatre milions de persones moren anualment en tot el món degut a la MSC (289,290), encara que els números poden variar en funció dels criteris d'inclusió (per exemple, el temps des del primer símptoma o la falta de testimonis) i la manera de comptabilitzar els casos, ja sigui a partir d'una anàlisi retrospectiva –a partir de certificats mèdics–, o prospectiva (291,292).

La causa més comuna de MSC en l'adult és la malaltia coronària (75-80%), però les malalties cardíques arritmogèniques hereditàries són la causa principal en la població jove, menor de 35 anys (Taula 1-1). Entre un 10 i un 15% de les MSC són degudes a miocardiopaties, que inclouen la Miocardiopatia Hipertròfica (MCH), la Miocardiopatia Dilatada (MCD), la Miocardiopatia Arritmogènica (MCA), la No Compactació del Ventricle Esquerre (NCVE), la Miocardiopatia Restrictiva (MCR) i les anomalies congènites (especialment de l'artèria coronària). Finalment, entre el 5 i el 10% restant de MSC són causades per una alteració en el funcionament elèctric del cor en absència d'un defecte estructural cardíac. Aquesta categoria inclou la Síndrome de Brugada (SBr), la Síndrome del QT Llarg (SQTL), la Síndrome del QT Curt (SQTC), la Taquicàrdia Ventricular Polimòrfica Catecolaminèrgica (TVPC) i la

Fibril·lació Auricular familiar (FA) (290,293). D'altra banda, els factors de risc associats a la MSC són la hipertensió arterial, la diabetis mellitus, la hipercolesterolèmia i l'obesitat.

La MS en infants majors d'un any i en joves adolescents representa un 10% de les morts en etapa pediàtrica, el que equival a 1-5 nens/joves per cada 100.000 a l'any. Tot i que una tercera part d'aquestes morts segueix com inexplicada després de l'autòpsia, la major part s'atribueixen a malalties cardíaques arritmogèniques hereditàries com les de l'adult, a defectes cardíacs congènits (per exemple anomalies de l'aorta) o a miocarditis (294,295).

Quan la MSI afecta a infants menors d'un any d'edat, aquesta és catalogada com a Mort Sobtada del Lactant (MSL) si l'autòpsia és inconclusiva i s'han descartat antecedents clínics de malalties cardíaques arritmogèniques tant del nadó com de la família. La MSL és la primera causa de mortalitat en nadons de menys d'un any en països industrialitzats, representant un 70-80% de les morts en etapes lactants, la major part abans dels 6 mesos de vida (296). Tot i que fins al moment s'ha plantejat un ampli ventall de mecanismes fisiopatològics per explicar aquestes morts, l'etiologia de la MSL continua sense ser del tot clara. És considerada com un desordre multifactorial, amb 3 grups de factors de risc ben establerts, com la vulnerabilitat de l'infant (prematuritat, infeccions), l'exposició a factors ambientals nocius, els defectes genètics associats a les malalties arritmogèniques heretables i un factor supressor extern, com dormir en una posició propensa (296–299).

1.4.2 – Les malalties cardíaques arritmogèniques

Les malalties no isquèmiques associades a la MSC són desordres poc freqüents entre la població general, amb unes incidències aproximades d'entre 5 i 10 individus per cada 10.000 (289). La majoria són considerades malalties d'herència mendeliana, tot i que amb la implantació de les tècniques de seqüenciament d'alt rendiment al llarg dels últims deu anys i amb la gran quantitat d'informació obtinguda després de seqüenciar un alt nombre de pacients, l'evidència suggereix un patró d'herència complex per algunes d'elles (300,301).

La gran majoria de malalties cardíaques arritmogèniques comparteixen característiques comunes, tant clíniques com genètiques, com són: **a)** l'heterogeneïtat genètica, quan mutacions en diferents gens poden causar la mateixa malaltia (per exemple, les mutacions al canal de potassi i al canal de sodi poden provocar SQTL); **b)** l'heterogeneïtat fenotípica, quan diferents mutacions al mateix gen poden donar lloc a malalties diferents. S'ha reportat una mutació a *SCN5A* com la causa genètica d'un SQTL, una SBr i una MCD (81); i mutacions a *SCN5A* poden causar SQTL, SBr, MCD i MCA en funció del tipus d'alteració que provoquin al canal iònic (302–305); **c)** l'expressivitat variable, per la qual els portadors de la mateixa mutació, inclús dins de la mateixa família, poden expressar diferents graus de severitat per un mateix fenotip, independentment de l'edat o el sexe; **d)** la penetrància incompleta, que succeeix quan portadors de la mateixa mutació, inclús dins de la mateixa família, presenten diferents

fenotips (per exemple, en famílies afectades per la SBr poden trobar-se portadors asimptomàtics amb electrocardiogrames normals, mentre que altres poden presentar elevació del segment ST i arrítmies ventriculars. En aquests casos clínics també és possible trobar-se amb fenocòpies, aquelles malalties que provoquen un mateix fenotip, però per les que l'origen genètic és totalment diferent (i, per tant, poden produir-se errors en la diagnosi). Un exemple paradigmàtic de fenocòpia és la malaltia de Fabry –un trastorn d'emmagatzematge lisosomal a causa de mutacions poc freqüents i recessives al gen *GLA* lligades al cromosoma X– amb la MCH, causada principalment per mutacions als gens codificants per proteïnes sarcomèriques (306).

Taula I-1 | Prevalença, percentatge de casos resolts i gens associats a les principals malalties cardíques arritmogèniques.

Malaltia	Prevalença	Causa genètica	Gens associats
MCH	1 : 500	32-63%	<i>ACTC1, ACTN2, ANKRD1, BAG3, CALM3, CALR3, CASQ2, CAV3, CSRP3, DMPK, FHL2, FLNA, FLNC, GLA, JPH2, LAMP2, LDB3, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYLK2, MYOZ2, MYPN, NEXN, PDLIM3, PLN, PRKAG2, RYR2, TCAP, TNNC1, TNNI3, TNNT2, TPM1, TRIM63, TTN, TTR, VCL</i>
MCD	1 : 2500	50-60%	<i>ABCC9, ACTC1, ACTN2, ANKRD1, BAG3, CAV3, CHRM2, CRYAB, CSRP3, CTF1, DES, DMD, DSC2, DSG2, DSP, EMD, FHL2, FKTN, FLNC, LAMA4, LAMP2, LDB3, LMNA, MYBPC3, MYH6, MYH7, MYPN, NEBL, NEXN, PKP2, PLN, RBM20, SCN5A, SDHA, SGCA, SGCD, TAZ, TCAP, TMPO, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL</i>
MCA	1 : 5000	75%	<i>CTNNA3, DES, DSC2, DSG2, DSP, FLNC, JUP, LMNA, PKP2, PLN, SCN5A, TGFB3, TMEM43, TTN</i>
NCVE	ND*	35-75%	<i>ACTC1, CASQ2, DTNA, LDB3, LMNA, MYBPC3, MYH7, NOTCH1, SCN5A, TAZ, TNNT2, TPM1, VCL</i>
MCR	ND*	ND*	<i>ACTC1, BAG3, DES, FLNC, MYBPC3, MYH7, TNNI3, TNNT2, TTN</i>
SBr	1-5 : 10000	30-35%	<i>ABCC9, CACNA1C, CACNA2D1, CACNB2, GPD1L, HCN4, KCND3, KCNE1L, KCNE3, KCNJ8, PKP2, RANGRF, SCN1B, SCN1Bβ, SCN2B, SCN3B, SCN5A, SCN10A, SLMAP, TRPM4</i>
SQTL	1 : 2000	75-85%	<i>AKAP9, ANK2, CACNA1C, CALM1, CALM2, CALM3, CAV3, KCNE1, KCNE2, KCNH2, KCNJ2,</i>

			<i>KCNJ3, KCNJ5, KCNQ1, NOS1AP, RYR2, SCN1Bβ, SCN4B, SCN5A, SCN10A, SLC8A1, SNTA1, TRDN</i>
SQTC	<1 : 10000	50-60%	<i>CACNA1C, CACNA2D1, CACNB2, KCNH2, KCNJ2, KCNQ1</i>
TVPC	1 : 10000	65%	<i>ANK2, CALM1, CALM2, CALM3, CASQ2, KCNJ2, RYR2, TRDN</i>
FA	1 : 100	30%	<i>ABCC9, ANK2, Cx43, GJA1, GJA5, KCNA5, KCND3, KCNE1, KCNE1L, KCNE2, KCNE3, KCNE4, KCNH2, KCNJ2, KCNJ8, KCNQ1, NPPA, NUP155, PITX2, RYR2, SCN1B, SCN1Bβ, SCN2B, SCN3B, SCN4B, SLN</i>
SM i DATA	1 : 5000	75-90%	<i>FBN1, FBN2, FLNA</i>

*ND: No disponible

I – La diagnosi genètica de les malalties cardiovasculars arritmogèniques

Els pacients afectats per miocardiopaties o canalopaties poden patir una MSC com a primer dels símptomes. El solapament fenotípic i de gens causals entre els diferents desordres dificulta tant la diagnosi com el consell genètic. Per tant, la identificació d'una mutació causal en un pacient és crucial per la confirmació del diagnòstic en els casos límit, pel maneig precoç dels familiars en situació de risc, per l'assessorament genètic i per evitar el seguiment innecessari dels familiars asimptomàtics no portadors (si es detecten VSI si que cal fer-ne el seguiment), punt que comporta l'estalvi de costos sanitaris importants (307,308).

Per totes aquestes raons, les directrius clíniques actuals recomanen el cribratge genètic en pacients que compleixin els criteris establerts pels diferents desordres, seqüenciant les regions codificants dels principals gens associats a cada una de les malalties a estudi, aquells que presenten una freqüència superior de mutacions reportades com a causals (308). Per exemple, en el cas de la SBr, es recomana la seqüenciació de *SCN5A*, el gen que codifica per la subunitat del canal de sodi cardíac Nav1.5 –tot i que s'han associat uns 20 gens més a la patologia–. De manera comparable, en el cas de la MCH, es recomana la seqüenciació de les regions codificants dels principals 5 gens sarcomèrics (*MYBPC3, MYH7, TNNT3, TNNT2* i *TPM1*), tot i que el número total de gens associats a la MCH –amb diferents graus d'evidència– siguin aproximadament uns 25. La situació és extrapolable a totes les malalties arritmogèniques associades a la MSC, ja siguin canalopaties o miocardiopaties.

Aquest conservadorisme aparent de les recomanacions clíniques es deu a la necessitat d'optimitzar i transformar la inversió destinada a la identificació de la causa genètica en un resultat útil derivat de la investigació. Aquest resultat del cribratge, el llistat de mutacions, serà resumit en un

informe que es farà arribar al cardiòleg i/o al forense per la presa de decisió clínica o diagnòstica. Per aquest motiu, la identificació de variants prèviament no reportades ha d'anar acompanyada d'evidència, quanta més millor, per ser associada a la malaltia del pacient en qüestió. Aquesta evidència s'obté mitjançant les prediccions de patogenicitat d'algoritmes *in silico*, o per l'associació de la variant a la malaltia, conclosa a partir d'estudis funcionals *in vitro* (en models cel·lulars) o *in vivo* (en models animals).

En aquest sentit, l'adveniment de les tecnologies de seqüenciació d'alt rendiment ha millorat indiscutiblement el camp de la diagnosi genètica. S'ha reduït el temps i la mà d'obra de personal qualificat capaç de dur a terme la tasca de seqüenciació, així com la inversió en reactius i en sous per seqüenciar les diferents peces a demanda del metge. Però per altra banda, han sorgit altres problemes a considerar. La quantitat d'informació capaç de derivar-se d'un experiment de seqüenciació d'alt rendiment és enorme (des de panells de gens fins a genomes complets), i la informació que es té de la majoria de variants identificades és més aviat poca. Si s'hagués de fer l'estudi funcional de totes les variants candidates a tenir un rol patogènic en la malaltia del pacient, la inversió econòmica en la diagnosi genètica seria màxima i els llargs temps d'espera (de varis mesos fins a anys) farien que el sistema fos insostenible. Aquest problema remarca la necessitat de trobar un equilibri entre la diagnosi clínica i la recerca. Habitualment aquest equilibri es troba amb la seqüenciació de panells de gens associats (i candidats) i amb l'anotació exhaustiva de tota la informació present a les bases de dades de les variants identificades; juntament amb uns criteris de classificació de variants prèviament establerts i comuns per la comunitat mèdica i científica.

Tot i les millores en la diagnosi genètica derivades del desenvolupament de les tecnologies de seqüenciació d'alt rendiment, el percentatge de casos que continuen sense una causa genètica de la malaltia després del cribratge de mutacions en els principals gens associats a les malalties arritmogèniques cardíques és encara molt alt (309–316). Aquests percentatges varien d'acord a com de complet sigui el cribratge en quant a número de gens revisats o a la quantitat de pacients inclosos en la cohort d'estudi. Els casos no resolts poden explicar-se per variants patogèniques en gens no associats a la malaltia en el moment del cribratge i, per tant, no inclosos a l'estudi, però també a partir de variants en regions reguladores (317), o per alteracions epigenètiques (318). Aquestes regions no s'inclouen a l'estudi genètic per motius de cost-efecte i per la complexitat en la interpretació i associació amb la malaltia.

Un altre tipus de variants potencialment patogèniques pels casos negatius són les variants estructurals, no detectables per la seqüenciació Sanger capil·lar convencional. En els últims deu anys, els científics han identificat un gran número de variants estructurals ubiqües al genoma humà, tant en població sana com en grups de malalts (267,319). D'entre els diferents tipus de variants estructurals, s'ha reportat evidència que recolza el rol de les CNVs en les malalties associades a la MSC, tot i que fins a data d'avui continua sent un camp relativament inexplorat. Falten estudis robusts, que incloguin grans

cohorts de pacients i un ampli ventall de gens en el cribratge. Fins la data, els estudis realitzats o bé han investigat un número molt reduït de gens, o s'han realitzat en famílies de malalts afectats, o en cohort realment petites.

II – Les miocardiopaties

Les miocardiopaties es caracteritzen per la presència d'una estructura anòmala de la paret cardíaca en absència de malaltia cardíaca isquèmica, que impedirà el correcte funcionament del cor (320). Les miocardiopaties representen del 10 al 15% del total de MSC. Afecten especialment alguns col·lectius de la població, com els atletes o esportistes d'elit, que a base de l'esforç continuat –i sempre amb una predisposició genètica prèvia– presenten una mortalitat superior (321,322). Les més freqüents són les que es detallen a continuació:

a. La Miocardiopatia Hipertròfica (MCH)

La MCH és una malaltia caracteritzada per la hipertròfia asimètrica del ventricle esquerre, amb característiques histològiques d'hipertròfia cel·lular, desordre miofibril·lar i fibrosi intersticial. És la malaltia cardiovascular arritmogènica més comuna, afectant 1 de cada 500 persones, principalment gent jove i sobretot a esportistes, individus en els que l'arrítmia pot ser desemmascarada amb més freqüència a causa de l'esforç, mentre que en un individu no esportista aquesta pot restar latent fins al moment de la MSC (323). Presenta una marcada variabilitat fenotípica, inclús dins la família, i penetrància incompleta. Les manifestacions clíniques oscil·len des de cursos clínics asimptomàtics fins la insuficiència cardíaca greu i la MSC. En esportistes és comú observar un cor d'aparença hipertròfica. No obstant, això acostuma a ser una adaptació a la resposta fisiològica aguda a la que és sotmès el cor durant l'activitat esportiva –i, per tant, és una manifestació no patològica–. És el que es coneix com la Síndrome del cor d'atleta, i no és considerada una malaltia cardíaca.

La MCH s'hereta com un tret autosòmic dominant en la majoria dels pacients adults que compleixen els criteris diagnòstics; i tot i que són poc comunes, hi han descrites mutacions *de novo*. En general, el cribratge genètic condueix a la identificació de mutacions causals en el 32-63% dels casos (309), depenent de les característiques clíniques dels pacients, el nombre de gens estudiats, i els criteris utilitzats per la classificació de variants (308,324). En la majoria dels casos, la MCH és causada per mutacions als gens que codifiquen proteïnes sarcomèriques (325–327). Entre ells, aproximadament el 85% de les mutacions es troben a *MYBPC3* i *MYH7*, el 10% a *TNNT2* i *TNNI3*, fins a un 2% a *TPM1*, i menys d'un 3% en altres gens codificants per proteïnes del sarcòmer (*MYL2*, *MYL3*, *ACTC1*, *TNNC1* i *FLNC*).

La primera CNV identificada en un pacient de MCH va reportar-se al 1992; consistia en una

deleció heterozigota de 2.4 Kb al gen *MYH7*, que incloïa l'últim exó codificant de la proteïna, l'exó 40, i la regió UTR (de l'anglès, *Untranslated Region*) associada. Va ser identificada per *southern blot* i es va considerar patogènica per ser capaç d'interferir en el correcte ensamblatge i en el bon funcionament del sarcòmer (328). L'any 2002 Jouven et al. publiquen un estudi sobre la relació entre la durada del QT i el gruix del múscul cardíac en MCH familiar; a l'estudi són inclosos dos pacients amb delecions a *MYBPC3*, una de l'exó 25 i l'altra del 33, però no es dona més informació dels casos i tampoc s'aprofundeix més en la relació d'aquests defectes en la MCH (329). Al 2009 es publica un article metodològic en el qual es proposa una variant del mètode de seqüenciació per captura. Un dels pacients seqüenciats era el portador d'una duplicació d'11 Kb heterozigota en tàndem de l'exó 13 al 27 de *MYBPC3*, juntament amb una deleció heterozigota de 215 pb a l'exó 29 del mateix gen (330). Des d'aleshores s'han publicat poques sèries estudiant la presència de CNVs en pacients de MCH, i la majoria d'estudis han investigat tant sols 1 o 2 gens. A Coto et al. s'investiga la regió codificant de *MYH7* en 150 pacients mitjançant PCR de llarg abast, però no van poder identificar cap CNV (331). En tres estudis independents es van dur a terme cribratges amb MLPA a *MYBPC3* (i en alguns casos a *TNNT2*), reportant una taxa de detecció de CNVs del 0% (0/108) (332), 1% (1/100) (333) i 0.5% (1/185), respectivament (334). Curiosament, la CNV identificada en els dos últims estudis és una deleció a *MYBPC3* idèntica, que involucra diversos exons (començant a l'intró 27 i acabant 485 pb després del codó *stop* de *MYBPC3*) i considerada com la causa genètica de la malaltia per provocar haploinsuficiència. Al 2012, Herman et al. investiguen el gen *TTN* mitjançant seqüenciació d'alt rendiment en 231 pacients de MCH en busca de mutacions puntuals o estructurals, però no troben cap CNV (335). Recentment, en un estudi de 22 pacients de MCH en els quals es va fer el cribratge de 23 gens mitjançant tècniques de seqüenciació per captura no es va trobar cap alteració estructural (336).

El primer estudi exhaustiu en incloure una gran cohort de pacients de MCH i un ventall més ampli de gens associats a la malaltia va publicar-se al 2015 per Lopes et al. Mitjançant tècniques de seqüenciació modernes van analitzar-se 19 gens associats a la malaltia en 505 pacients, detectant-se 4 CNVs (0.8%): una deleció de 4 exons a *MYBPC3*, la deleció dels primers 4 exons de *PDLIM3*, la duplicació de tot el gen *TNNT2*, i una duplicació de 5 exons a *LMNA*. Les delecions es van considerar variants probablement patogèniques, mentre que les duplicacions es van considerar com VSI (337). Recentment, Ceyhan-Birsoy et al. van fer el cribratge de CNVs en 708 pacients de MCH mitjançant l'ús d'un panell de captura de 46 gens associats a cardiomiopaties; van detectar 4 CNVs (0.56%): la duplicació de l'exó 2 de *MYOZ2*; la deleció de l'exó 12 al 20 de *MYBPC3*; la duplicació de tot el gen *NEXN* i la duplicació dels gens *GLA*, *LAMP2*, *EMD* i *TAZ*, detectades en un pacient amb trisomia del cromosoma X. D'aquestes, tant sols la CNV a *MYBPC3* va ser classificada com a patogènica (338).

b. La Miocardiopatia Dilatada (MCD)

La MCD és caracteritzada per la dilatació ventricular, que comporta l'alteració de la funció

sistòlica, principalment per l'engruiximent de la paret del ventricle esquerre. La prevalença de la MCD és d'1 individu per cada 2.500 (339) i les manifestacions clíniques més comunes són les palpitations, la insuficiència cardíaca o la MSC.

La MCD pot ser causada per factors externs, com l'abús continuat d'alcohol, per exemple (340) o per una base genètica (MCD familiar). Existeixen 2 patrons d'herència associats a la MCD familiar: l'autosòmica dominant i la lligada al cromosoma X (gen *DMD*). La penetrància depèn de l'edat, això vol dir que un portador d'una variant rara i causal és més probable que presenti una manifestació clínica de MCD amb l'edat, i que l'avaluació fenotípica normal d'un altre individu mitjançant ecocardiograma i electrocardiograma no exclou la possibilitat d'aparició de la malaltia en edats més avançades. L'expressivitat també pot ser variable, provocant que les manifestacions clíniques entre individus de la mateixa família (i portadors de la mateixa variant) oscil·lin des de les més lleus, com la disfunció sistòlica mínima o l'ejecció mínima del ventricle esquerre fins a fenotips agressius característics de la malaltia totalment desenvolupada, podent arribar a ser necessari un transplantament de cor.

Amb el cribratge s'aconsegueix trobar la causa genètica més plausible de la malaltia en el 50-60% dels casos (310); un percentatge que varia en funció de gens inclosos i de les característiques fenotípiques dels individus sotmesos a estudi, sent més resolutiu pels individus que presentin problemes en la conducció, o concentracions elevades de creatina-cinasa. En general, els principals gens associats amb la MCD (sobretot *LMNA* i *TTN*) codifiquen per proteïnes integrants del citoesquelet del sarcòmer, l'embolcall nuclear i el sarcolemma (341). A més, s'han identificat mutacions rares *missense* en més de 30 gens codificants per proteïnes del citoesquelet, de miofilaments i canals iònics, però la majoria acostuma a explicar un percentatge marginal dels casos (342–344).

La MCD s'associa amb la distròfia muscular, i qualsevol pacient amb una forma desconeguda de miopatia esquelètica ha de ser avaluat per criteris de MCD. La malaltia és predominantment associada a les distrofinopaties, com a resultat de mutacions en la distrofina, i pot conduir a la distròfia muscular de Duchenne, a la distròfia muscular de Becker, i a la miocardiopatia lligada al cromosoma X. També pot manifestar-se clínicament en un pacient amb distròfia miotònica, amb miopatia miofibril·lar, i moltes de les distròfies musculars d'extremitats i cintura (288).

Les CNVs reportades en pacients de MCD són menys nombroses que per la MCH. Al 2010, Gupta et al. identifiquen mitjançant MLPA un portador de la deleció de l'exó 3 al 12 al gen *LMNA* en una cohort de 25 pacients de MCD (4.0%). A partir d'estudis immunohistoquímics conclouen que la CNV és la causa genètica de la malaltia del pacient, provocant una disrupció de l'embolcall nuclear dels miòcits cardíacs deguda a una menor expressió de laminina (345). Norton et al. detecten mitjançant aCGH la deleció de l'exó 4 de *BAG3* en el proband d'una família afectada de MCD; la CNV cosegrega amb els familiars afectats i és considerada patogènica per causar una pèrdua de funció de la proteïna, o haploinsuficiència; el mateix grup va analitzar per MLPA la regió codificant de *LMNA* en una cohort de 58 probands de MCD sense trobar cap anomalia estructural (346). Herman et al. identifiquen una

duplicació patogènica en tàndem de 28 Kb que s'allarga des de l'intró 71 al 124 del gen *TTN* en un proband d'una cohort de 163 pacients de MCD (0.6%) (335).

A l'estudi de Ceyhan-Birsoy, la cohort de MCD estudiada és de 479 pacients; en aquesta s'identifiquen 3 CNVs: la deleció de l'exó 1 de *LMNA*, la duplicació dels exons 193-224 de *TTN* i la deleció dels exons 8 i 9 de *LAMP2*. Les variants són considerades patogèniques per causar una pèrdua de funció de la proteïna, a excepció de la detectada a *TTN*, que és considerada VSI al no poder caracteritzar els punts de trencament de la CNV i desconèixer si es tracta d'una duplicació en tàndem (338).

c. La Miocardiopatia Arritmogènica (MCA)

La MCA es caracteritza per la substitució fibroadiposa progressiva del miocardi, especialment del ventricle dret, tot i que quasi en un 50% dels casos afecta també el ventricle esquerre, donant lloc a l'afectació biventricular. (347,348). Tot i així, també s'han descrits casos esporàdics on només hi ha una afectació del ventricle esquerre. La presència de teixit fibro-adipós (i en alguns casos inflamació) provoca alteracions en la transmissió elèctrica cardíaca que, en ocasions, pot estar també associada a deficiències en la funció mecànica, l'aparició d'arrítmies ventriculars, síncope i MSC (349,350).

La MCA presenta una elevada heterogeneïtat en la seva manifestació clínica; des de cors macroscòpicament normals a alteracions estructurals severes, presentant des d'una absència de símptomes fins a bloqueigs de branca dreta, arrítmies malignes i MSC (350). És per aquest motiu que no existeix una prova diagnòstica definitiva, sinó que es requereix d'un mètode diagnòstic complex basat en un sistema de puntuació, mitjançant criteris morfològics, funcionals, clínics, genètics i d'història familiar –els criteris *Task Force*– (351). Malgrat que aquests paràmetres són molt útils en els casos greus, encara són limitats a l'hora de diagnosticar pacients en estadis incipients de la malaltia. Per aquest motiu, l'ús d'altres tècniques de diagnòstic (no lliures de controvèrsia) agafen cada vegada més protagonisme com a eines complementàries per la diagnosi de la patologia, com per exemple, l'estudi immunohistoquímic en biòpsies (352).

S'estima que la MCA és la responsable d'un 5% del total de les MSC que succeeixen anualment, afectant 1 de cada 5000 persones. Presenta una major incidència entre els homes (80% dels casos) diagnosticats abans dels 40 anys per l'aparició d'arrítmies, síncope o MSC, i molt especialment entre atletes joves, ja que l'esport és un inductor d'arrítmies letals en casos predisposats genèticament a patir la patologia (353). Presenta un origen genètic en aproximadament el 60% dels casos, mostrant un patró d'herència autosòmic dominant, tot i que la penetrància incompleta en conjunt amb l'expressió variable i la dependència de l'edat pot enfosquir els patrons d'herència mendeliana (354–356). Tot i ser molt poc freqüent, s'ha reconegut una forma autosòmica recessiva de la malaltia. És l'associada al fenotip de Naxos, caracteritzat per presentar queratoderma palmoplantar, (l'engruiximent de la capa externa de la pell dels palmells de mans i peus), i cabell arissat; a més de les alteracions cardíques pròpies de

la MCA (357,358). L'heterozigositat composta (l'herència conjunta de diferents al·lels associats a la malaltia per un únic gen) i l'heterozigositat digènica (l'herència conjunta d'al·lels associats a la malaltia per dos gens diferents) s'identifica en un 10% dels casos (359) i pot contribuir a la penetrància variable i a la complexitat de l'herència de la malaltia.

La pèrdua de l'estructura normal del desmosoma és un factor crucial en la patogènesi de la MCA. La causa genètica s'identifica principalment en mutacions als gens que codifiquen per proteïnes desmosòmiques, com *PKP2* (30-40% dels casos), *DSP* (10-15%), *DSC2* (1-5%), *DSG2* (3-8%) i *JUP* (<1%). Conjuntament, són responsables d'entre un 50-60% del total de casos de MCA, mentre que la resta de gens associats representen menys d'un 5% (311).

La identificació de CNVs en pacients de MCA s'ha centrat quasi exclusivament en el cribratge del gen *PKP2*; de ser detectades, són considerades patogèniques per pèrdua de funció de la proteïna o per haploinsuficiència. Cox et al. investiguen la presència de CNVs mitjançant MLPA en una cohort de 149 probands; identifiquen 3 CNVs (2%): la deleció de l'exó 8, una deleció dels 4 primers exons del gen i la deleció dels exons 1-14 de *PKP2* (360). S'han reportat varis casos clínics de pacients de MCA amb defectes a *PKP2*. Roberts et al. reporten dues grans deleccions detectades per MLPA i/o *arrays* de SNPs, una del gen sencer i l'altra de tot el gen a excepció de l'exó 1 (361). D'altra banda Li-Mura et al. detecten, mitjançant un *array* de SNPs, una deleció de 122 Kb que inclou el gen sencer de *PKP2*. L'equip de Ceyhan-Birsoy, al seu torn, detecta per NGS la deleció de l'exó 8 en una cohort de 90 probands (1.1%) (338). Finalment, Sonoda et al. reporten, per una cohort de 71 probands, la deleció del gen sencer de *PKP2* (1.4%), detectada per MLPA. Al caracteritzar-la per qPCR se'n adonen de que la deleció s'allarga fins 1.23 Mb i que inclou els gens *SYT10* i *ALG10*, sense associació aparent al fenotip (362).

d. La No Compactació del Ventricle Esquerre (NCVE)

La NCVE és una de les incorporacions més recents al grup de les miocardiopaties arritmogèniques heretables. És caracteritzada per una aparença morfològica esponjosa resultant de l'excessiva trabeculació del interior del ventricle esquerre que es fa més evident a la porció apical i mediolateral inferior del ventricle esquerre. És el resultat del fracàs del cor en formar un miocardi compacte durant les etapes tardanes del desenvolupament cardíac (363,364). Les manifestacions clíniques abasten un ampli espectre, des de pacients asimptomàtics fins a altres amb fallada cardíaca severa o arrítmies ventriculars i pot anar associada amb característiques de MCH, MCD o vàries formes de malaltia cardíaca congènita. El miocardi pot demostrar funció sistòlica o diastòlica anòmala i la mida, el gruix i la funcionalitat pot canviar de manera inesperada sota el que es coneix com a "fenotip ondulant". A més, els malalts tenen una major incidència de tromboembolismes en comparació amb qualsevol altra miocardiopatia (365).

En comparació amb la MCH o la MCD, la NCVE és poc freqüent. La incidència i la prevalença

exacta es desconeix. La NCVE afecta a nadons, nens petits i adults, sent els infants els de pitjor prognosi, especialment en aquells amb malalties sistèmiques i trastorns metabòlics (365). La miocardiopatia ha estat identificada en famílies amb patrons d'herència lligada al cromosoma X, autosòmica dominant, autosòmica recessiva i d'herència mitocondrial materna (366). A més són freqüents els casos esporàdics, involucrats en aproximadament el 60-70% dels casos. Tot i que s'han identificat diversos gens de susceptibilitat a la NCVE, cap d'ells predomina per sobre els altres i tampoc s'han reportat avaluacions en cohorts grans. Les causes genètiques s'associen típicament a variants no sinònimes poc freqüents, principalment *missense* i en ocasions *nonsense*, de *splicing* o petites insercions o delecions en 15 gens entre els quals s'inclouen alguns codificants per proteïnes del citoesquelet (*ACTC1*), proteïnes del sarcòmer (*MYH7*, *MYBPC3*, *TNNT2*), canals iònics i mutacions al gen *TAZ* (367–369). La malaltia mitocondrial és una característica prominent en els infants i joves afectats per NCVE, i per tant requereix una avaluació independent (370).

Diversos estudis han vinculat la presència de CNVs al receptor de la rianodina (*RYR2*) amb la NCVE en combinació amb fenotips agressius i complexes de TVPC amb altres defectes (371,372). Ja que les CNVs han estat identificades en un gen tradicionalment associat amb la TVPC, aquestes s'expliquen a l'apartat corresponent a la malaltia. A més, Ceyhan-Birsoy et al. identifiquen un portador de la duplicació del gen sencer de *PKP2* en una cohort de 54 pacients de NCVE (1.8%), considerada com VSI (338).

e. La Miocardiopatia Restrictiva (MCR)

La MCR és molt poc freqüent i no es té coneixement sobre la prevalença en població general. Es caracteritza per l'alteració de l'entrada de sang al ventricle i una disminució del volum diastòlic en un ventricle esquerre que no presenta anomalies en la fracció d'ejecció ni a les parets musculars (373). El fenotip de MCR sovint es solapa amb el de MCD o MCH, per aquest motiu, la caracterització de la malaltia és més funcional que estructural (364).

Els símptomes inicials més comuns acostumen a manifestar-se durant la infància, i inicialment no solen estar relacionats directament amb problemes cardíacs, sinó com a manifestacions cardíques de malalties sistèmiques (374). Els nens amb MCR sovint presenten història d'infeccions pulmonars repetides o asma, i no són derivats al cardiòleg fins que no es fa una radiografia de tòrax en la que s'identifica alguna anomalia cardíaca. Entre els afectats també es detecta ascites (líquid a l'abdomen), hepatomegàlia (engrandiment del fetge) i edema. També pot ser que la primera manifestació sigui un so anòmal del cor, que suggereixi insuficiència cardíaca, desmais (en un 10% dels casos) i MSC (375).

La identificació de familiars afectats és crítica per l'assignació del patró d'herència, ja que de ser classificada com MCR familiar, pot presentar un patró autosòmic dominant, recessiu, lligat al cromosoma X o ser heretada per llinatge matern. Un dels factors a investigar en l'afectat i els familiars

és la presència d'una miopatia esquelètica que es manifesti simultàniament amb els altres símptomes, ja que en casos de MCR familiar acostuma a trobar-se entre els afectats en diferents graus de severitat (374). Si no es detecten familiars afectats, el pacient és diagnosticat amb MCR idiopàtica.

La MCR familiar és genèticament heterogènia. Les poques mutacions que s'han reportat han estat identificades principalment als gens sarcomèrics, a *MYH7* i a *TNNI3*, suggerint la causa genètica de la malaltia en tant sols un 5% dels casos (288).

Fins a data d'avui, no s'han reportat CNVs en pacients diagnosticats amb MCR.

III – Les canalopaties

Les canalopaties constitueixen un grup heterogeni de desordres arritmogènics causats per canals iònics cardíacs disfuncionals. Els canals iònics són les proteïnes que regulen el flux de ions a través de la membrana plasmàtica cel·lular i dels orgànuls intracel·lulars. El potencial de membrana de la cèl·lula en repòs i el potencial d'acció de membrana de les cèl·lules excitable depenen en gran mesura de la funció d'aquests canals iònics (376).

a. La Síndrome de Brugada (SBr)

La SBr es caracteritza per l'elevació del segment ST en les derivacions V1-V3 de l'electrocardiograma. També predisposa a un alt risc de MSC, taquicàrdies ventriculars polimòrfiques secundàries i fibril·lació ventricular en absència de defectes estructurals del cor (377–379). La mitjana d'edat d'inici de les manifestacions clíniques es situa al voltant dels 40 anys malgrat que la MSC pot afectar a persones de totes les edats. La SBr sol afectar especialment als homes, al voltant d'un 75% dels casos, motiu pel que es sospita que els factors hormonals puguin tenir un paper rellevant. Dels pacients afectats per aquesta síndrome, entre el 20-50% presenten antecedents de MSC familiar, que juntament amb la identificació de variants associades a la síndrome, els síncope i les arrítmies ventriculars completen els criteris diagnòstics per la SBr (379). En molts pacients, l'elevació del segment ST es manté oculta fins que es desemmascara la síndrome a causa d'algun factor extern. Aquests factors inclouen la febre, el consum de cocaïna, els antidepressius tricíclics, els antihistamínics de primera generació o els fàrmacs antiarrítmics bloquejadors del canal de sodi (com l'ajmalina i la flecainida). Aquests últims s'utilitzen en la pràctica clínica de manera controlada per a desemmascarar la malaltia en aquells casos en que hi hagi indicis de SBr (380).

S'estima que la SBr és la responsable d'entre el 4 i el 12% del total de MSC. La prevalença de la malaltia varia segons la regió geogràfica; així, és d'1-5 pacients per 10.000 habitants europeus i de 12 pacients per cada 10.000 habitants al sud-est asiàtic. La canalopatia presenta un patró d'herència

autosòmic dominant (381). La causa genètica més coneguda de la SBr són les mutacions al gen codificant de la subunitat del canal de sodi cardíac dependent de voltatge (Nav 1.5) *SCN5A*, amb la identificació de les quals s'expliquen un 20-25% dels casos (382). Aquestes mutacions es relacionen amb una pèrdua de funció del canal que provoca el seu tancament prematur, o bé que aquest no s'activi, provocant un escurçament de la fase 1 del potencial d'acció cardíac i donant lloc a arrítmies per reentrada (383).

S'han associat fins a 19 gens més a la SBr, però en conjunt expliquen tant sols entre el 5 i el 10% dels casos. Generalment aquests gens codifiquen per proteïnes moduladores de la funció del canal de sodi, o per canals de calci i de potassi i les seves unitats reguladores. Per tant, la majoria de pacients de SBr (65-75%) queden sense una causa genètica de la malaltia coneguda després de realitzar l'estudi genètic (384).

Les úniques CNVs identificades fins la data en pacients de SBr poden resumir-se en dos estudis. El primer, el realitzat per Eastaugh et al. en el que es presenta el cas clínic d'un pacient de 14 anys amb la deleció dels exons 9 i 10 de *SCN5A* identificada per MLPA i considerada patogènica, assumint que causa una pèrdua de funció del canal iònic (385). En el segon, Mademont-Soler et al. –estudi publicat pel nostre grup durant l'elaboració d'aquesta investigació (Annex 4)– analitzen per seqüenciació d'alt rendiment 20 pacients i investiguen per MLPA la presència d'anomalies estructurals al mateix gen en una cohort de 220 pacients de SBr. Les dues tècniques convergeixen en la identificació d'un portador (0.5%) amb la duplicació dels exons 15 a 28 en estat aparent de mosaïcisme. D'expressar-se en teixit cardíac, l'anomalia podria ser la causant del fenotip (386).

b. La Síndrome del QT Llarg (SQTL)

La SQTL es caracteritza per la prolongació de l'interval QT a l'electrocardiograma, degut a un allargament del temps de repolarització ventricular. L'interval QT és la distància entre el pic QRS, corresponent a la despolarització ventricular, i el final de l'ona T, corresponent al final de la repolarització ventricular. És la canalopatia amb més prevalença entre els adults menors de 35 anys, i afecta 1 de cada 2000 individus (387). Clínicament es manifesta amb síncope i aturades cardíques –principalment desencadenades per l'estrès emocional o físic–, convulsions, arrítmies ventriculars, fibril·lació ventricular i *torsades de pointes*. (388).

Les causes de la SQTL poden ser genètiques (SQTL familiar) o adquirides (per exemple, a partir d'un tractament amb fàrmacs bloquejadors dels canals de potassi) (389). La forma congènita s'associa a alteracions genètiques en canals de potassi, per una disminució dels corrents repolaritzadors, o en canals de sodi, per un retard en l'entrada del flux de ions al miòcit. El patró d'herència és autosòmic dominant i tot i que s'han associat fins a 15 gens diferents a aquesta malaltia, en la immensa majoria de casos la causa genètica s'identifica en mutacions rares a *KCNQ1* (40-55%), *KCNH2* (30-45%) i *SCN5A*

(5-10%). El cribratge identifica la causa genètica de la malaltia en un 75-85% els casos (314).

En el cas de la SQT, les CNVs sempre s'han buscat als gens associats amb la malaltia, especialment a *KCNQ1* i *KCNH2*, mitjançant MLPA. Sempre es consideren la causa genètica de la malaltia per provocar una pèrdua de funció del canal de potassi cardíac o l'haploinsuficiència. La freqüència de detecció de CNVs per la SQT és relativament més elevada en comparació amb la resta de patologies. Al 2006, Koopmann et al. reporten la identificació d'una deleció de 3.7 Kb que involucra l'exó 6 de *KCNH2* en una cohort de 21 pacients (4.7%) (390). Eddy et al. identifiquen 3 CNVs en una cohort de 26 pacients (11.5%): la deleció dels exons 13 i 14 de *KCNQ1*, la deleció dels exons 6 a 14 de *KCNH2* i la duplicació dels exons 9 a 14 també a *KCNH2* (391); Tester et al. troben 2 delecions a *KCNQ1* en una cohort de 42 pacients (4.7%), la primera afectant l'exó 3 i provocant un *frameshift* amb l'aparició d'un codó *stop* prematur; la segona afecta l'exó 7 i l'intró 8 del mateix gen (392); Barc et al. amplien l'estudi per MLPA a tres gens associats amb la SQT, *KCNQ1*, *KCNH2* i *SCN5A* en una cohort de 93 pacients, trobant 3 CNVs (3.2%): la deleció del gen sencer *KCNH2*, una altra afectant als exons 4-14 del mateix i una tercera que afecta els exons 7 i 8 de *KCNQ1*. Aquesta és la mateixa deleció reportada pel nostre grup, a Campuzano et al., (Annex 5) detectada per seqüenciació d'alt rendiment en un cas sever de SQT amb segregació en diversos membres de la família (393,394). Stattin et al. identifiquen per primera vegada en una cohort de 200 pacients una deleció dels exons 9 i 10 de *SCN5A* en un cas de SQT, juntament amb una duplicació de l'exó 2 de *KCNH2* en dos individus diferents (1.0%) després d'afegir al cribratge els gens *KCNE1* i *KCNE2* (395). Finalment, Williams et al. identifiquen en una cohort de 90 pacients una deleció de 2.7 Kb que involucra els exons 13 i 14 de *KCNQ1*, juntament amb la detecció en un pacient control de la duplicació –aparentment no patogènica– dels gens *KCNE1* i *KCNE2*. Una altra deleció del gen sencer de *KCNH2* va ser detectada en un pacient que prèviament havia patit una leucèmia. Al interpretar que probablement la deleció reflectís un esdeveniment somàtic associat a la leucèmia, no es va relacionar amb l'adquisició d'un QT llarg i no es va considerar més en la interpretació dels resultats del pacient (396).

c. La Síndrome del QT Curt (SQTC)

La SQTC es caracteritza per presentar un interval de QT inferior als 330 ms, amb una ona T alta i acabada en punxa. Els fenotips són molt agressius i l'edat d'aparició de les primeres manifestacions clíniques oscil·la des de la infància fins als 40 anys, amb símptomes com la fibril·lació auricular, les arrítmies ventriculars, les síncope i la MSC (397).

La prevalença de SQTC és d'un 0.02-0.1% en la població adulta i d'un 0.05% en població pediàtrica, amb una major prevalença entre els individus de sexe masculí (398). Diversos estudis han intentat establir la prevalença d'interval QT curts en població general sana. L'estudi de Moriya et al. en població general va demostrar que es poden trobar interval QT curts amb una prevalença del 0.01%

en individus sans, sense risc de patir una MSC, i que aquests pertanyien a la cua de la distribució normal poblacional del paràmetre (399). La síndrome presenta un patró d'herència autosòmic dominant i fins la data s'han associat 3 gens codificants per proteïnes de canals de potassi, també associats amb la SQT: *KCNH2*, *KCNQ1* i *KCNJ2* (400); i 3 gens codificants per canals de calci: *CACNA1C*, *CACNB2*, i *CACNA2D1*. Les mutacions en aquests tres gens comporten un guany de funció del canal de potassi. En conjunt, les mutacions en aquests sis gens aconseguen donar una explicació genètica al 15-40% dels casos reportats, tot i que per la majoria dels pacients existeix una associació familiar. No obstant, aquest percentatge pot ser més baix degut al biaix en la publicació dels fenotips més severos i de la baixa prevalença de la malaltia (288,401–403).

Fins a data d'avui, no s'han reportat CNVs en pacients diagnosticats amb SQT.

d. La Taquicàrdia Ventricular Polimòrfica Catecolaminèrgica (TVPC)

La TVPC es caracteritza per l'aparició de taquicàrdies ventriculars polimòrfiques i/o bidireccionals induïdes per catecolamines durant l'estrès emocional o físic, i per poder causar episodis de síncope i MSC en individus joves (316,404). La prevalença estimada de la malaltia és d'1 entre 10.000 individus (405). El principal símptoma dels pacients són les síncope induïdes per l'esforç físic o emocional i, en menor mesura, la presència de palpitations i mareigs en les mateixes condicions. En un 30% dels casos, la MSC pot ser la primera manifestació clínica de la malaltia (288). Donat que el desencadenant de la malaltia és l'estrès i que el patró de l'electrocardiograma en repòs és completament normal, és necessari realitzar una prova d'esforç o un monitoratge amb Holter per poder-la diagnosticar.

La canalopatia presenta, al igual que la majoria de canalopaties, penetrància incompleta, expressivitat variable i dos patrons d'herència: l'autosòmic dominant, causada per alteracions genètiques en el gen del receptor de rianodina (*RYR2*) en un 50-55% dels casos i el gen codificant de la calmodulina (*CALM1*), en menys d'un 1% dels casos; i autosòmic recessiva, causada per alteracions genètiques en el gen de la calsequèstrina i triadina (*CASQ2* i *TRDN*), en un 5% i 2% dels casos, respectivament. Sense excepció, les proteïnes codificades pels gens associats tenen un paper important en la regulació del calci intracel·lular; l'alteració genètica provoca un guany de funció, provocant un increment de la sortida de calci provinent del reticle sarcoplasmàtic, que conduirà a una despolarització tardana que facilitarà l'arrítmia (406). Per tant, el cribratge genètic identifica la causa de la malaltia en un 65% dels casos, aproximadament (316). Alguns estudis han associat mutacions als gens *KCNJ2* i *ANK2* amb taquicàrdies ventriculars induïdes per catecolamines, però és un tema que suscita certa controvèrsia, ja que aquests dos gens s'han associat a formes de SQT i es creu que podria tractar-se de fenocòpies de TVPC (407).

La identificació de CNVs en pacients de TVPC s'ha centrat sempre en el gen codificant pel

receptor de la rianodina *RYR2*, pel que s'han reportat un total de 10 pacients portadors de la deleció de l'exó 3, flanquejat per seqüències *Alu*, causants de la deleció (408). En la majoria dels casos, el fenotip és sever i complex, combinant característiques de TVPC amb altres miocardiopaties: a l'estudi de Bhuiyan et al. identifiquen 2 probands portadors de la deleció que també presenten característiques fenotípiques de MCD i MCA, a banda de cosegregació familiar (408); els pacients de TVPC portadors de la deleció –identificades per MLPA– reportats en els altres estudis o bé tenen característiques fenotípiques o bé van ser diagnosticats també de NCVE (371,372,409,410). En el primer s'identifiquen 2 portadors d'un grup de 17 pacients; en el segon són 3 els portadors identificats entre 110 pacients (2.7%); en el tercer 2 de 24 (8.3%) i el quart és un cas clínic d'un pacient de 20 anys que veu agreujat el fenotip de TVPC al desenvolupar una NCVE.

A Tang et al. es reporta que la deleció de l'exó 3 de *RYR2* en miòcits cardíacs provoca uns nivells més elevats de calci citosòlic, en comparació amb els controls. El mateix estudi conclou que mutacions puntuals a l'extrem N-terminal de *RYR2*, prèviament associades a MCD i MCA també resulten en concentracions més elevades de calci citosòlic (411). Aquestes observacions porten a postular a Ohno et al. que la gestió aberrant del calci citosòlic, atribuïble a un alliberament aberrant de calci, és un mecanisme comú per les miocardiopaties associades a *RYR2* (372,411). L'estudi de Campbell et al. suggereix la possibilitat d'un nou mecanisme d'aparició de NCVE: la malaltia reflecteix una anomalia de la morfogènesi primerenca del miocardi; no obstant, la NCVE pot desenvolupar-se més endavant, al llarg de la vida, tal i com passa amb el seu pacient. Això suggereix que la NCVE pot ser el resultat d'un error en el desenvolupament i un procés latent mitjançant el qual un miocardi ostensiblement normal esdevé no compactat. El desenvolupament de NCVE en aquest pacient coincideix amb l'empitjorament clínic. El pacient va desenvolupar trabeculacions un any abans de la primera aturada cardíaca. Tot i així, són necessaris més estudis per determinar si aquestes mutacions que causen els nivells anòmals de calci citosòlic són suficients per causar NCVE *in vivo*, i per aclarir si el desenvolupament d'una NCVE és un predictor clínic d'un fenotip més sever de TVPC (371).

e. La Fibril·lació Auricular (FA)

La FA és l'arrítmia més comuna observada en la pràctica clínica i els mecanismes fisiopatològics que la provoquen són complexes. Pot ser induïda per hipertensió, problemes valvulars cardíacs i fallada cardíaca i s'associa a un major risc d'accident cerebrovascular, amb insuficiència cardíaca, a demència, i a MSC (412). Entre les manifestacions clíniques s'inclouen les palpitations, la fatiga, la dispnea com a conseqüència d'un esforç físic, i el dolor toràcic, però la majoria de pacients són diagnosticats de FA de manera accidental (413).

La prevalença és d'1% en la població, i augmenta amb l'edat. Presenta un patró d'herència autosòmic dominant (414) i malgrat que sembla que els factors ambientals són determinants en la seva

manifestació (415), els estudis clínics publicats en les últimes dues dècades recolzen el paper clau de la genètica en la FA. Així doncs, s'identifica història familiar de la malaltia en un 30% dels casos diagnosticats clínicament. Les causes genètiques acostumen a ser variants molt poc freqüents en gens associats a canals iònics (sobretot *KCNQ1*, *KCNE2*, *KCNJ2* i *KCNH2*), a proteïnes relacionades amb el metabolisme del calci o amb la predisposició a la fibrosi, a malalties del sistema de conducció cardíac i a processos inflamatoris (412). S'han associat a la malaltia altres gens codificants per canals iònics, però sempre en formes esporàdiques de FA (*GJA1* i *GJA5*, els gens codificants per la connexina 40 i 43, respectivament) (416). Tot i el número de gens associats a FA, les mutacions identificades semblen ser úniques per famílies individuals i, per tant, són considerades com causes rares de l'arrítmia. No es recomana el cribratge genètic per la seva diagnosi.

L'únic estudi que relaciona la presència de CNVs amb la susceptibilitat a patir FA és el de Tsai et al. En aquest s'identifica –mitjançant un *array* de SNPs– la duplicació de 4.5 Kb a l'intró 1 de *KCNIP1*, el gen codificant per una proteïna d'unió a calci, una subunitat auxiliar moduladora del complex proteic Kv4.3 que forma el canal de potassi neuronal. Després d'estudiar més de 430 pacients de FA amb diferents graus de severitat i aproximadament 1000 controls taiwanesos, determinen que la freqüència de la mutació estructural és més elevada en els pacients amb FA (*odd ratio* = 2.27). Conclouen un efecte deleteri de la CNV demostrant en un estudi funcional *in vivo* que la mutació provoca una major expressió de *KCNIP1*, gen associat al mecanisme de manteniment de freqüències auriculars més elevades, convertint-lo en una possible diana terapèutica (417).

IV – La Síndrome de Marfan

La Síndrome de Marfan (SM) és una malaltia sistèmica del teixit connectiu caracteritzada per una combinació variable de manifestacions cardiovasculars, músculo-esquelètiques, oftalmològiques i pulmonars. Les manifestacions clíniques poden aparèixer a qualsevol edat i són molt variables inclús dins de la mateixa família. L'afectació cardiovascular es caracteritza per la dilatació progressiva de l'aorta, acompanyada d'un risc elevat de dissecció i aneurisma toràcic i d'aorta (DATA) i d'insuficiència de la vàlvula aòrtica –la principal causa de mort entre els afectats–; i per insuficiència mitral, que pot complicar-se amb arrítmies, endocarditis o insuficiència cardíaca. Per aquest motiu, tot i que no sigui considerada una malaltia arritmogènica sota la mateixa classificació que les miocardiopaties o les canalopaties, i considerant el paper rellevant de les variants estructurals en la genètica dels pacients, s'ha inclòs un conjunt de pacients de SM a la cohort d'estudi d'aquesta tesi.

La prevalença estimada de la SM és d'1 per cada 5000 individus i afecta per igual els dos sexes. En la immensa majoria dels casos (75-90%), la causa genètica es deu a mutacions a *FBN1*, el gen codificant per la fibril·lina-1, una proteïna essencial del teixit connectiu. Les mutacions en aquest gen també s'han associat amb defectes en la conducció elèctrica del miocardi i predisposició dels pacients

a taquicàrdies ventriculars, fibril·lació auricular i taquicàrdia intranodal (418). S'han identificat algunes formes de la malaltia en pacients portadors de mutacions als gens *TGFBR2*. La diagnosi es basa en l'evidència clínica i en la història familiar; no obstant, com el quadre clínic pot ser molt variable, s'han establert uns criteris diagnòstics internacionals basats en l'evidència clínica.

Fins a data d'avui s'han detectat i reportat més de 3000 mutacions a *FBN1*, que poden consultar-se a la *Universal Mutation Database* (419). Les CNVs identificades a *FBN1*, un total de 51 (delecions en un 98% dels casos), representen un 1.6% del total de mutacions causals de la SM i la seva detecció mitjançant MLPA, forma part dels protocols rutinaris de diagnosi genètica per aquest tipus de pacients (420).

II. Justificació de la recerca, hipòtesis i objectius

JUSTIFICACIÓ DE LA RECERCA

Al llarg de la introducció s'ha descrit el marc i les bases clíniques de la MSI i les bases genètiques de les malalties arritmogèniques associades a la MSC. S'ha exposat la importància d'una diagnosi genètica complerta i acurada, tant pel suport al fenotip –i l'ajuda al facultatiu en la presa de decisions sobre el tractament–, com pel consell genètic, la detecció precoç de familiars portadors asimptomàtics i, en alguns casos, la determinació del risc a patir una MSC del pacient.

La majoria de malalties cardíques arritmogèniques d'etiologia genètica són considerades d'herència mendeliana. Tot i així, l'evidència acumulada en l'última dècada (a partir de la seqüenciació massiva i el cribratge de milers d'individus) suggereix patrons d'herència complexes per algunes d'elles (300,301). Aquestes malalties comparteixen característiques clíniques i genètiques comunes: l'heterogeneïtat genètica –la majoria són considerades oligogèniques o poligèniques– i fenotípica (81); l'expressivitat variable i la penetrància incompleta. En alguns casos, el solapament fenotípic i dels gens associats entre els diferents desordres enfosqueix els patrons d'herència i dificulta la diagnosi i el consell genètic (354,356).

Per aquests motius, en ocasions la diagnosi genètica pot ser fal·lible, o insuficient per poder emetre una diagnosi apropiada. Sovint, després de la investigació de les regions codificants dels gens associats a la malaltia del pacient, tant sols s'hauran identificat variants de significat incert o classificades com benignes. En aquests casos no s'aconsegueix determinar una causa genètica clara de la malaltia. El percentatge de casos no resolts varia per cada patologia i en funció de l'extensió del cribratge realitzat (288,309–316). Al ser malalties poc freqüents es necessiten cohorts de mida extensa per poder extreure conclusions a nivell poblacional.

Hi han diverses causes per les que un cas clínic pot no resoldre's després del cribratge genètic. Deixant de banda que es pugui tractar d'un cas complex, en funció dels punts exposats anteriorment (i, per tant, més lligat a la interpretació del genetista), és possible que la variant patogènica estigui localitzada en un gen fins al moment no associat a la malaltia (motiu pel que pot haver estat exclosa de l'estudi). La causa genètica pot estar localitzada en una regió reguladora no codificant –no incloses a les anàlisis genètiques rutinàries (317)–; o inclús no ser identificable en la seqüència de DNA del pacient, sinó que el fenotip tingui una causa epigenètica (318). Una altra possibilitat és la que ocupa un tema central en aquesta tesi: la presència de variants estructurals no detectables per les tècniques de seqüenciació capil·lar convencionals. En els últims anys s'han identificat una gran quantitat de variants estructurals al genoma humà (319), tant en la població general com en cohorts de malalts (267). Fins ara, però, la prevalença de les CNVs –reestructuracions genòmiques desequilibrades superiors als 50 pb (421)– en pacients diagnosticats amb malalties cardíques associades a la MSC és un tema poc explorat. S'han reportat nombroses sèries de pacients (petites) i diversos casos puntuals de portadors de CNVs (per una revisió exhaustiva consultar l'apartat 1.4 de la Introducció). Manquen estudis robustos que incloguin un ampli ventall de gens i un número elevat de pacients, per tal de poder conèixer

l'impacte real d'aquestes variants en els malalts.

HIPÒTESIS

Tenint en compte el marc teòric exposat i la implantació, a l'inici d'aquest treball, de les tecnologies de seqüenciació d'alt rendiment al laboratori de diagnosi genètica del Centre de Genètica Cardiovascular, s'han formulat les següents hipòtesis que emmarquen la present tesi doctoral:

1. Un cert percentatge de pacients diagnosticats amb malalties arritmogèniques associades a la MSC o que hagin patit una mort sobtada inexplicada continuaran sense causa genètica després del cribratge convencional de variants puntuals i *indels*. La nostra principal hipòtesi és que un percentatge d'aquests pacients negatius seran portadors de CNVs en gens associats o candidats a la malaltia per la que han estat diagnosticats.
2. En els casos en els que s'identifiqui un portador de CNV, la variant podrà explicar el fenotip clínic, contribuint així a ampliar el coneixement de les causes genètiques de les diferents patologies.
3. Les recomanacions actuals per la interpretació clínica de variants genètiques (422) es centren en mutacions puntuals i en *indels*. Una classificació específica de les CNVs pot aportar informació valuosa als casos clínics de portadors d'aquest tipus d'alteracions genètiques.

OBJECTIUS

Per tal de donar resposta a les nostres hipòtesis, s'han formulat els següents objectius:

1. Desenvolupar un algoritme informàtic per la detecció de CNVs a partir de dades de seqüenciació d'alt rendiment obtingudes amb panells de captura. Aquest ha de ser robust, fiable i fàcil d'implementar en la rutina d'un laboratori de diagnosi genètica.
2. Realitzar el cribratge genètic exhaustiu d'una gran cohort de 2073 pacients diagnosticats amb malalties associades a la MSC hereditària i a la MSI. Aquest ha de combinar la detecció de variants puntuals, *indels* i CNVs per les regions codificants dels principals gens associats a aquestes malalties, els minoritaris i els candidats.
3. Fer la translació a la clínica dels resultats obtinguts en base a uns criteris de classificació de variants estructurals propis, amb la intenció de facilitar una avaluació clínica i/o forense més concisa, millorant l'assessorament genètic i les mesures preventives per als pacients i els seus familiars.

III. Material i Mètodes

3.1 – Declaració de contribució

Degut a la transversalitat d'aquest estudi, ha estat necessari el treball col·laboratiu de l'equip de professionals del Centre de Genètica Cardiovascular, que ha contribuït activament en les tasques que es llisten a continuació:

El Sr. Ferran Picó va extreure el DNA procedent de les mostres de sang i de saliva. Tant els processos de control de qualitat del DNA com la preparació de llibreries de DNA genòmic van ser duts a terme per la Dra. Irene Mademont, la Dra. Mònica Coll i la Sra. Alexandra Pérez, el mateix equip que també va encarregar-se de la confirmació de les variants identificades per l'anàlisi bioinformàtica. El Dr. Carles Ferrer va crear i optimitzar els dissenys de sondes dels panells utilitzats en la seqüenciació.

Jo he seleccionat i gestionat les dades de la cohort d'estudi. He dut a terme el control de qualitat i el processament bioinformàtic de les mostres seqüenciades. He dut a terme la detecció de variants puntuals i he desenvolupat un algoritme de detecció de CNVs, amb el qual s'han analitzat totes les mostres de la cohort d'estudi. També he donat suport informàtic a l'equip de confirmació de resultats per resoldre els problemes que anaven apareixent. Conjuntament amb la Dra. Irene Mademont hem proposat uns criteris de classificació per CNVs, hem classificat les variants detectades i hem escrit els manuscrits que es desprenen d'aquest treball.

3.2 – Consentiment informat

Cada actuació en matèria de salut requereix de la firma d'un consentiment per part del pacient. Aquest document demostra que el pacient, o un representant legal autoritzat, dóna el vistiplau al facultatiu per sotmetre'l al procediment d'estudi, de manera conscient i lliure de coacció. Si del procediment en qüestió se'n desprenen resultats genètics, el consentiment guanya encara més importància, degut a la naturalesa única i permanent de les dades genètiques d'un individu, que el fan susceptible de ser identificat.

Dels resultats d'un estudi genètic poden derivar-se implicacions reproductives del investigat i són dades amb potencial poder discriminatori, tant a nivell familiar, com social o ocupacional. Donat que poden detectar-se alteracions genètiques heretables, els resultats poden tenir repercussions en els integrants de la família. D'altra banda, els resultats poden ser de difícil interpretació, reinterpretables i rebel·lar malalties en individus asimptomàtics en el moment de l'estudi.

Al document de consentiment informat hi ha de constar que el pacient ha estat informat dels següents punts:

- ❖ El propòsit del procediment al que s'està a punt de sotmetre
- ❖ Els detalls del procediment d'estudi

- ❖ Els possibles riscos associats al procediment
- ❖ La confidencialitat i la gestió de les dades personals per la part sol·licitant

També hi ha d'aparèixer la informació proporcionada al pacient, la comprensió de la informació per part d'aquest, i la voluntarietat de la cessió de la mostra biològica per l'estudi genètic.

En aquest treball s'ha utilitzat el consentiment informat de Ferrer inCode S.L (Barcelona, Espanya), realitzat en col·laboració amb el Centre de Genètica Cardiovascular de Girona (Annex 1). Tant aquest treball com el consentiment informat van ser aprovats pel Comitè Ètic de l'Hospital Josep Trueta de Girona, en conformitat amb els principis de la Declaració de Hèlsinki de 2008 per les investigacions mèdiques en éssers humans.

3.3 – Cohort d'estudi

En aquest treball s'ha estudiat una cohort de 2073 pacients caucàsics, sense relació de parentesc (Figura 3-1). Entre aquests, 1369 casos clínics (65'9% de la cohort) van ser diagnosticats de malalties cardíaques arritmogèniques associades a MSC per cardiòlegs de 15 hospitals repartits pel territori de l'estat espanyol. De tots els pacients es va extreure o bé una mostra de sang perifèrica o de saliva. D'aquestes mostres es va extreure el DNA per procedir a la investigació genètica. L'edat mitja d'aquesta porció de la cohort és de $43'5 \pm 20'8$ anys, amb una relació de sexes 3:1, amb major incidència en els homes. La cohort es divideix en tres subgrups:

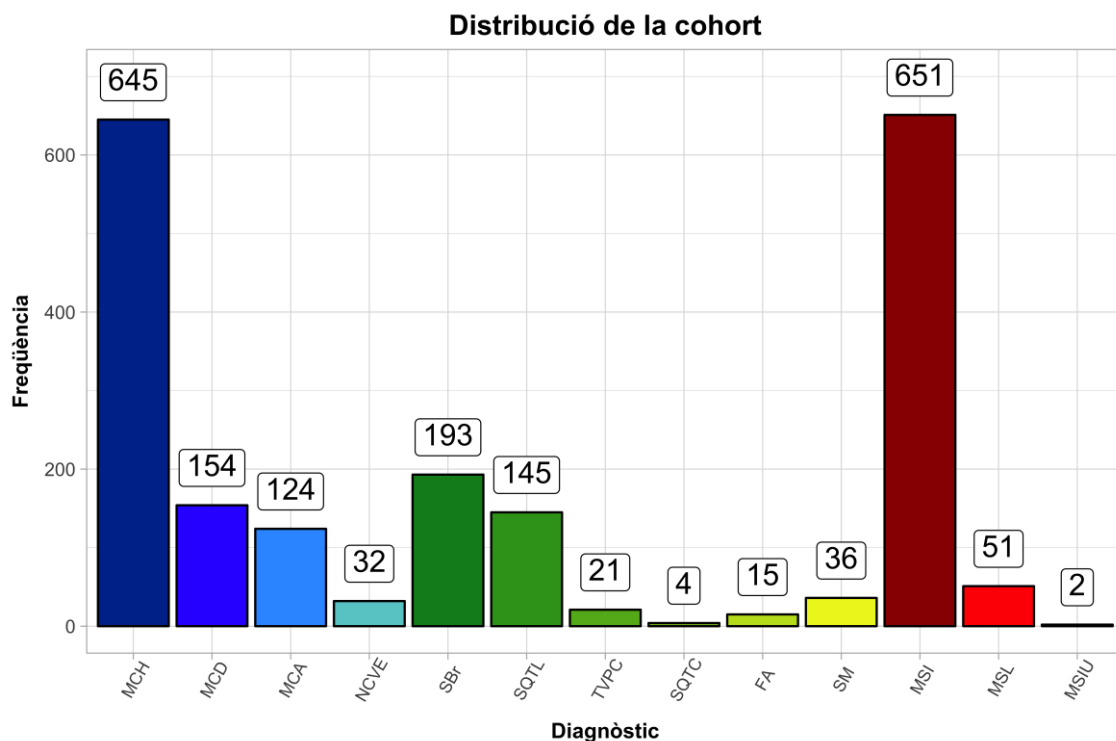


Figura 3-1 | Distribució dels pacients de la cohort d'estudi en funció del diagnòstic.

3.3.1 – Pacients diagnosticats amb miocardiopaties

El grup el formen 955 pacients (46'0% de la cohort) diagnosticats amb miocardiopaties associades a MSC: 645 (31'1%) de MCH, 154 (7'4%) de MCD, 124 (6'0%) de MCA i 32 (1'5%) de NCVE. Tot i no tractar-se exactament d'una miocardiopatia, aquest grup també inclou 36 pacients de SM i DATA (1'7%) –els pacients de SM tenen predisposició a patir taquicàrdies ventriculars, fibril·lacions auriculars i taquicàrdies intranodals (418)–.

3.3.2 – Pacients diagnosticats amb canalopaties

El grup és format per 378 pacients (18'2%) diagnosticats amb canalopaties associades a la MSC: 193 (9'3%) de SBr, 145 (7'0%) de SCTL, 4 (0'2%) de SQTC, 21 (1'0%) de TVPC i 15 (0'7%) de FA.

3.3.3 – Casos de MSI

El projecte MOSCAT va iniciar-se al 2012 pel Centre de Genètica Cardiovascular en col·laboració amb el Institut de Medicina Legal i Ciències Forenses de Catalunya (IMLCFC). El IMLCFC supervisa i centralitza tots els casos de MSI que s'esdevenen a Catalunya i que requereixen d'una autòpsia legal degut a la naturalesa imprevista de la mort. El projecte té com a objectiu principal determinar si la causa de la MSI en pacients joves, en els quals l'autòpsia no resulta conclusiva, té una base genètica que pot posar en risc altres membres de la família.

Per ser inclosos dins del projecte, els casos han d'acomplir uns certs requisits, com ser víctimes de mort sobtada, ser menors de 50 anys i haver patit una mort natural. A tots els individus se'ls hi va practicar una autòpsia completa d'acord amb la regulació internacional actual per part de metges forenses (423,424). En absència d'alteracions macroscòpiques, se'ls hi va realitzar una investigació histològica i toxicològica completa, que incloïa la recollida de mostra sanguínia per l'estudi genètic. Tots els casos inclosos a la cohort d'estudi van resultar negatius per l'anàlisi toxicològica i per tots ells la investigació microscòpica no va identificar cap alteració cardíaca que pogués explicar la causa de la mort.

La cohort inclou 704 casos de MSI (34'2% de la cohort). La major part del grup (72%) la formen els casos de MSI provinents del projecte MOSCAT, tots ells sense causa conclusiva de la mort després d'un examen complet, motiu pel qual es sospita que la causa més probable sigui cardíaca. Aquest grup inclou 51 casos de MSL (2'5%) i 2 casos de Mort Sobtada Intrauterina (MSIU) (0'1%). La resta de casos corresponen a pacients de Mort Sobtada Recuperada (MSR), sense diagnòstic de miocardiopatia o canalopatia després d'una exploració completa. L'edat mitja d'aquesta porció de la cohort és de 32'3 ± 14'9 anys, amb una relació de sexes 5:1, amb major incidència en els homes.

3.3.4 – Pacients diagnosticats amb hipercolesterolèmia familiar

Amb l'objectiu de validar els resultats de l'algoritme de detecció de CNVs, i sense computar pel global de la cohort estudiada, es va seleccionar un grup control de 108 mostres provinents d'un banc de DNA de pacients diagnosticats amb hipercolesterolèmia familiar. Aquests pacients havien estat prèviament caracteritzats mitjançant seqüenciació Sanger i MLPA, de manera que es coneixia quins d'ells eren portadors de variants estructurals. Les llibreries genòmiques van ser preparades per tècnics del laboratori de Gendiag S.L (Esplugues del Llobregat, Espanya), lloc on també van ser seqüenciades en un MiSeq. Les mostres d'hipercolesterolèmia van ser processades i analitzades pel Dr. Carles Ferrer amb l'algoritme de detecció de CNVs que he desenvolupat en aquesta tesi, per tal de poder-ne avaluar la sensibilitat –la capacitat de l'algoritme per detectar les CNVs en els portadors– i l'especificitat –la capacitat de l'algoritme per detectar l'absència de CNVs en individus que no siguin portadors–.

3.4 – Disseny de panells de captura

Com s'ha comentat al llarg de la introducció, les tècniques de seqüenciació d'alt rendiment han estat incorporades progressivament durant l'última dècada a la diagnosi genètica. Entre totes les opcions de seqüenciació actuals, el disseny de sondes per la captura de regions genòmiques amb una certa rellevància clínica associada es presenta com la més resolutiva i fàcilment implementable en un laboratori de diagnosi genètica (425). Les aplicacions clíniques han de considerar com d'informatiu i pragmàtic resultarà l'estudi, les restriccions de cost per pacient, el temps de processament dels resultats i el guany de cobertures de les regions d'interès (426) –ja que la cobertura mínima acceptada per considerar vàlid l'estudi d'una regió és de 30x (427)–. Els panells de captura resulten l'opció amb millor relació cost-efecte en comparació amb la seqüenciació d'un exoma o genoma sencer. Per aquestes opcions (i sempre dins de l'àmbit clínic) encara es requereix una inversió important per obtenir unes cobertures acceptables.

En aquest treball s'han utilitzat diversos panells de gens (Taula 3-1). El primer incloïa les regions codificants i UTR dels 55 gens associats a la MSC amb major prevalença. Aquest primer disseny va evolucionar i ampliar-se en altres panells amb gens de menor prevalença (minoritaris) o candidats, perdent en la transformació les sondes dissenyades per la captura de les regions UTR (els panells de 78 i 85 gens).

En el panell de 118 gens, contemporani al de 78, s'inclouen gens associats a la SM i a la DATA. Aquest panell va evolucionar a un de 147 gens en el que s'afegien sondes dissenyades per les regions codificants de gens associats a rasopaties, a banda d'ampliar els gens candidats per la MSL. Per últim també va dissenyar-se un panell d'hipercolesterolèmia familiar, que inclou els principals gens associats amb aquesta malaltia familiar hereditària.

Els diferents panells utilitzats es recullen a la Taula 3-1. A l'Annex 2 s'hi pot trobar la relació de gens i isoformes incloses en cadascun d'ells.

Taula 3-1 | Llista dels panells de gens utilitzats per la seqüenciació de mostres.

Panell	Gens	Exons	Isoformes	UTR	Bases cobertes (Kb)
Sudd.v1.55	55	1201	55	Si	243'89
Sudd.v2.78	78	1569	79	No	300'87
Sudd.v3.85	85	1723	90	No	328'58
SuddXL.v2.118	118	2190	119	No	403'64
SuddXL.v3.147	147	2672	159	No	442'13
HF.v2	5	91	5	No	31'31

3.4.1 – Selecció de regions d'interès

El primer pas d'un projecte de seqüenciació que inclogui la selecció de regions diana és el de decidir quines són aquestes regions. Aquest punt dependrà clarament de l'objectiu de cada projecte, ja que és possible incloure gairebé qualsevol regió del genoma, ja sigui una regió codificant, reguladora o sense funció descrita. Cal tenir en compte, però, que no totes les regions poden seqüenciar-se de manera apropiada amb la tecnologia de seqüenciació d'alt rendiment de fragments curts (inferiors a 600 pb). Així, algunes regions repetitives del genoma (entre elles els telòmers i centròmers) i regions amb continguts GC extrems resultaran molt complicades de seqüenciar adequadament.

Actualment, i gràcies als projectes internacionals de seqüenciació dels que s'ha parlat a la introducció, existeixen diverses bases de dades públiques i gratuïtes. Des d'aquestes, l'usuari pot descarregar-se la informació necessària per dissenyar a mida els projectes de seqüenciació. Les coordenades seleccionades per aquest treball van ser descarregades de la base de dades ENSEMBL 75 (428); concretament aquelles vinculades com a mínim a o bé un codi RefSeq (429) o bé a un CCDS (430), i sempre basades en la versió hg19 (NCBI GRCh37) del genoma humà de referència.

Els panells de gens utilitzats en aquesta tesi tenen com a finalitat el cribratge genètic per la diagnòsi clínica de malalties cardiovasculars. En alguns gens concrets van fusionar-se les coordenades de diferents isoformes. D'aquesta manera es disposa d'una isoforma artificial que inclou totes aquelles susceptibles de ser seqüenciades. El processament d'aquestes coordenades resulta en l'arxiu de regions a partir del qual es dissenyen les sondes de captura.

3.4.2 – Criteris pel disseny de sondes de captura

El sistema *SureSelectXT Target Enrichment* (Agilent Technologies, Califòrnia, USA) accepta un màxim de 57.000 sondes de 120 pb d'RNA complementari per disseny. Les sondes van ser dissenyades

pel Dr. Carles Ferrer-Costa a partir de l'arxiu de regions prèviament generat i de la seqüència del genoma humà de referència.

Per un disseny òptim, el primer paràmetre a considerar és el *tiling* –o solapament de sondes–, que equival a la quantitat mínima de sondes diferents amb la que quedarà coberta cada base de DNA que es pretén capturar. A l'hora de considerar el *tiling* apropiat per cada regió entren en joc diversos factors: a) la quantitat total de bases que es vulguin capturar amb el panell (quantes més bases s'hagin de cobrir, menys sonda disponible per cada regió); b) la longitud de les regions en les que dissenyar sonda (el guany de cobertura és, fins a cert punt, cooperatiu; per tant, el *tiling* per regions llargues ha de complir criteris d'economització de sondes sense posar en risc l'homogeneïtat de la cobertura a la regió); i c) l'eficiència d'hibridació predita de les sondes en funció de les característiques intrínseques de la seqüència de DNA que es vol capturar.

Per altra banda, l'eficiència d'hibridació d'una sonda amb la cadena de DNA ve determinada pel contingut GC de la regió, la presència de variants en el lloc d'hibridació, la repetitivitat intrínseca de la seqüència de la regió d'unió, o la identitat de seqüència d'aquesta amb altres parts del genoma on també podrien hibridar les sondes. En aquest escenari, hi hauria un descens significatiu de sondes hibridants a la regió d'interès original. Aquest fet podria comportar un guany insuficient de cobertura o la identificació ambigua de variants, amb l'agreujant de no poder descobrir quina és la localització real de la variant mitjançant els mètodes de confirmació convencionals. Per aquest motiu és preferible que les regions amb homologies significatives siguin excloses del panell en les etapes inicials del disseny.

En regions amb continguts de GC propers al 50% el *tiling* es va establir inicialment en 5x amb un solapament amb la sonda veïna de 20 nucleòtids. De manera progressiva, en aquelles regions on el contingut GC varia fins assolir la qualitat de GC extrem (regió inferior a un 35% o superior a un 60%), o bé es detecten repeticions de seqüència, el *tiling* s'incrementa per compensar la baixa eficiència d'hibridació. S'equilibra també la quantitat total de sondes per regió. Així, en aquelles regions en les que l'eficiència d'hibridació no es veu compromesa, la densitat de sondes dissenyades és inferior a la de les regions més problemàtiques. En aquestes, la quantitat final de sondes vindrà determinada per un factor multiplicador dependent de les característiques de la regió.

El conjunt provisional de seqüències és aleshores avaluat en funció de la qualitat de mapeig que presentin. Sota paràmetres d'alineament conservadors, en els que tant sols es permet un alineament únic i sense *missmatches*, les seqüències que no compleixen els requisits són substituïdes per altres de diferents, o bé per duplicats de seqüències que han demostrat un comportament acceptable (l'alineament de seqüències s'explica a l'apartat 3.6.3).

Finalment, les seqüències són sintetitzades i distribuïdes en forma de solucions de sondes de captura biotinitades d'RNA complementari (Agilent Technologies).

3.5 – Processament de mostres

El processament de les mostres, ja siguin de sang perifèrica (fresca o *post-mortem*) o saliva, comença amb l'extracció del DNA. Aquest procés es va dur a terme mitjançant l'extractor *Chemagic Magnetic Separation Module I* (PerkinElmer, Massachussets, USA), seguint les recomanacions del fabricant. Posteriorment, el DNA va quantificar-se per fluorimetria amb Qubit™, utilitzant el kit comercial *dsDNA Broad Range Assay Kit* (Invitrogen™, Califòrnia, USA). La puresa del DNA va avaluar-se amb NanoDrop® ND-1000 (Thermo Fisher Scientific, Massachussets, USA): tant la proporció 260/280 com la 260/230 han de ser properes a 1.8 (valors allunyats informen de contaminació per proteïnes, o sals i fenols, respectivament). La integritat del DNA va avaluar-se en un gel d'agarosa al 0'8%.

3.5.1 – Preparació de llibreries genòmiques

En primer lloc, el DNA extret per la preparació de llibreries genòmiques ha de ser fragmentat. En funció del protocol de preparació, aquests fragments varien en la mida, però han de mantenir-se dins d'un rang de dispersió moderada. D'aquesta manera s'eviten artefactes en les anàlisis posteriors. En el nostre cas, la fragmentació del DNA va ser física, mitjançant un Bioruptor® NGS (Diagenode, Lieja, Bèlgica). Els fragments obtinguts van ser purificats amb *beads* magnètiques en suspensió aquosa *Agencourt® AMPure® XP* (Beckman Coulter Inc., Califòrnia, USA). Per comprovar la distribució i mida dels fragments obtinguts es va utilitzar un *Bioanalyzer* i el kit *Agilent High Sensitivity DNA Chip o DNA 1000* (els dos d'*Agilent Technologies*). En una bona fragmentació, el rang de mida dels fragments ha d'oscil·lar entre els 150 - 200 pb.

Les llibreries genòmiques van ser preparades seguint el protocol *G7530-90000 SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library* (*Agilent Technologies*). L'única diferència significativa a considerar en relació al protocol comercial és la reducció al 50% del volum de llibreria de captura utilitzada per cada mostra (prèviament validat); aquesta variació va implementar-se amb l'objectiu de reduir costos sense renunciar a uns resultats de qualitat. Un cop finalitzat el protocol de preparació es torna a avaluar la llargada dels fragments al *Bioanalyzer*; aquesta vegada, el pic de la distribució s'ha de visualitzar entorn els 300 - 400 pb, ja que s'han lligat als fragments els adaptadors que permetran la fixació a la cel·la de flux del seqüenciador.

3.5.2 – Pooling i càrrega al seqüenciador MiSeq

Com ja s'ha comentat, en el transcurs dels últims cinc anys hi han hagut millores significatives en el mètode de seqüenciació per síntesi d'Illumina. Aquestes millores, a banda de permetre

l'assoliment de densitats de clústers més elevades i seqüències de fiabilitat superior, han permès l'optimització dels *runs* del MiSeq, permetent l'abaratiment dels costos sense renunciar a una seqüenciació homogènia i de qualitat. Durant aquesta tesi s'ha pogut incrementar el número de mostres combinades per carrera de manera progressiva (passant de poder-ne carregar 6, en un inici, a 10-14 en l'actualitat, en funció del panell utilitzat) i s'han forçat al màxim certs paràmetres, com la densitat de clústers –1200-1400 K/mm² recomanats contra 1600-1700 K/mm² assolits a la pràctica–. A densitats de clúster més elevades s'incrementa la probabilitat de que un major número de seqüències no assoleixin la qualitat requerida per passar els filtres del seqüenciador. Amb una densitat de clústers propera als 1700 K/mm² es maximitza la quantitat de seqüències de qualitat amb un percentatge baix de seqüències de mala qualitat. El resultat habitual són uns 33'5 milions de seqüències generades per *run*, dels quals 30 milions són de qualitat òptima.

Per la càrrega, les llibreries són combinades i normalitzades a una concentració de 10 nM. La concentració de càrrega requerida pel MiSeq és de 12'5 pM. Per arribar a concentracions tant petites minimitzant l'error es fa un banc de dilucions de les mostres combinades –o *pool*–. Les seqüències obtingudes, aparellades, ja que el protocol és *paired-end*, són de 151 pb per mostres seqüenciades amb el panell de 55 gens (i amb la versió antiga del cartutx de seqüenciació) i de 76 pb per la resta de panells.

3.6 – Anàlisi bioinformàtica dels resultats

3.6.1 – Controls de qualitat

La seqüenciació d'alt rendiment de fragments curts és una tècnica efectiva, però limitada. Hi han moltes etapes del procés susceptibles a acumular errors o dades de baixa qualitat que poden condicionar l'anàlisi posterior i inclús alterar-ne els resultats. Per aquest motiu, el primer pas del processament de dades és l'avaluació de la qualitat tant del mateix procés de seqüenciació com de les seqüències generades. Degut a la gran quantitat d'informació obtinguda en cada experiment, aquesta tasca ha de ser executada mitjançant eines informàtiques.

I – Control de qualitat de la seqüenciació

El *software* intern del MiSeq genera mètriques a partir del rendiment del *run* per detectar possibles problemes durant el procés de seqüenciació (Figura 3-2). També realitza gràfics en els que es representen els índexs associats a cada mostra (quant més similars siguin els percentatges, més homogènia serà la combinació de mostres); la intensitat de llum captada en tota la superfície de la cel·la de flux (per desemmascarar problemes de fluídica, com la presència de bombolles d'aire o una distribució poc homogènia dels fragments); histogrames de distribució de les qualitats associades a

cada seqüència i diagrames que resumeixen la quantitat total de clústers generats.

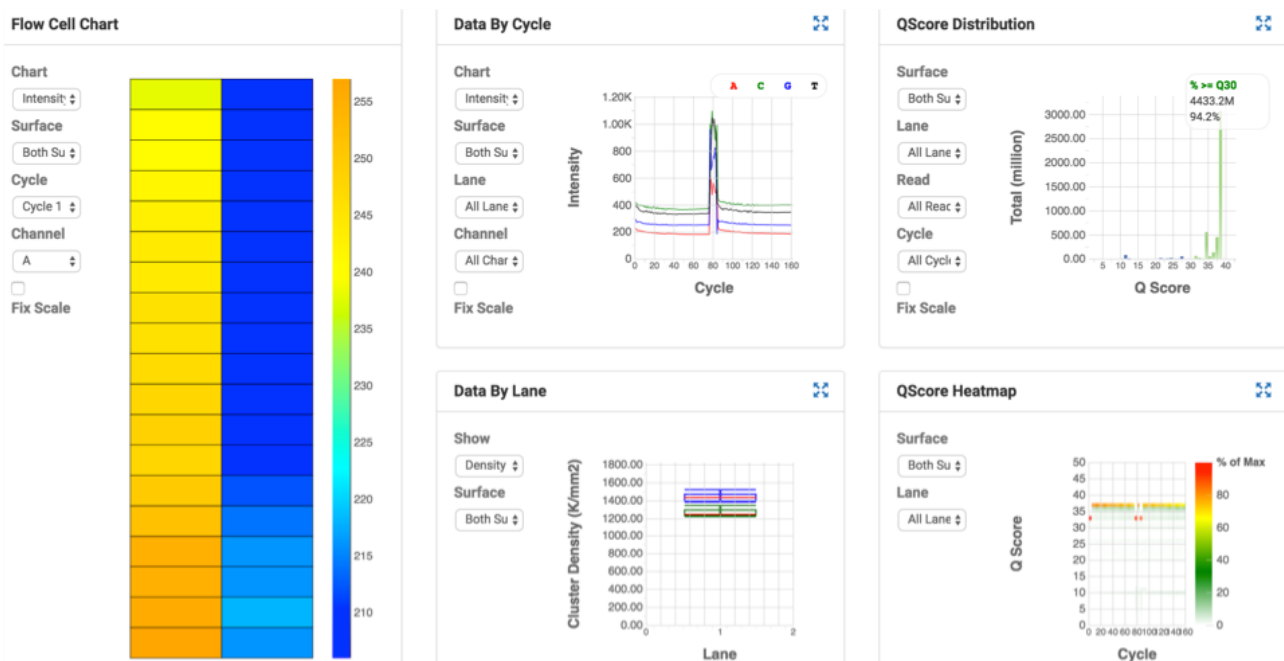


Figura 3-2 | Gràfics de control de qualitat generats en un *run* de MiSeq.

A més, amb cada combinació de llibreries també és seqüenciat el petit genoma del bacteriòfag *PhiX-174*. Aquest genoma, processat de la mateixa manera que la resta de mostres, es subministra com a llibreria a punt per ser seqüenciada i serveix com a control intern per solucionar problemes inherents del mètode de seqüenciació per síntesi, com la generació de clústers, o l'acumulació d'errors a través dels cicles de seqüenciació. Al ser un genoma tan petit pot alinear-se i analitzar-se ràpidament per tal de determinar si els possibles problemes detectats als arxius resultants són causats per un funcionament anòmal del seqüenciador o estan relacionats amb el procés de preparació i càrrega de llibreries.

II – Control de qualitat de les seqüències

Bàsicament hi han dos aspectes a considerar: l'eliminació de les bases associades a una baixa qualitat (Figura 3-3/A) i l'eliminació de possibles seqüències d'adaptadors dels extrems dels fragments seqüenciats.

Les qualitats associades als nucleòtids (el format fastq s'explica a l'apartat 3.6.2) tendeixen a decaure a mesura que incrementa la llargada del fragment seqüenciat. Els fragments es sintetitzen des de l'extrem 5' cap a 3' en base a la iteració de processos bioquímics que no són lliures d'error. Per aquest motiu, a mesura que el fragment creix, els errors s'acumulen en el procés de *phasing*, que pot interpretar-se com la desincronització en la síntesi de les molècules constitutives dels clústers (23). Els clústers són formats per aproximadament unes 1500 molècules idèntiques, tot i que aquest número pot variar en funció de la densitat de clústers esperada per les característiques específiques del *run*. Si els processos bioquímics d'addició de nucleòtids i de rentat demostrassin una estequiometria perfecta, a cada cicle d'elongació s'afegirien exactament un nucleòtid, un fluoròfor i un bloquejador per cada molècula. Aleshores, després de la presa d'imatge, exactament un fluoròfor i un bloquejador per molècula serien eliminats en el rentat. A la pràctica, però, els rendiments no són perfectes i una petita fracció de molècules no són processades com haurien. Poden saltar-se un cicle d'elongació, o se'ls hi pot afegir més d'una base, per error del bloquejador. Aquest *phasing* resulta en un reconeixement de nucleòtid poc clar i amb una certa tendència a l'error. El progressiu guany de mobilitat en tres dimensions que adquireix el fragment a mesura que creix també dificulta la presa d'imatge. En aquests casos, al nucleòtid se li associa una baixa qualitat.

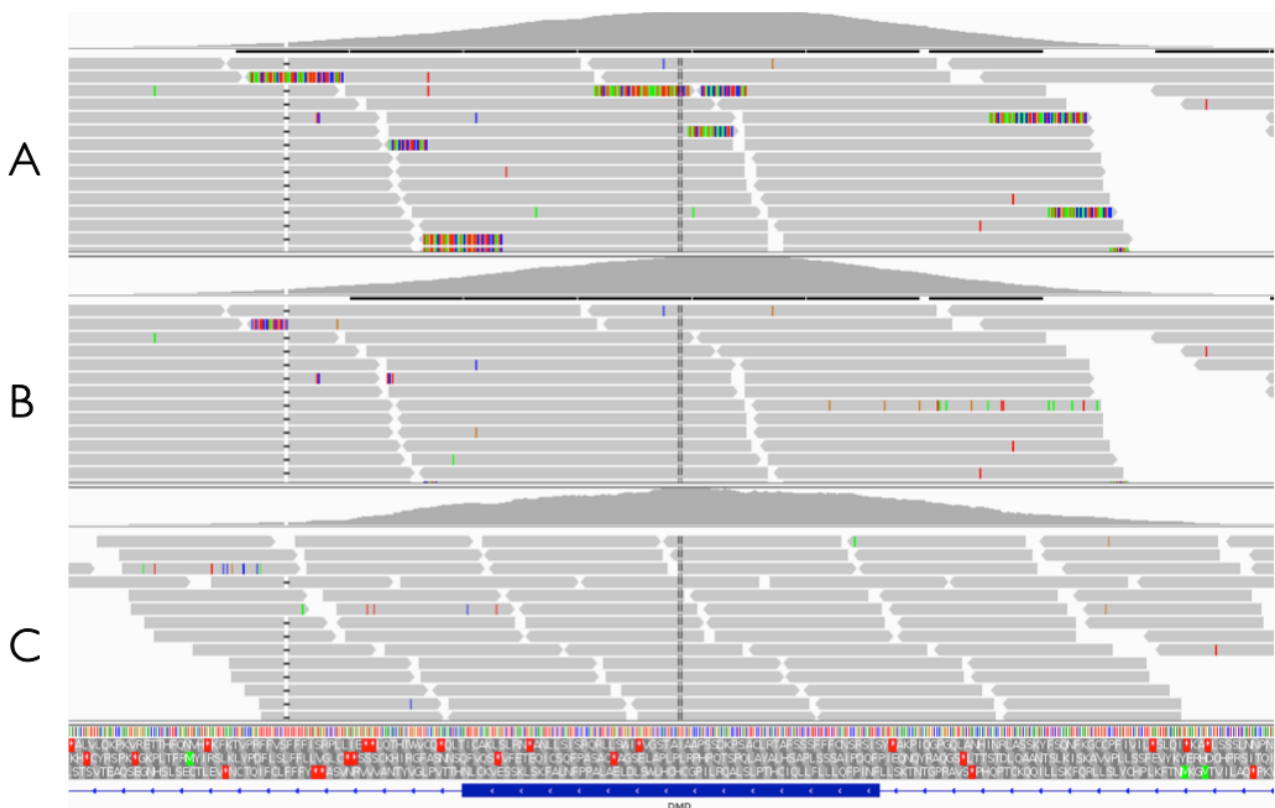


Figura 3-3 | Captures de pantalla del visor genòmic IGV (Broad Institute, Cambridge, USA). S'hi mostren les seqüències alineades a una regió exònica del cromosoma X. **A)** Seqüències de 151 pb sense processar, amb adaptadors i bases de baixa qualitat incloses; **B)** Les mateixes seqüències mostrades a A, un cop processades; **C)** Seqüències de 76 pb amb cartutxos de versió 3, sense processar.

Els adaptadors inclosos a les seqüències són afegits durant la preparació de llibreries, als extrems 5' i 3' de tots els fragments de DNA. La seva funció és essencial, ja que són els encebadors de la seqüenciació *paired-end*; a més, són els oligonucleòtids que immobilitzen les molècules de DNA a la superfície de la cel·la de flux, permetent l'amplificació en pont. També són els codis de barres, o índexs, que determinen la pertinença de cada molècula a la seva respectiva llibreria.

L'insert de DNA, o el fragment "original" que es vol seqüenciar, queda situat entre els dos adaptadors en posició *downstream* en relació a l'adaptador a 5'. Per tant, aquests adaptadors no apareixen mai inclosos a la molècula seqüenciada. Però si es dona el cas en que l'insert és més curt que el número de cicles de seqüenciació, la polimerasa avançarà sobre la seqüència de l'adaptador lligat a 3' (Figura 3-4), incloent els nucleòtids a la seqüència final. Si aquests fragments no són eliminats, el *software* d'alineament pot discriminar aquestes seqüències, o associar-les a qualitats d'alineament baixes. A més, en cas d'acumular-se en localitzacions concretes, la detecció de variants en aquesta seqüència pot complicar-se per la interferència amb el material exogen.

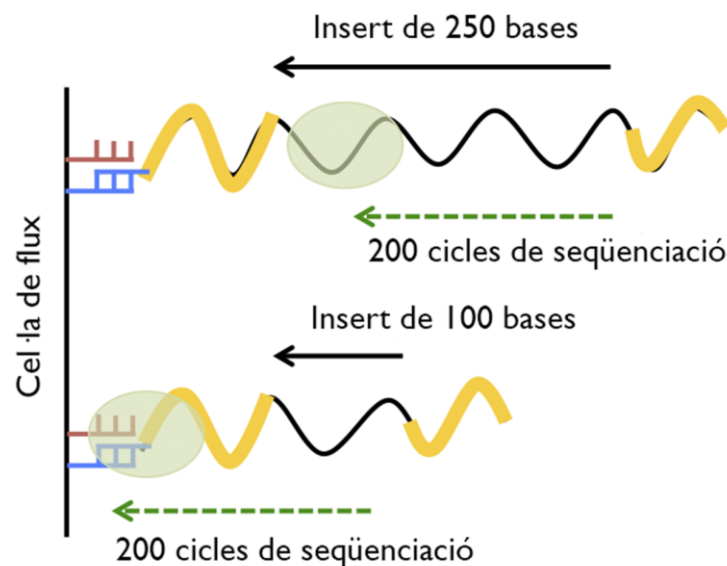


Figura 3-4 | Representació esquemàtica de dos fragments, amb llargades d'insert diferents, immobilitzats a la cel·la de flux del seqüenciador. Els adaptadors es representen en groc. En vermell i blau es representen les seqüències universals per la hibridació dels adaptadors. La polimerasa es representa amb l'oval de color verd.

A la pràctica, tant les bases de baixa qualitat com els adaptadors són exclosos mitjançant l'ús de *softwares* públics o de *scripts* propis (com en aquest treball), que detecten i tallen les seqüències conegudes dels adaptadors comercials i utilitzen una finestra mòbil per la detecció i discriminació d'extrems de seqüències en les que els nucleòtids presentin qualitats baixes.

3.6.2 – Els arxius fastq

La matèria primera de l'anàlisi bioinformàtica són els arxius de seqüències fastq. Com el protocol de preparació de llibreries és *paired-end*, s'obtenen dos arxius fastq per cada mostra, un per cada seqüència del parell. El format d'aquests arxius, que de fet els hi dóna nom, és una variació del format FASTA convencional, dissenyada per l'emmagatzematge d'informació biològica amb anotacions de qualitat (Figura 3-5/A).

A diferència del format FASTA clàssic, el format fastq és format per quatre línies. La primera és l'identificador de la seqüència, i s'inicia amb el símbol '@'; en aquesta línia apareix informació sobre la plataforma que ha generat la seqüència, l'experiment de seqüenciació i la localització espacial del clúster al que pertanyia el fragment dins la cel·la de seqüenciació. Els camps d'informació que s'hi distingeixen són resumits a la Taula 3-2.

Taula 3-2 | Camps d'informació presents a la línia identificadora d'una seqüència fastq.

Element	Descripció
<instrument>	Identificació de l'instrument
<assaig>	Identificació de l'assaig de l'instrument
<cel·la_de_flux>	Identificació de la cel·la de flux
<carril>	Número de carril de la cel·la de flux
<cel·la>	Número de subcel·la on es genera el clúster
<x_pos>	Coordinada X del clúster de lectures
<y_pos>	Coordinada Y del clúster de lectures
<lectura>	Número de lectura dins del parell (1 o 2)
<filtratge>	La lectura ha estat filtrada ('Y' si filtrada; 'N' si no és filtrada)
<control>	La lectura ha estat considerada control (0 si no ho ha estat)
<índex>	Número d'índex de la mostra

A la segona línia hi ha la informació biològica en forma de seqüència. En funció de la plataforma de seqüenciació emprada i l'aplicació biològica de les seqüències, aquestes poden oscil·lar entre les 50 i les 600 bases. La tercera línia fa la funció de separador i en ella hi apareix únicament el símbol '+'. A la quarta i última línia hi apareixen les qualitats (Q) dels nucleòtids seqüenciats, en notació Phred i codificades en llenguatge ASCII; per tant, la segona i quarta línies han de tenir la mateixa llargada.

La notació Phred és la relació logarítmica que indica la probabilitat (P) de que el nucleòtid incorporat s'hagi identificat erròniament. Es calcula mitjançant la següent fórmula:

$$Q = -10 \log_{10} P$$

Per tant, una qualitat $Q=10$ equivaldrà a una probabilitat d'error d'1 entre 10, i a una precisió de base del 90%; una $Q=20$ té associada una probabilitat d'1 entre 100 i una precisió de base del 99%; per una $Q=30$ la precisió és de 99'9%, etcètera. Aquesta mesura de qualitat va utilitzar-se per primera vegada en el Projecte Genoma Humà, com a suport en l'automatització de la seqüenciació del DNA. Va esdevenir un estàndard de qualitat en el camp i va utilitzar-se en comparacions d'eficàcia entre diferents mètodes de seqüenciació (431,432).

En un primer moment, la qualitat es derivava a partir de la comparació dels paràmetres (amplitud, alçada...) extrets dels pics dels electroferogrames amb els pics d'altres bases de seqüències conegudes. En l'actualitat, aquesta notació avalua la qualitat de la identificació de la base efectuada pel seqüenciador a partir de la intensitat de la fluorescència emesa pel clúster. També s'utilitza per la determinació de seqüències consens d'alta precisió.

3.6.3 – Alineament de seqüències

Les seqüències fastq per si mateixes no aporten cap mena d'informació, però un cop analitzades col·lectivament guanyen una gran versatilitat. L'alineament és el procés computacional pel qual es determina la localització més probable de cada seqüència al genoma de referència. També pot tenir com a objectiu l'alineament de seqüències contra genomes de referència d'altres espècies, mentre s'assumeixi la distància evolutiva entre les dues. Fa dues dècades, els algorismes d'alineació de seqüències més utilitzats, com el BLAST (231), estaven explícitament dissenyats per aquesta finalitat.

A diferència dels algorismes primerencs, que requerien de grans bases de dades de seqüències homòlogues a la seqüència problema, els *softwares* d'alineació actuals s'utilitzen generalment per l'alineament de seqüències de DNA d'una espècie d'interès contra el genoma de referència de la mateixa espècie. Aquesta diferència té conseqüències en el disseny i en la implementació final dels algorismes, com per exemple, basar l'assumpció de discordances esperades en la taxa de polimorfisme de l'espècie i en la taxa d'error de la tecnologia emprada, deixant de banda les consideracions de substitucions evolutives. En general, aquests criteris permeten un processament optimitzat de les dades, característica molt desitjada en l'actualitat. Els temps de processament informàtic han passat a ser el coll d'ampolla més important de tot el procés de seqüenciació, donada la quantitat massiva de dades que poden generar-se en un únic experiment (433).

En general, els algorismes d'alineament segueixen un procés multi etapa en el que el primer pas té com objectiu principal i, mitjançant mètodes heurístics, acotar al màxim un grup de localitzacions del genoma de referència en les que sigui més probable trobar la millor alineació per la seqüència problema. Tot i que existeix un ampli ventall de *softwares* que implementen aquests algorismes, la

varietat de mètodes fonamentals en els que es basen és reduïda. Generalitzant, existeixen dos mètodes principals per resoldre la primera etapa de l'anàlisi.

El primer mètode va ser utilitzat per la primera fornada d'alineadors de seqüències curtes. Es basa en la construcció d'una estructura de dades anomenada *hash-table*. Aquesta estructura permet la indexació de dades complexes i no seqüencials de manera que facilita les cerques ràpides a través d'ella. És un mètode especialment apropiat per organitzar seqüències de DNA, un tipus de dades en el que és extremadament poc probable trobar-hi totes les combinacions possibles de nucleòtids i que, alhora, és molt probable que contingui seqüències duplicades. Els inconvenients d'aquest mètode depenen de les dades triades per construir la *hash-table*. Si s'utilitza el genoma de referència per escanejar l'estructura construïda a partir de les seqüències problema, el procés requerirà poca memòria –encara que variable en funció de la quantitat de seqüències i la seva diversitat–, però el temps de computació necessari per escanejar el genoma de referència serà elevat encara que la quantitat de seqüències per alinear sigui reduïda. En canvi, el procés d'escaneig de la *hash-table* construïda a partir del genoma de referència requerirà un ús de memòria constant i elevat, en funció de la mida i la complexitat de la referència, independentment de la quantitat de seqüències que vulguin alinear-se, i el temps computacional dedicat serà inferior (433).

El segon mètode, utilitzat per programes més contemporanis, és el que es serveix de l'índex de Ferragina i Manzini (434), amb el que es millora la rapidesa de la cerca de subseqüències al genoma de referència. La creació de l'estructura de dades subjacent requereix dos passos. Durant el primer, el genoma de referència és reorganitzat mitjançant la transformació de Burrows-Wheeler (BWT) (435), de manera que aquelles seqüències que existeixen múltiples vegades apareixen agrupades en l'estructura de dades. En el segon pas, es crea l'esmentat índex, que serà utilitzat per la ràpida localització de les seqüències al genoma. La creació de l'índex és el pas més intensiu, computacionalment parlant, però allargant poc més el temps de processat s'aconsegueix rebaixar la demanda de memòria. Les implementacions de BWT són molt més ràpides que les seves homòlogues basades en *hash-table* amb un mateix nivell de sensibilitat, i poden ser-ho encara diverses vegades més reduint-la sensiblement; aquesta és precisament la limitació del mètode, l'equilibri entre rapidesa i sensibilitat (433).

Independentment del mètode utilitzat per identificar el subconjunt de regions al genoma de referència i, per tant, exclosa una immensa fracció de regions, en una segona fase de l'anàlisi s'executen algoritmes més lents però precisos, com el de Smith-Waterman (208), que d'altra manera seria impossible aplicar al conjunt de la referència en un termini de temps acceptable.

Les polítiques concretes d'alineament poden distingir algunes implementacions per sobre d'altres. Aquestes són, per exemple, les decisions arbitràries dels desenvolupadors en quant al tractament de seqüències *multimap* (aquelles que poden alinear en diverses localitzacions del genoma). Alguns programes situen les seqüències a qualsevol lloc on aquestes tinguin l'homologia requerida (436), mentre que altres seleccionen aleatòriament una única localització d'alineament i descarten la

resta (42). Per millorar la qualitat dels alineaments, els programes actuals poden utilitzar informació addicional durant el procés. Tal i com s'ha comentat a la introducció, si la llibreria de DNA que es vol alinear és *paired-end*, és possible associar una seqüència a un lloc on hi ha certa ambigüïtat sempre i quan la seva parella hagi estat alineada de manera exacta. Per aquest motiu, l'alineament de seqüències *paired-end* supera als alineaments de seqüència única, tant en sensibilitat com en especificitat (42). Una altra informació susceptible de ser utilitzada és la qualitat de base de les seqüències, de la que ja s'ha parlat anteriorment (433).

En aquest treball s'ha triat l'alineador *Burrows-Wheeler Aligner* (BWA) (42) per diversos motius: és un *software* de codi obert i gratuït, que permet una anàlisi ràpida sense posar en compromís la sensibilitat dels alineaments; a més, permet l'anàlisi tant de seqüències úniques com *paired-end* i mesura la qualitat de l'alineament per cada una de les seqüències processades. L'algoritme té en compte les marques de *soft-clipping*; quan una seqüència no apareix alineada des del primer nucleòtid fins l'últim. Als extrems poden trobar-se subseqüències marcades amb senyals de no-coincidència amb la referència. Aquestes subseqüències, processades de manera adequada, resultaran d'utilitat pel desenmascarament de reestructuracions del DNA. A més, el resultat de BWA és un arxiu SAM (de l'anglès *Sequence Alignment / Map*) i, per tant, fàcilment implementable a l'anàlisi.

3.6.4 – El format SAM / BAM

Els arxius SAM reben el nom del format de text separat per tabulacions amb el que es presenta la informació que contenen. Va ser ideat pel Dr. Heng Li (437) i en l'actualitat és el format més estès per l'emmagatzematge de seqüències de DNA de fins a 128 Mb de llargada, un cop ja han estat alineades a una referència. A la Figura 3-5/B pot veure's una seqüència d'exemple en format SAM.

Les primeres línies, les de l'encapçalament, comencen amb el símbol '@'. En aquest apartat es presenta informació molt diversa: la versió de SAM amb la que s'està treballant; informació sobre el centre, les llibreries i el seqüenciador utilitzat; un resum de les regions en les que s'han trobat alineaments; els programes i les comandes utilitzades per la generació del fitxer i un llarg etcètera. El gruix de la informació, però, el formen les línies de la secció d'alineaments. Cada línia és constituïda per 11 camps obligatoris, resumits a la Taula 3-3, i un número variable de camps opcionals.

no supera els controls de qualitat de la plataforma de seqüenciació. Aquest paràmetre numèric s'utilitza per filtrar de manera ràpida totes les seqüències d'un arxiu SAM amb propietats idèntiques, o per excloure'n les que no resultin d'interès per l'anàlisi. A la Figura 3-5/C es pot veure la representació esquemàtica de la seqüència de la Figura 3-5/B i el seu parell. En aquest exemple, l'etiqueta de bits és 163, que informa de que la seqüència és aparellada i és la número 2, que el parell ha pogut alinear-se de manera adequada, i que la seva parella prové d'un fragment de la cadena *reverse*.

El quart camp informa de la primera posició (la de més a l'esquerra) en que la seqüència ha alineat amb la referència. Pels posteriors anàlisis s'ha de tenir en compte que el sistema de coordenades del format SAM és basat en 1 (en comptes de 0). Això equival a dir que els intervals de coordenades que es presenten són inclusius. És un punt important a considerar ja que si s'identifiquen variants en un sistema de coordenades basat en 1 i després s'anoten o es comparen amb bases de dades basades en 0 poden haver-hi discrepàncies importants.

El sisè camp, el codi CIGAR, informa de les incongruències i les coincidències identificades en l'alineament en comparar-lo amb la referència. Si tots els nucleòtids de la seqüència coincideixen amb la referència, el CIGAR serà format pel número representatiu de la llargada de la seqüència seguit d'una M, de l'anglès *match* (Figura 3-5/B). Les insercions, delecions i els salts són representats amb les lletres I, D i N, respectivament. El *soft-clipping* es representa amb la lletra S.

Per millorar el rendiment computacional durant el processament dels arxius SAM, es va idear la versió binària del mateix format. Els arxius BAM (de l'anglès *Binary Alignment / Map*) contenen la mateixa informació que els SAM, però comprimida, ordenada i indexada. D'aquesta manera s'evita la càrrega excessiva d'informació a la memòria de l'ordinador i s'optimitza la recuperació d'alineaments en regions específiques (437).

3.6.5 – Eliminació de duplicats òptics i de PCR

Els duplicats òptics són seqüències originàries d'un únic clúster que el *software* de detecció ha identificat, per error, en més d'un clúster adjacent. Aquest tipus de duplicats poden ser identificats sense necessitat d'alinear les seqüències, tant sols tenint en compte les coordenades dels clústers presents a la informació de la seqüència fastq.

Els duplicats de PCR, en canvi, s'originen per altres causes. Els clústers, com ja s'ha dit, són formats per la replicació d'un únic fragment de DNA que hibrida a la cel·la de flux per complementarietat amb els adaptadors. Idealment, cada clúster ha de ser únic, i ha d'estar format per un únic fragment de DNA, diferent a tota la resta de fragments generadors de clústers (tant el disseny de les cel·les de flux com els protocols de càrrega de llibreries ho propicien). A la pràctica, però, no hi ha manera de controlar quins fragments hibriden a cada localització de la cel·la de flux, com tampoc es

poden controlar els biaixos en l'amplificació de les molècules originals de DNA. Aquests biaixos, un número elevat de cicles de PCR o els passos dels protocols en els que es redueix considerablement la complexitat de la llibreria (normalment la captura de fragments amb *beads*), incrementen les probabilitats de tenir clústers duplicats, originats de fragments idèntics, que donaran lloc a seqüències duplicades.

Per norma general, percentatge de duplicats de PCR varia en funció de la quantitat de DNA inicial requerida per la preparació de llibreries. A major quantitat de DNA de partida, menor percentatge de duplicats de PCR esperats. Per llibreries de la mida de les utilitzades en aquest treball, quan s'utilitza un protocol que inicialment pren 3 µg de DNA, els duplicats esperats ronden el 3%, mentre que en protocols de 200 ng, el percentatge de duplicats puja fins al 25%.

Quan es processen dades de seqüenciació d'alt rendiment és habitual eliminar els duplicats. Així s'elimina la interferència que poden causar durant l'etapa d'identificació de variants. Si un fragment en el que s'introdueix un error durant l'amplificació és representat en excés, la presència de la base errònia serà proporcionalment molt superior a la que hauria de ser, podent ser detectada com una variant i generant un fals positiu. Per l'efecte contrari, si l'error inclòs al fragment reverteix la variant i el fragment és sobrerepresentat, durant la identificació de variants la presència elevada de fragments sense la variant poden arribar a emascarar-la, causant un fals negatiu.

En aquest treball, per l'eliminació de duplicats, s'ha utilitzat el *software* gratuït i de codi lliure Picard v1.119 (438).

3.6.6 – Detecció de variants puntuals i *indels*

En el procés de detecció de variants puntuals s'interroguen les posicions genòmiques seqüenciades per identificar aquelles en les que l'individu presenta nucleòtids diferents als que apareixen en la seqüència de referència del genoma humà (Figura 3-6). Degut a la gran quantitat de posicions que s'han de revisar, la tasca requereix de mètodes computacionals.

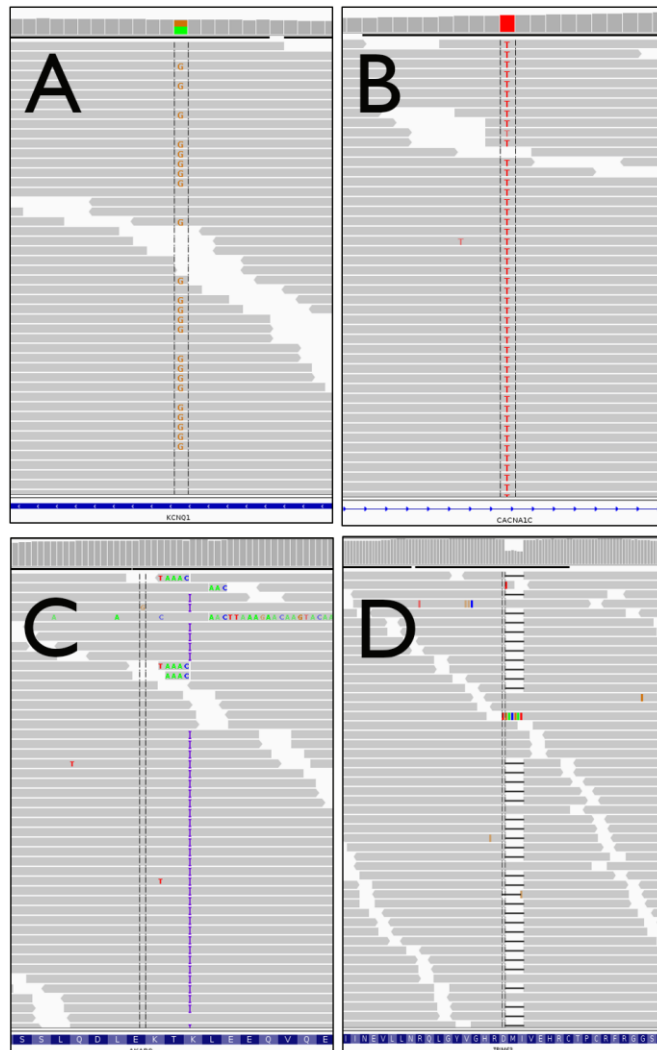


Figura 3-6 | Captures de pantalla del visor genòmic IGV (Broad Institute) en les que es mostra: **A)** Una SNV heterozigota; **B)** Una SNV homozigota; **C)** Una inserció heterozigota; **D)** Una deleció heterozigota.

En un sistema ideal i lliure d'error, el procés seria tant senzill com comptabilitzar a cada posició el número d'ocurrències dels diferents nucleòtids solapants de tots els fragments alineats en aquella localització. Si tots els nucleòtids fossin A, sabríem que el genotip del individu és AA, de la mateixa manera que si tots fossin B, el genotip seria BB. En cas de que s'observés una barreja de dos nucleòtids, el individu seria heterozigot AB. Si en comparar la posició amb la referència es veïés discrepància, s'hauria identificat una variant. No obstant, tal i com ja s'ha donat a entendre, tant la seqüenciació d'alt rendiment de fragments curts com els *softwares* d'alineament no són tecnologies lliures d'error. Un mètode com el plantejat anteriorment no pot aplicar-se, ja que és incapaç de gestionar de manera eficient el soroll present a les dades generades. Si s'intentés actuar de manera conservadora, rebaixant els valors de tall de percentatge d'al·lel alternatiu per a considerar una variant, s'identificaria un número massa elevat de falsos positius. Per altra banda, si els valors de tall fossin laxes seria molt probable que

cometéssim errors de tipus II (o falsos negatius) que en un context clínic han d'evitar-se a tota costa, ja que equival a una diagnosi genètica incompleta del pacient. La complexitat de la qüestió rau en discriminar en quins dels casos les diferències de nucleòtids són degudes a errors comesos durant la seqüenciació o l'alineament. Per fer-ho, els *softwares* utilitzen dues aproximacions diferents: els mètodes heurístics i els mètodes probabilístics.

Els mètodes heurístics basen la identificació de variants en un criteri d'intuïció específic (o en l'experiència) més que en l'observació simple de les mètriques resultants de l'aplicació d'una tecnologia. Aquests criteris poden ser, per exemple: el percentatge de l'al·lel alternatiu per una posició concreta; el número mínim d'ocurrències que ha de presentar l'al·lel alternatiu per ser considerat; la cobertura mínima que ha de tenir una regió per ser analitzada o la qualitat d'alineament que han de presentar les seqüències solapants a la posició d'estudi. Els *softwares* basats en heurística no han estat gaire populars en comparació amb els mètodes probabilístics però, a la pràctica, poden resultar tan robustos i fiables com aquests. Són especialment fiables en assaigs en els que les regions acumulen cobertures elevades, com en la seqüenciació de panells de captura. En funció dels paràmetres de cada *run* i del grau d'optimització dels dissenys de sondes, la mitjana de cobertura per mostra pot arribar a 700-900x. En canvi, quan es seqüencia un exoma, en el que generalment la cobertura mitjana es troba propera a 70x, o un genoma, que en els millors dels casos ronda el 50x, es requereix d'inferència estadística.

Amb els mètodes probabilístics es calcula la probabilitat de que una posició concreta sigui polimòrfica. Aquesta és definida com la probabilitat d'observar unes dades de seqüenciació determinades per una posició (el comptatge d'ocurrències dels diferents al·lells, les qualitats de base associades i la presència d'altres variants a les proximitats de la posició interrogada) donat un conjunt de possibles genotips subjacents (439). Aleshores, mitjançant inferència bayesiana es calculen les probabilitats de cada genotip potencial (440). Les dues eines més utilitzades pel descobriment de variants puntuals a partir de dades de seqüenciació d'alt rendiment són *SAMtools* (437) i el *Genome Analysis ToolKit*, també conegut com GATK (441,442).

En aquest treball, per la identificació de variants, s'ha utilitzat el *software* gratuït i de codi lliure *SAMtools* v1.2 (437), juntament amb *scripts* propis basats en paràmetres heurístics.

3.6.7 – Anotació de variants puntuals i *indels*

L'objectiu de l'etapa d'anotació consisteix en obtenir la major quantitat d'informació possible de les diverses bases de dades o repositoris públics. En aquesta etapa de l'anàlisi s'acumula l'evidència que recolza la patogenicitat de les variants identificades. A partir d'aquesta informació, les variants podran ser classificades i presentades al facultatiu, que decidirà en base a criteris clínics quines són les mesures a prendre amb el pacient. Aquesta informació pot ser de caire poblacional, com la freqüència de l'al·lel minoritari –MAF– (de l'anglès *Minor Allele Frequency*) observada als diferents subgrups de

poblacions, o pot aportar dades sobre el canvi que provoca, com el canvi d'aminoàcid o l'alteració del patró de *splicing*. També es recull informació sobre articles o projectes relacionats amb les variants, o a quines malalties ha estat associada, ja sigui per estudis funcionals o per GWAS. Addicionalment en la mateixa etapa d'anotació també s'han utilitzat els predictors de patogenicitat *in silico* PolyPhen-2 (443), Provean (444) i *Mutation Taster 2* (445).

En aquest treball les variants puntuals van ser anotades amb la informació continguda als repositori de dbSNP –Build 142– (446), l'*Exome Sequencing Project* –ESP6500SI-V2– (447), el *1000 Genomes Project Phase 3* (448), l'*Exome Aggregation Consortium* –ExAC– v.0.3 (449), la *Human Gene Mutation Database* –HGMD– (450), ClinVar (451) i ENSEMBL (428). També es va utilitzar la informació d'una base de dades del propi Centre de Genètica Cardiovascular formada en base a la informació obtinguda en els últims anys a partir de la seqüenciació massiva de pacients.

3.6.8 – Detecció i anotació de variants estructurals

A l'apartat de resultats s'explica detalladament l'algorisme bioinformàtic de detecció i anotació de variants estructurals que s'ha desenvolupat en aquesta tesi. Per altra banda, a la introducció es parla dels mètodes de detecció de variants estructurals tradicionals i contemporanis.

3.7 – Classificació de les variants genètiques

Les SNVs i els *indels* van ser classificats com Variants Patogèniques (VP), Variants Probablement Patogèniques (VPP), Variants Probablement Benignes (VPB), Variants Benignes (VB) o Variants de Significat Incert (VSI); d'acord amb les recomanacions de l'*American College of Medical Genetics and Genomics* i l'*Association for Molecular Pathology* (422). Els criteris de classificació es basen en l'acumulació d'evidència que recolzi una de les etiquetes llistades anteriorment. Les combinacions d'evidències que conclouen la classificació de les variants puntuals apareixen llistades a la Taula 3-4.

Taula 3-4 | Taula de classificació de variants en funció de l'evidència de patogenicitat acumulada.

Classificació	Combinació de criteris de sospita
Variant Patogènica (VP)	<p>(i) 1 Molt elevada [PME] (a més de)</p> <p>(a) ≥ 1 Elevada [PE1-PE4] (o bé)</p> <p>(b) ≥ 2 Moderada [PM1-PM6] (o bé)</p> <p>(c) 1 Moderada [PM1-PM6] i 1 Suport [PS1-PS5] (o bé)</p> <p>(d) ≥ 2 Suport [PS1-PS5]</p> <p>(ii) ≥ 2 Elevada [PE1-PE4] (o bé)</p> <p>(iii) 1 Elevada [PE1-PE4] (a més de)</p> <p>(a) ≥ 3 Moderada [PM1-PM6] (o bé)</p> <p>(b) 2 Moderada [PM1-PM6] (a més de) ≥ 2 Suport [PS1-PS5] (o bé)</p> <p>(c) 1 Moderada [PM1-PM6] (a més de) ≥ 4 Suport [PS1-PS5]</p>

Variant Probablement Patogènica (VPP)	<p>(i) 1 Molt elevada [PME] (a més de) 1 Moderada [PM1-PM6] (o bé)</p> <p>(ii) 1 Elevada [PE1-PE4] (a més de) 1–2 Moderada [PM1-PM6] (o bé)</p> <p>(iii) 1 Elevada [PE1-PE4] (a més de) ≥ 2 Suport [PS1-PS5] (o bé)</p> <p>(iv) ≥ 3 Moderada [PM1-PM6] (o bé)</p> <p>(v) 2 Moderada [PM1-PM6] (a més de) ≥ 2 Suport [PS1-PS5] (o bé)</p> <p>(vi) 1 Moderada [PM1-PM6] (a més de) ≥ 4 Suport [PS1-PS5]</p>
Variant Probablement Benigna (VPB)	<p>(i) 1 Elevada [BE1-BE4] (a més de) 1 Suport [BS1-BS5] (o bé)</p> <p>(ii) ≥ 2 Suport [BS1-BS5]</p>
Variant Benigna (VB)	<p>(i) 1 Autònoma [BA] (o bé)</p> <p>(ii) ≥ 2 Elevada [BE1-BE4]</p>
Variant de Significat Incert (VSI)	<p>(i) Exclusió de criteris (o bé)</p> <p>(ii) Criteris contradictoris</p>

S'associa una sospita de patogenicitat molt elevada [PME] a les variants radicals (*nonsense*, canvis de pauta de lectura, variants que localitzen al codó d'inici o a ± 2 d'un lloc de *splicing*) en un gen en el que la pèrdua de funció sigui un mecanisme conegut com causant de la malaltia. Les evidències que recolzen una sospita de patogenicitat elevada [PE] són: que la variant provoqui un canvi d'aminoàcid idèntic a una variant per la que prèviament s'ha establert la seva patogenicitat, independentment del canvi de nucleòtid [PE1]; que la variant sigui *de novo* en un pacient afectat i sense història familiar. S'ha de confirmar que tant el pare com la mare no en són portadors [PE2]; que es disposi d'estudis funcionals *in vitro* o *in vivo* ben establerts per la variant i que recolzin un efecte deleteri pel gen o pel seu producte [PE3]; que la prevalença de la variant als individus afectats sigui significativament superior en comparació amb l'observada als controls [PE4]. Es considera sospita de patogenicitat moderada [PM] quan la variant localitza en un *hot spot* mutacional i/o en un domini funcional ben establert, com per exemple el centre actiu d'un enzim [PM1]; que la variant sigui absent en controls als repositoris d'*Exome Sequencing Project*, *1000 Genomes Project* i ExAC [PM2]; per desordres recessius, que la variant sigui detectada en *trans* juntament amb una variant patogènica [PM3]; que la llargada de la proteïna canviï com a resultat: a) d'una inserció o deleció que no faci variar la pauta de lectura en una regió sense repeticions o b) d'una variant que causi la pèrdua del codó *stop* [PM4]; que es tracti d'una variant *missense* no descrita prèviament en un aminoàcid en el que un canvi *missense* diferent hagi estat prèviament considerat com a patogènic [PM5]; que s'assumeixi que la variant sigui *de novo*, però no pugui confirmar-se l'absència en la mare i en el pare [PM6]. Les variants que no compleixen els criteris anteriors però per les que es té sospita d'un rol patogènic [PS] són aquelles que segreguen en diversos membres afectats de la mateixa família en un gen associat a la malaltia [PS1]; les variants *missense* en un gen que té una taxa baixa de variació *missense* benigna i en el qual les variants són un mecanisme de patogenicitat comú [PS2]; les variants per les que diversos predictors *in silico* prediuen un efecte deleteri pel gen o pel seu producte [PS3]; variants en pacients en

els que el fenotip o la història familiar és altament específic per una malaltia amb una única etiologia genètica [PS4]; o variants descrites com a patogèniques per fonts reputades, però per les que sigui impossible pel laboratori dur a terme una avaluació independent [PS5].

Els criteris d'associació de sospita de benignitat són els següents: una variant amb freqüència poblacional superior al 5% a la base de dades d'*Exome Sequencing Project*, *1000 Genomes Project* o ExAC és considerada benigna [BA]; les variants amb sospita de benignitat elevada [BE] són aquelles amb una freqüència al·lèlica major de l'esperada pel desordre [BE1]; les variants identificades en un individu sa i adult per un desordre recessiu (sent la variant homozigota), dominant (sent la variant heterozigota) o bé en un desordre lligat al cromosoma X (sent la variant hemizigota), amb penetrància completa i a una edat primerenca [BE2]; les variants amb estudis funcionals *in vivo* o *in vitro* ben establerts en els que no es conclouï un efecte deleteri en la funció de la proteïna o que causi un *splicing* alternatiu [BE3] i les variants que no segreguin entre els membres afectats d'una mateixa família [BE4]. Les variants que no compleixen els criteris anteriors, però per les que es té sospita que d'un rol benigne [BS] són: variants *missense* en un gen en el que es coneixen variants truncades que causen la malaltia [BS1]; les variants observades en *trans* juntament amb una variant patogènica per un trastorn dominant amb penetrància total o observada en *cis* amb una variant patogènica en qualsevol patró d'herència [BS2]; les insercions o delecions que no variïn la pauta de lectura en una regió repetitiva sense funció coneguda dins del gen [BS3]; les variants per les que diversos predictors *in silico* no prediuen cap impacte en el gen o el seu producte [BS4]; les variants trobades en casos clínics amb una base molecular alternativa per la malaltia que no impliqui el gen o el seu producte [BS5]; les variants descrites com benignes per fonts reputades però per les que sigui impossible pel laboratori dur a terme una avaluació independent [BS6]; les variants silents per les que els algorismes de predicció de *splicing* alternatiu no suggereixin cap impacte en la seqüència consens ni la creació d'un nou lloc de *splicing* i, a més, el nucleòtid no sigui altament conservat [BS7].

En quant a les CNVs, al no existir unes recomanacions públiques sobre com classificar aquestes variants quan són petites i intragèniques, els criteris escollits van ser propis. Una CNV va ser considerada patogènica (VP) si: a) havia estat prèviament publicada com a patogènica i associada a la malaltia del pacient (o pels casos de MSC, si la malaltia reportada és compatible amb un cor estructuralment normal –aquesta observació aplica per les següents classificacions–); b) si era una deleción en un (o d'un) gen pel que s'hagi establert la pèrdua de funció com un mecanisme de patogenicitat en la malaltia del pacient; c) si era una duplicación intragènica en tàndem (que no involucrava l'últim exó del gen) en un gen pel que la pèrdua de funció s'ha establert com mecanisme de patogenicitat en la malaltia del pacient; o d) era una duplicación d'un gen sencer pel que s'ha establert la triplosensitivitat com un mecanisme causal de la malaltia. Una CNV va ser considerada com probablement patogènica (VPP) si: a) era una deleción en (o d'un) gen associat amb la malaltia del pacient o una duplicación intragènica (que no involucra l'últim exó del gen) en un gen associat amb la malaltia, i la variant no s'havia identificat entre els controls dels repositoris públics de DGV (de l'anglès *Database of Genomic Variants*) o *1000*

Genomes Project; o b) la CNV segrega en més de 5 familiars afectats. Les CNVs van ser considerades probablement benignes (VPB) si: a) s'havien identificat en més de 10 individus de la població general (per duplicacions o delecions de gen sencer es van considerar en la comparació casos de la població general que involucraven gens contigus), o b) la CNV no segrega entre els afectats de la família. Van ser considerades benignes (VB) aquelles CNVs prèviament descrites com a benignes. Les CNVs que no complien els criteris anteriors van ser classificades com variants de significat incert (VSI).

3.8 – Confirmació de variants

Degut a les limitacions de les tècniques de seqüenciació d'alt rendiment, fins al moment i sempre dins de l'àmbit clínic es precisa de la confirmació de totes les variants identificades mitjançant una tècnica *gold standard* que corrobore la presència de les variants detectades bioinformàticament. Tot i que comencen a publicar-se estudis que desafien aquesta pràctica establerta (452), en aquest treball les variants van confirmar-se amb tres tècniques diferents, en funció del tipus de variant en cada cas.

3.8.1 – Seqüenciació Sanger

Les variants puntuals, SNVs i *indels*, van ser confirmades mitjançant seqüenciació Sanger convencional. Les regions d'interès van ser amplificades per PCR (*Verities PCR*, Applied Biosystems). Un cop purificat, el producte de la PCR va ser seqüenciat en dues direccions utilitzant el kit comercial *BigDye Terminator v3.1 Cycle Sequencing Kit* (Applied Biosystems) i carregat en un seqüenciador d'electroforesi capil·lar *ABI3130XL Genetic Analyzer* (Applied Biosystems). Les seqüències obtingudes van analitzar-se visualment mitjançant el *software SeqScape v2.5* (Life Technologies).

Les regions que després de ser seqüenciades presentessin una cobertura inferior a 30x també van ser seqüenciades per Sanger

3.8.2 – MLPA

Sempre que hi hagués disponible un kit comercial, les CNVs identificades van ser confirmades per MLPA. Els diferents assaigs de quantificació van dur-se a terme d'acord amb els protocols descrits als kits comercials (MRC-Holland, Àmsterdam, Països Baixos). Després de l'etapa de PCR, la seqüenciació per electroforesi capil·lar va dur-se a terme en un seqüenciador *ABI3130XL Genetic Analyzer* utilitzant *LIZ500® Size-Standard* (els dos d'Applied Biosystems). Els resultats van ser analitzats mitjançant el *software Coffalyser.Net* (MRC-Holland).

3.8.3 – PCR quantitativa

Quan no hi havien kits comercials de MLPA, les CNVs van confirmar-se mitjançant qPCR. Els assaigs es van fer mitjançant *QuantStudio 7 Flex System* i es va utilitzar el kit *PowerUp™ SYBR® Green Master Mix* (els dos de Life Technologies). Els resultats van ser analitzats amb el programa *QuantStudio Real-Time PCR Software v1.2*, de la mateixa empresa.

IV. Resultats i discussió

4.1 – Caracterització i discussió dels resultats de la seqüenciació d'alt rendiment

Les 2073 mostres que formen la cohort d'estudi van ser seqüenciades al Centre de Genètica Cardiovascular entre els anys 2012 i 2017. Les llibreries genòmiques van ser preparades en 202 tandes independents, i seqüenciades en un total de 261 *runs* de MiSeq. A la Taula 4-1 s'hi resumeix el número de mostres seqüenciades per cada panell. També s'hi recullen algunes mètriques descriptives de la distribució de cobertures, útils per poder caracteritzar la qualitat de la seqüenciació que s'ha dut a terme.

Taula 4-1 | Mètriques descriptives relatives a la seqüenciació de la cohort d'estudi.

Panell	Mostres	Call Rate 30x	Mitjana	Seqüències Totals*	Seqüències RI*	Exons <30x	Enriquiment (%)	Duplicats (%)
55	772	99,6±0,55	941±280	7,1±3,5	3,0±0,9	12±7	43±11	6,5±0,4
78	714	99,9±0,20	543±145	5,7±1,1	2,7±0,7	7±7	49±7	4,6±0,3
118	103	99,8±0,14	778±126	10,3±1,7	5,1±1,0	8±9	50±6	3,5±0,2
85	383	99,9±0,06	320±90	4,6±0,8	2,0±0,5	3±5	46±8	3,2±0,3
147	101	99,7±0,17	266±53	5,5±1,0	2,4±0,7	19±14	46±9	3,1±0,2

*En milions de seqüències.

Inicialment, les llibreries es preparaven en tandes de 8 a 10. A mesura que s'obtenien resultats amb la qualitat desitjada, el número de mostres va incrementar-se fins a 16, coincidint amb la capacitat del suport magnètic utilitzat durant el protocol de preparació de llibreries. De manera paral·lela, els pools de MiSeq eren constituïts inicialment per 5 mostres a seqüenciar amb el panell de 55 gens. Amb la comprovació de la potència i el progressiu grau d'optimització dels panells, aquest número va incrementar fins a 14 mostres amb el panell de 85 gens, o les 8 mostres que formen els pools de seqüenciació actuals amb els panells de 118 o 147 gens.

Fins al moment, els panells de gens que més temps s'han fet servir (i amb els que s'han seqüenciat més mostres) han estat el de 55 i el de 78 gens (772 i 714 mostres, respectivament). El segon va desenvolupar-se al cap d'un any i mig que el de 55 gens, conjuntament amb el de 118 gens (103 mostres seqüenciades). Aquest últim ha estat menys sol·licitat ja que s'associa a un perfil específic del servei de seqüenciació, relacionat amb la recerca, o per la seqüenciació de pacients amb SM i DATA. El panell de 85 gens és en actiu des de finals de 2015 i és l'evolució dels panells de 55 i 78 gens. Paral·lelament, el de 147 gens és l'evolució del panell de 118 gens.

Els dos paràmetres principals per caracteritzar qualsevol resultat de seqüenciació per captura són el **percentatge d'enriquiment** (o la relació que s'estableix entre les seqüències totals i aquelles que han hibridat a les regions d'interès), i la **densitat de cobertura**. L'enriquiment assolit per les mostres, amb independència del panell utilitzat, és entorn al 40-50%, amb un màxim proper al 60%

(per aquelles poques mostres que hi hagin arribat.) Per altra banda, la distribució de cobertures és dependent del número de mostres seqüenciades en un mateix run i de la mida del panell. A mesura que incrementa el número de mostres o de regions a cobrir, l'*output* potencial del seqüenciador (reflectit en el número total de seqüències per regió) ha de repartir-se entre més regions diana. Per tant, la cobertura disminueix de manera proporcional (Figura 4-1/A). Així doncs, les cobertures més elevades són les que presenten les mostres seqüenciades amb el panell de 55 gens (Taula 4-1), amb una cobertura mitjana per mostra de 941x, seguides per les del panell de 118 gens (778x), el panell de 78 gens (543x), el de 85 gens (320x) i el de 147 gens (266x).

La informació que aporten els paràmetres de densitat de cobertura i d'enriquiment es complementa amb el **call rate**, el **número d'exons perduts** i el **percentatge de duplicats òptics i de PCR**. El *call rate* informa de la fracció total de bases pertanyents a les regions d'interès que han assolit una cobertura específica. Així, que la mitjana del *call rate* a 30x per les mostres seqüenciades amb el panell de 55 gens sigui d'un 99,6% equival a dir que, de mitjana, un 0,4% de les bases que volien seqüenciar-se no han assolit la cobertura mínima desitjada de 30x. El *call rate* a 30x és un paràmetre important en l'àmbit de la diagnosi genètica per concloure si una mostra s'ha seqüenciat de manera satisfactòria. Qualsevol variant detectada ha d'estar coberta com a mínim a 30x per ser validada. De manera similar, aquells exons que no s'hagin cobert al 100% a 30x seran considerats com "perduts" i hauran de seqüenciar-se mitjançant seqüenciació Sanger per assegurar la presència o l'absència de variants genètiques. Els percentatges de *call rate* són elevats i varien en un marge estret. No hi ha cap panell pel que la mitjana del *call rate* a 30x resulti inferior al 99%. Això es tradueix en un número d'exons perduts a 30x molt moderat (menys de 20, de mitjana, per panell). A les mostres seqüenciades amb el panell de 85 gens s'observa un percentatge de *call rate* a 30x del 99.9%, amb només 3 exons perduts, de mitjana. Aquestes dades donen a entendre que el panell de 85 gens té el disseny de sondes més optimitzat i efectiu, capaç d'assolir les cobertures més homogènies dels cinc panells utilitzats (Figura 4-1/B).

Els percentatges de duplicats òptics i de PCR també són baixos. Varien en un rang molt estret (entre un 3-6%). Tal i com s'ha explicat a l'apartat 3.6.5 de Material i Mètodes, això és degut a que (independentment del panell utilitzat) les llibreries genòmiques van preparar-se sempre amb la mateixa quantitat de DNA inicial –3 µgr–.

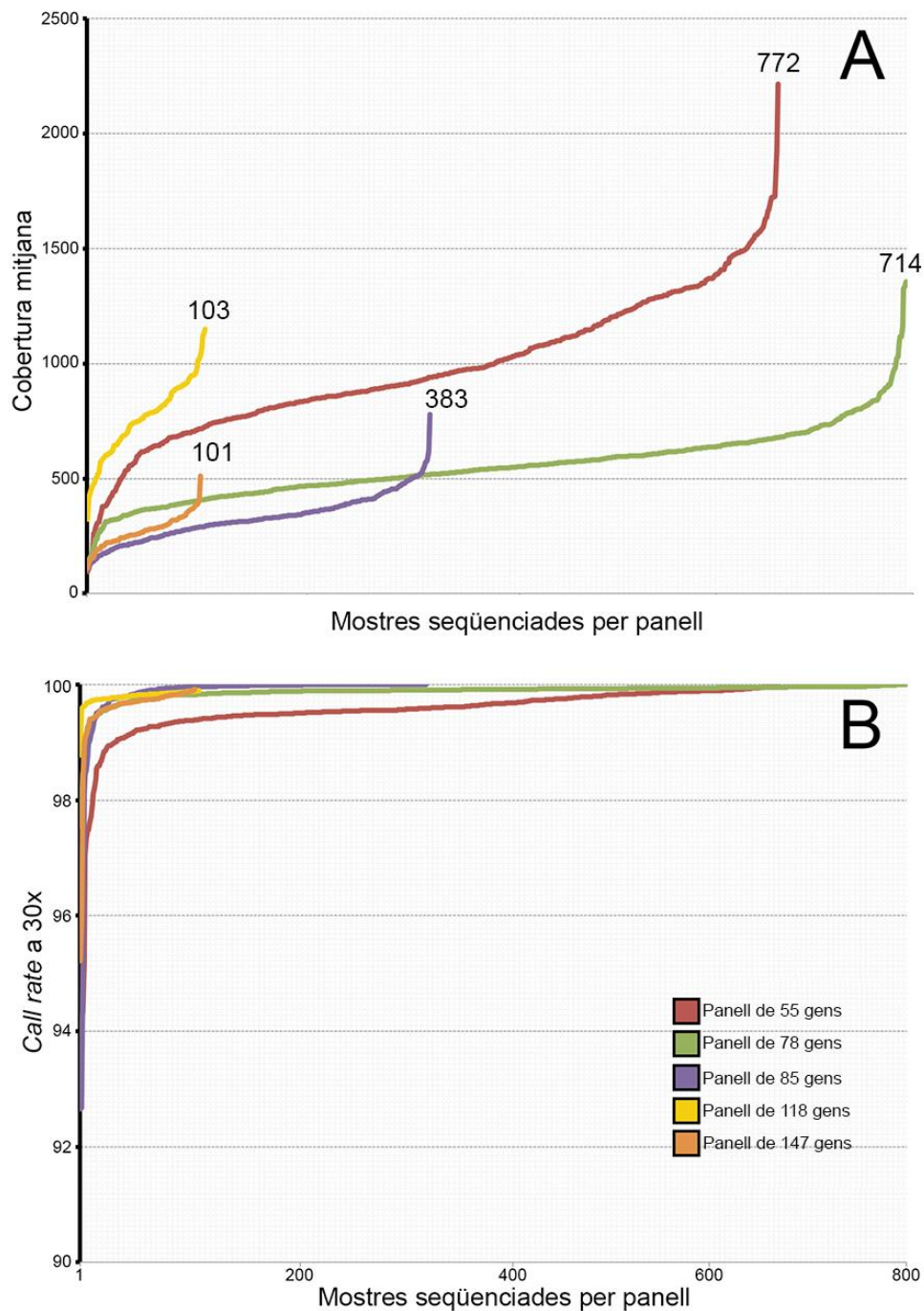


Figura 4-1 | A) Cobertures mitjanes assolides per les mostres de la cohort. Els números informen de la quantitat total de mostres seqüenciades per panell; **B)** Call rates a 30x de les mostres seqüenciades. Cada panell és representat per un color diferent.

4.1.1 – Evolució dels panells de seqüenciació

En aquest apartat es pretén fer una descripció del procés d'evolució dels dissenys de sondes per cada un dels panells de seqüenciació i dels motius lògics que hi han al darrere de cada millora. Com s'ha exposat a la declaració de contribució (apartat 3.1), la tasca d'optimització dels dissenys de sondes va correspondre al Dr. Carles Ferrer-Costa. Els criteris pel disseny de sondes s'expliquen a l'apartat 3.4.2 Els dissenys de captura optimitzats, juntament amb l'algoritme de detecció de CNVs que s'ha

desenvolupat en aquesta tesi (apartat 4.2) constitueixen el mètode de detecció de CNVs utilitzat en el cribratge de la cohort d'estudi.

El sistema utilitzat pel disseny de sondes (*SureSelectXT Target Enrichment, Agilent Technologies*) accepta una quantitat màxima de 57000 sondes de 120 bases de RNA complementari per projecte. Per l'optimització del primer disseny de sondes, el del panell de 55 gens (1201 exons –243,89 Kb–), es va considerar que les regions a capturar que presentessin un contingut GC més elevat podrien ser les més dificultoses de seqüenciar de manera apropiada. A la Figura 4-2/A es representa la cobertura acumulada de les sondes del disseny, de manera que s'hi reflexa la densitat de sondes per les diferents regions en funció del contingut GC. En verd clar es mostra la cobertura màxima per regió i en verd fosc la cobertura mitjana. El contingut GC es representa amb la línia de color groc. Es pot comprovar com la quantitat de sondes dissenyades es manté estable per les regions amb continguts de GC baixos (26 exons amb GC < 35%) o centrals. La densitat de sondes puja, però, quan el percentatge GC s'aproxima al 60% (160 exons amb GC > 60%).

A la Figura 4-2/B es representa (en blau) la cobertura mitjana de cada regió per totes les mostres seqüenciades amb aquest panell. Tot i que les cobertures assolides són altes, en comparació amb la resta de dissenys amb un major nivell d'optimització, es pot veure una marcada davallada de la cobertura en aquelles regions amb continguts GC elevats (superiors al 60%). Això posa en evidència que tot i que es va dissenyar una major quantitat de sondes per aquestes regions, les sondes encara resultaven insuficients per obtenir unes cobertures homogènies i comparables amb la resta de les obtingudes per les altres regions.

A la Figura 4-2/C s'hi representa la mitjana d'exons perduts a 30x per mostra. En les mostres seqüenciades per aquest panell és freqüent trobar sense cobrir de manera òptima una gran quantitat d'exons número 1. Els exons inicials acostumen a presentar uns percentatges de GC més elevats que la resta. Aquesta és una de les formes que té el genoma de regular l'expressió gènica. A percentatges elevats de GC cal una major energia per separar les cadenes de DNA. Sovint aquesta energia vindrà donada per l'efecte de factors de regulació en *cis* o en *trans*, minimitzant la possible expressió basal (453,454). Alguns dels exons que més freqüentment presenten cobertures baixes són: l'exó número 1 de *DSC2* (73%), *DSG2* (70%), *KCNQ1* (76%), *SGCB* (73%) o *HCN4* (73%), entre d'altres.

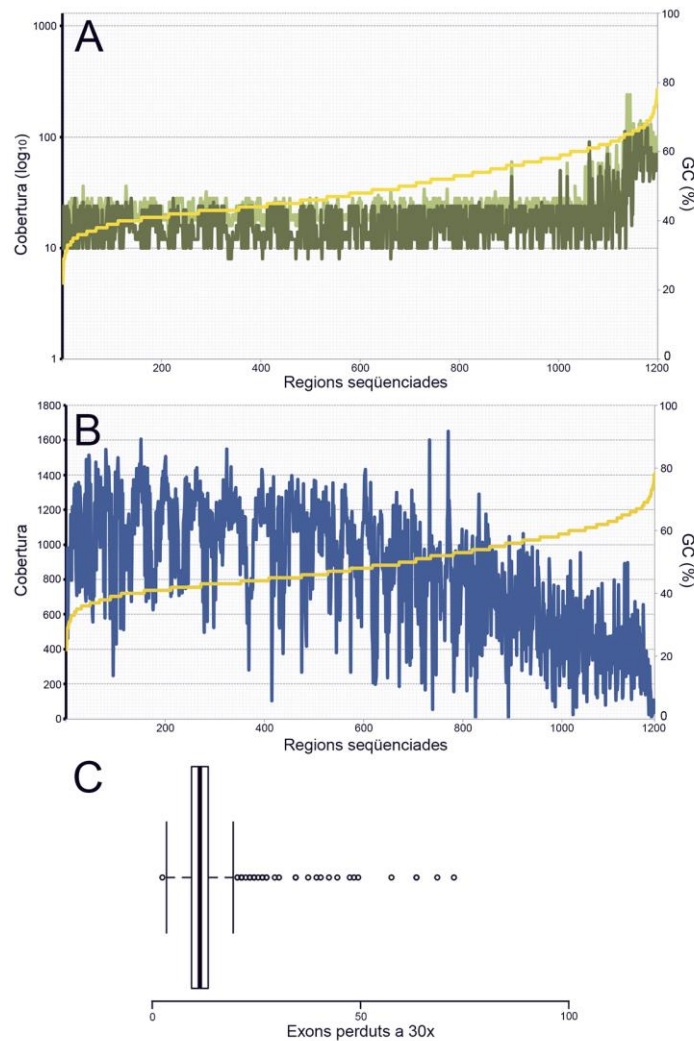


Figura 4-2 | Panell de 55 gens; **A)** Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertures mitjana. A les regions amb continguts GC elevats es distribueixen una major quantitat de sondes; **B)** Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). Es pot apreciar com les regions amb continguts GC elevats assolixen molta menys cobertures que la resta; **C)** Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.

Gràcies als resultats obtinguts amb les mostres seqüenciades amb el panell de 55 gens, els dissenys posteriors dels panells de 78 i 118 gens, desenvolupats en paral·lel (1569 exons –300,87 Kb– i 1723 exons –328,58 Kb–, respectivament), van poder-se optimitzar. Es van redistribuir sondes de regions amb continguts GC propers al 50% i que havien assolit nivells de cobertures elevats cap a regions que havien demostrat ser més difícils de seqüenciar de manera òptima (Figura 4-3/A-B). En general, aquelles regions amb un GC a partir del 60% o similars (194 i 355 exons amb GC > 60%, respectivament) van rebre una major densitat de sondes, corregint així la davallada de cobertures que s'apreciava pel disseny de 55 gens (Figura 4-3/C-D). Per primera vegada, els dissenys inclouen gens amb un percentatge de GC generalitzadament baix (69 i 80 exons amb GC < 35%, respectivament). Els gens *TRDN* i *AKAP9* són casos paradigmàtics, amb un contingut GC mig proper al 35%. Tot i que les regions amb GCs baixos

estiguin en minoria, a l'extrem esquerre de la Figura 4-3/C-D es pot observar la davallada de cobertura resultant de no haver redistribuït cap a aquestes regions una quantitat superior de sondes. Aquestes regions acaben assolint cobertures inferiors i molt variables al llarg dels diferents *runs*.

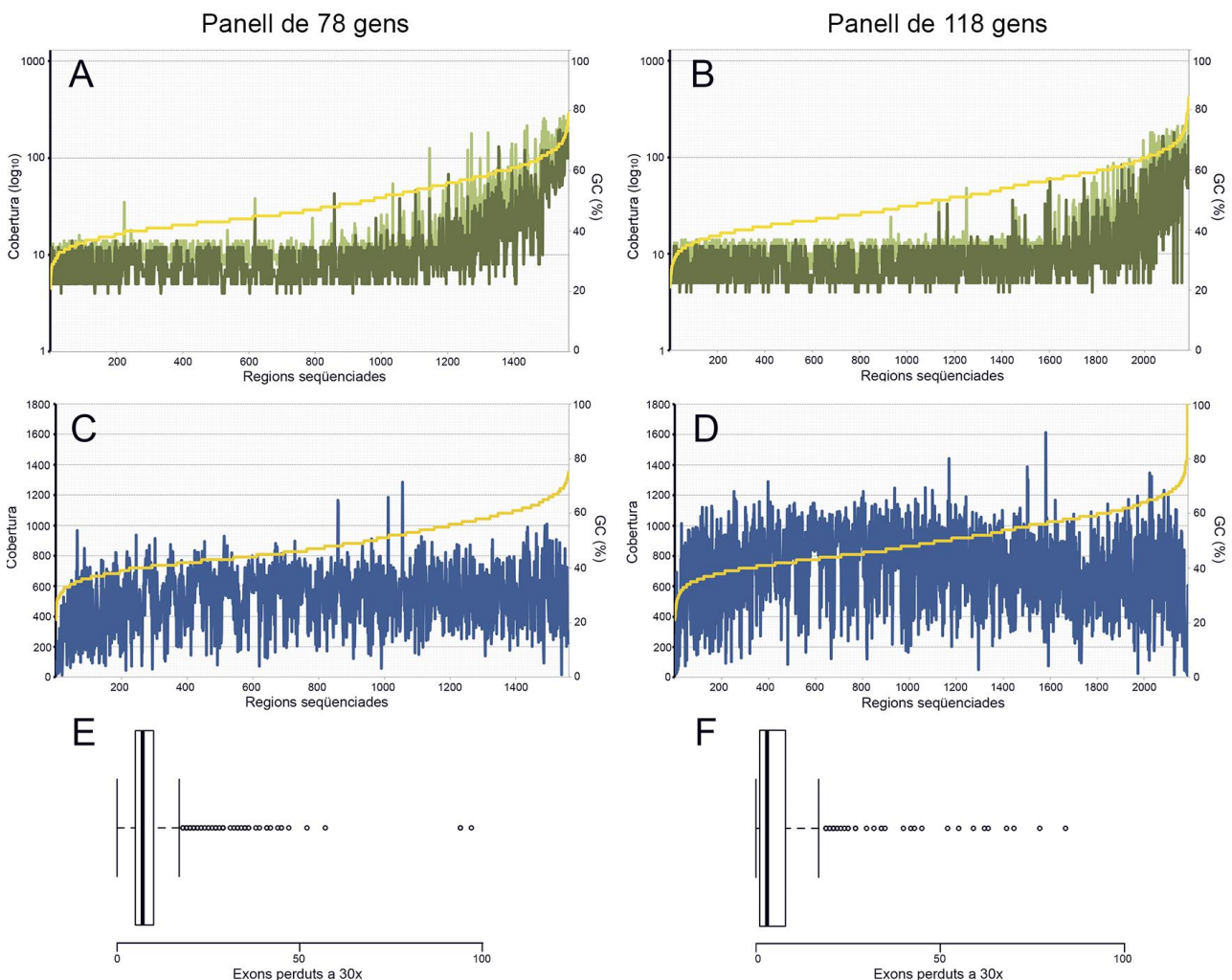


Figura 4-3 | Panell de 78 i 118 gens, respectivament. **A-B)** Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertura mitjana. S'aprecia l'increment de la densitat de sondes en les regions amb continguts GC elevats; **C-D)** Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb els panells de 78 i 118 gens, respectivament, en funció del contingut GC (groc). Les cobertures assolides per les regions d'elevat contingut GC són acceptables, però les regions amb un GC reduït presenten menys cobertura que la resta; **E-F)** Mitjanes d'exons perduts a 30x per les mostres seqüenciades amb el panell.

Per tal de solucionar aquest problema, al disseny de sondes del panell de 85 gens (2190 exons –403,64 Kb–, dels quals 73 exons presenten un GC < 35% i 271 tenen un GC > 60%) es va redistribuir una quantitat considerable de sondes cap a regions de baix contingut GC. A la banda esquerra de la Figura 4-4/A s'aprecia la major densitat als dos extrems de la gràfica. Amb la nova distribució de sondes, la cobertura s'assoleix amb una marcada homogeneïtat al llarg de totes les regions (Figura 4-4/B), fet

que comporta un menor número d'exons perduts a 30x (Figura 4-4/C).

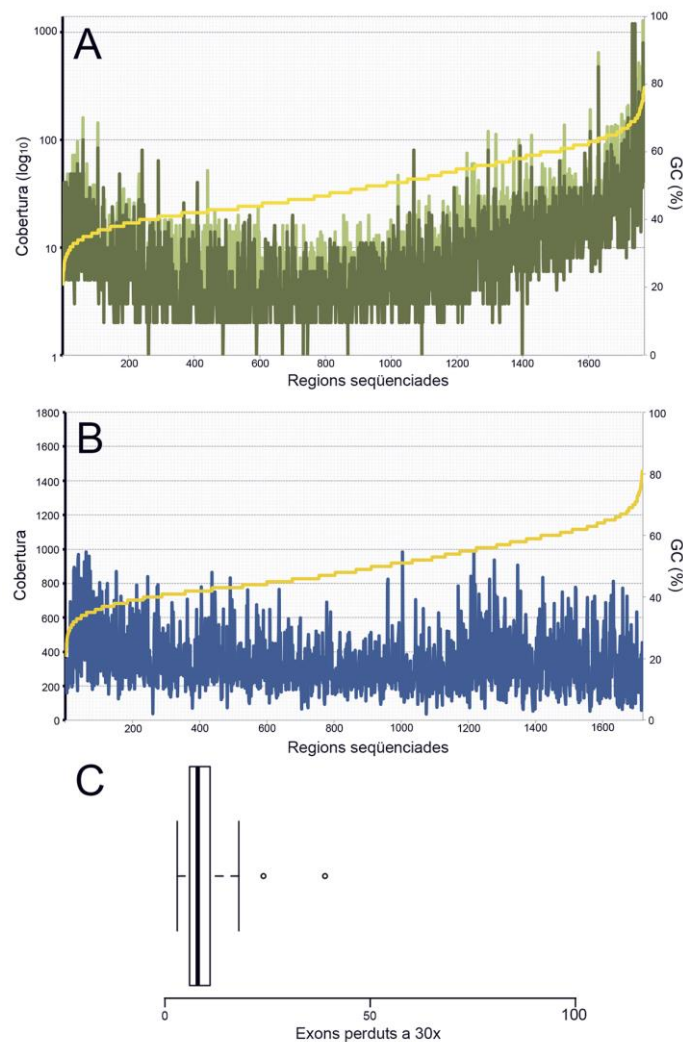


Figura 4-4 | Panell de 85 gens. **A)** Densitat de sondes per regió en funció del contingut GC (groc). En verd clar es mostren les cobertures màximes per regió, i en verd fosc la cobertura mitjana. En aquest disseny s'han redistribuït sondes cap a les regions amb GCs extrems ($\leq 35\%$ i $\geq 60\%$); **B)** Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). Pot apreciar-se una adquisició de cobertura homogènia per totes les regions seqüenciades, independentment del GC; **C)** Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.

El panell de 147 gens (2672 exons –442,13 Kb–) és el més ambiciós de tots els dissenys, després del de 118 gens. Un 21% de les regions incloses presenten un contingut GC superior al 60% (476 exons) i inferior al 35% (97 exons). La distribució diferencial de sondes segueix el mateix criteri que el triat pel panell de 85 gens (Figura 4-5/A), però per primera vegada es detecta una davallada en les cobertures de regions amb continguts GC propers al 50% (Figura 4-5/B), fet que no afavoreix l'adquisició de cobertures homogènies. Inicialment, el número de mostres seqüenciades per *run* amb aquest panell eren 10. Tot i així, degut a les cobertures ajustades assolides (Taula 4-1), la mida del run es va disminuir a 8. D'aquesta manera es volen ampliar sensiblement les cobertures i evitar la pèrdua d'homogeneïtat de les mateixes.

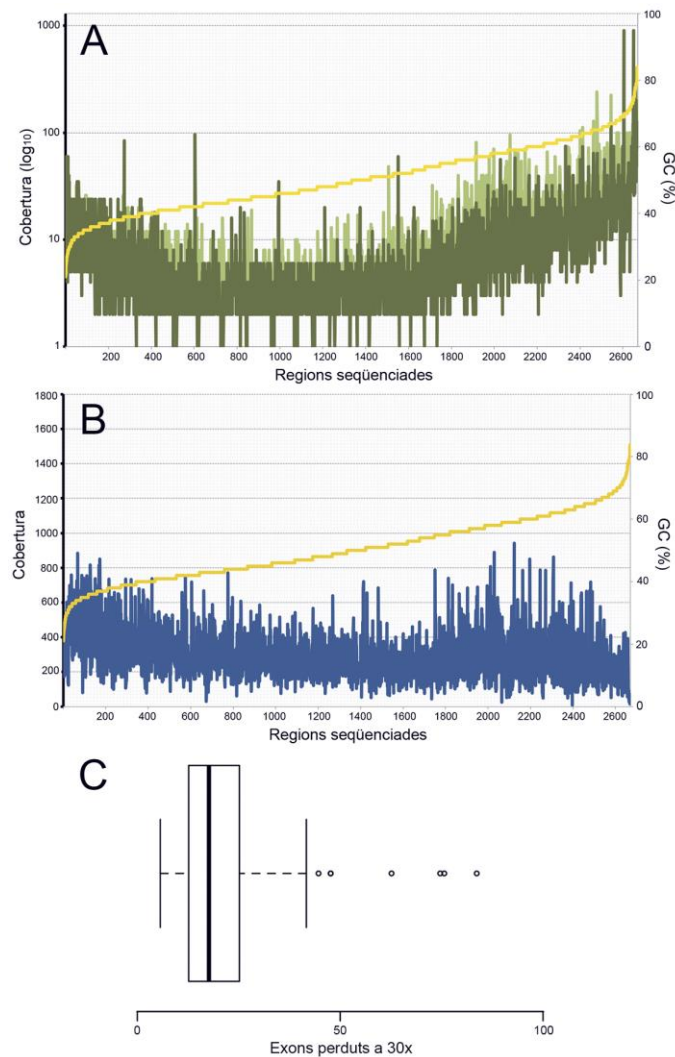


Figura 4-5 | Panell de 147 gens. **A)** Densitat de sondes per regió en funció del contingut GC (groc). Es segueix la mateixa estratègia que pel panell de 85 gens, però amb un número considerablement superior de regions per seqüenciar; **B)** Cobertures mitjanes per regió (blau) de totes les mostres seqüenciades amb el panell en funció del contingut GC (groc). S'intueix una lleugera davallada de cobertura a les regions amb GCs centrals, símptoma de que el disseny es troba al límit de la seva capacitat; **C)** Mitjana d'exons perduts a 30x per les mostres seqüenciades amb el panell.

En aquest treball, els resultats de la seqüenciació de llibreries genòmiques reflecteixen una evolució tècnica a múltiples nivells. Per una banda hi ha la millora deguda a l'experiència de l'equip tècnic que ha preparat i seqüenciat les llibreries. Aquests són uns protocols llargs i laboriosos, en els que les mostres estan sotmeses a molts punts crítics on fàcilment poden cometre's biaixos que afectaran de manera determinant el resultat final de la seqüenciació. Alguns d'aquests processos són: el pas de fragmentació física del DNA; les etapes de selecció de regions genòmiques amb les sondes lligades a les *beads* magnètiques, en els que la complexitat de la llibreria baixa de manera dràstica; o els passos

d'enriquiment dels fragments obtinguts mitjançant PCR al termociclador. Qualsevol error o imperfecció que pugui cometre's durant l'execució del protocol tindrà un efecte identificable en el resultat final de la mostra. Des del 2012 fins l'actualitat, la qualitat de les llibreries ha augmentat molt. En aquesta millora també hi ha intervingut el perfeccionament dels reactius químics de seqüenciació i la plataforma en sí mateixa. El canvi de versió 2 a versió 3 en els cartutxos de seqüenciació no només va comportar la disminució de la llargada de les seqüències obtingudes (de 151 a 76 parells de bases), sinó que va suposar un increment notable en la qualitat i quantitat d'aquestes. El MiSeq va guanyar capacitat a principis de 2015, amb la instal·lació de plaques de miralls accessoris. Amb aquestes millores, i arribant fins a una densitat de clústering més elevada que la proposada pels protocols comercials (fins a 1700 K/mm²), el MiSeq és capaç d'obtenir 33,5 milions de seqüències per *run*, dels quals 30 són d'alta qualitat (Q>30). Aquesta quantitat de seqüències assolibles en cada assaig permeten aprofitar al màxim el número de mostres per *run* (de 5 mostres inicials amb el panell de 85 gens fins a 14 mostres pel panell de 85, o 8 pel de 147 gens). Això, a banda de suposar un estalvi considerable en reactius, també provoca un descens notable en les cobertures de les mostres (sobretot en funció de la mida del panell amb el que es seqüencia la mostra). Per tant, els paràmetres de cobertura de les mostres seqüenciades fa 3 anys són poc comparables amb els que s'obtenen avui en dia.

Amb l'experiència obtinguda, s'han pogut millorar els dissenys de sondes dels diferents panells. L'objectiu de l'optimització és, en part, el de contrarestar els múltiples biaixos que poden cometre's durant la seqüenciació del DNA, com el causat per l'efecte de continguts GC extrems en certes regions genòmiques, o la inclusió de regions amb homologia de seqüència als dissenys. La millora en el disseny de les sondes es nota a l'hora d'avaluar de manera concisa el resultat de la seqüenciació. Aquests paràmetres són: l'enriquiment de les mostres, la cobertura amb la que s'han seqüenciat les regions d'interès, la uniformitat de la cobertura al llarg d'aquestes regions, la reproductibilitat dels resultats i, de manera secundària, la quantitat requerida de DNA inicial i el cost per base.

El percentatge d'enriquiment en assajos de seqüenciació per captura no és mai perfecte. Aquest es relaciona directament amb la qualitat de la mostra que es seqüencia (les mostres amb DNA degradat, encara que sigui de manera parcial, no s'enriqueixen de manera homogènia) i amb la qualitat del disseny de les sondes i l'eficiència de captura d'aquestes. El contingut GC extrem de les regions genòmiques que volen capturar-se és un dels principals causants de que l'enriquiment no sigui tant eficient com cabria esperar, sobretot amb la tecnologia d'Illumina (455). Es requereix molta menys energia per separar les cadenes de DNA de les regions amb percentatges baixos de GC (inferiors al 35%) durant el *melting*. Aquestes cadenes, un cop separades, tendeixen a degradar-se ràpidament, fet que impedeix la hibridació de les sondes. Al contrari, aquelles regions amb un percentatge de GC elevat (superior al 60-65%) poden no acabar-se de separar de manera satisfactòria, impeding també la hibridació de les sondes. Per aquest motiu, el comportament de la distribució de la cobertura en regions amb GC extrem no és comparable entre mostres del mateix *run*, però tampoc ho és entre replicats tècnics o biològics d'una mateixa mostra (456).

Un altre dels factors principals que provoquen la caiguda de l'enriquiment és l'homologia de seqüència. Les sondes degenerades contribueixen a la seqüenciació de regions que no interessen, provocant així una fuga en el potencial químic del seqüenciador. A la Figura 4-6 pot apreciar-se fins a quin punt és determinant l'absència de regions amb homologia de seqüència al disseny, per tal d'assegurar uns percentatges acceptables d'enriquiment. En aquest cas, l'exclusió d'una regió de 3 Kb (la regió 3'UTR del gen *LDLR*, amb un 10% d'homologia de seqüència –300 bases–) del panell de dislipèmia provoca que el percentatge total d'enriquiment de les mostres passi de ser del 20% al 40%.

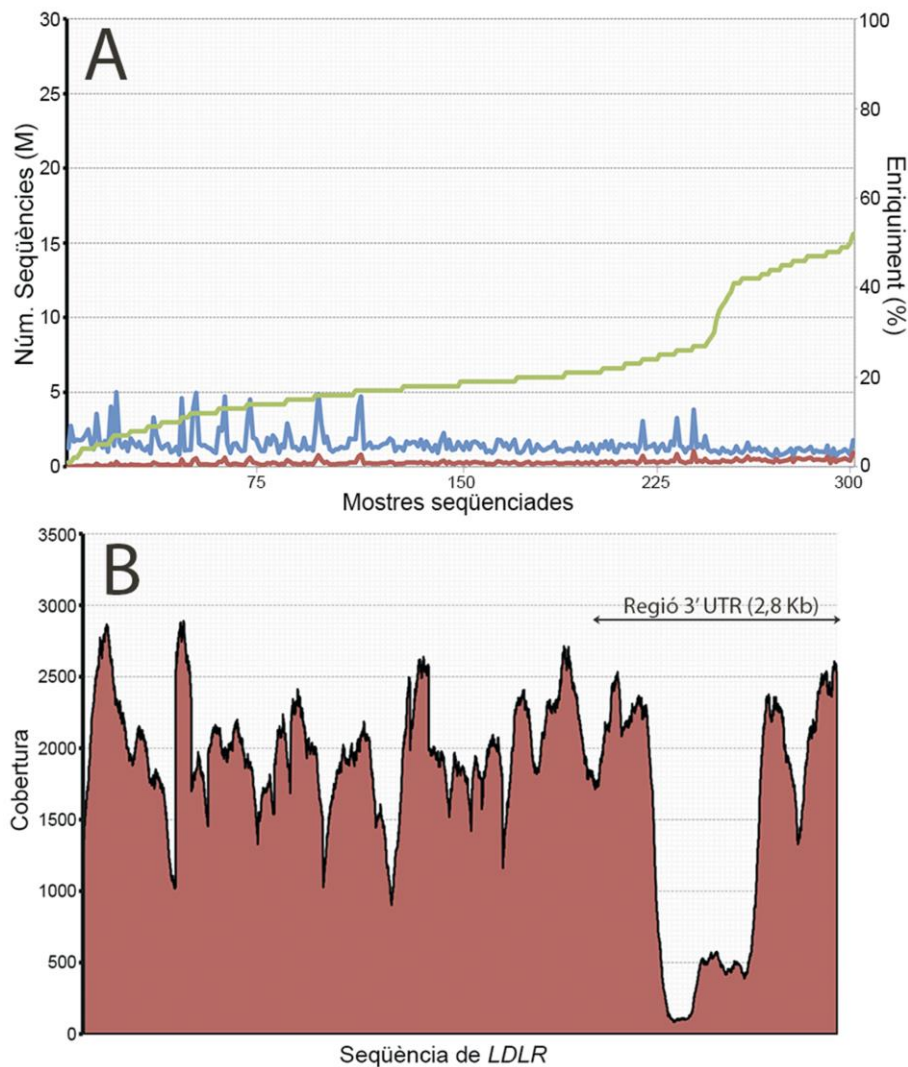


Figura 4-6 | Mostres seqüenciades amb el panell de dislipèmia. **A)** Número (en milions) de seqüències totals (blau) i solapants amb les regions d'interès (vermell) per mostra. L'enriquiment es mostra en verd; **B)** Cobertura acumulada en les regions codificants del gen *LDLR*. La regió 3' UTR apareix indicada amb una fletxa.

Les sondes dissenyades per aquesta regió hibridaven en un total de 20 regions genòmiques alternatives. La regió es va descartar en una versió posterior del mateix disseny. Pot ser que sondes teòricament perfectes hagin d'hibridar en una regió d'interès en la que el pacient té una variant

genètica. L'eficiència d'hibridació de la sonda en aquesta regió minva, provocant que no es capturi de la manera desitjada. Pel mateix motiu, algunes variants genètiques de l'individu poden provocar que la seqüència de determinades regions genòmiques es torni homòloga a la d'alguna de les sondes del disseny, fent que les sondes hibridin i no capturin la regió original d'interès (o que la capturin pitjor).

Amb independència del panell utilitzat, els percentatges d'enriquiment als que s'acostuma a arribar són al voltant del 40-50% i amb límit al 60% en algunes mostres extraordinàriament ben seqüenciades. Tenint com a referència que l'enriquiment esperat per un exoma complet és del 65-70% (457), i que el percentatge disminueix a mesura que es redueix la quantitat total de bases que es volen capturar, aquests són uns percentatges molt acceptables.

Normalment l'enriquiment es complementa amb les cobertures assolides per regió. En casos en que l'enriquiment sigui modest però la cobertura sigui elevada, la mostra probablement haurà estat seqüenciada de manera satisfactòria. En el cas de les mostres incloses en aquest treball, les cobertures són suficientment elevades com per obtenir un número d'exons perduts a 30x assumible per la dinàmica diària d'un laboratori de diagnosi genètica. Aquest punt es complementa amb la uniformitat relativa de les cobertures obtingudes amb les últimes millores en els dissenys de sondes (Figures 4-4/B i 4-5/B).

En quant a la reproductibilitat dels resultats de seqüenciació, aquesta s'avalua segons la capacitat d'identificar variants conegudes (i prèviament detectades) en una mostra estandarditzada – per exemple, una Coriell–. Amb cada un dels nous dissenys es va fer un run de prova (mostres no incloses en la cohort d'estudi) per tal d'avaluar la capacitat dels nous dissenys d'identificar variants prèviament detectades amb els dissenys anteriors. En tots els casos estudiats, 34 mostres amb variants prèviament conegudes –incloent 5 mostres Coriell–, la reproductibilitat va ser del 100%.

4.1.2 – Regions *multimap*

Com ja s'ha comentat, tant per la correcta interpretació de les variants com per la detecció de possibles artefactes és important conèixer quines regions dels panells que volen seqüenciar-se presenten homologia de seqüència (regions amb *multimapping*), ja sigui amb regions del mateix panell o amb altres regions genòmiques.

Per tal d'esbrinar quines regions dels dissenys són *multimap* es va utilitzar un *script* que genera sondes de 120 nucleòtids a partir de les regions d'interès de cada panell. Per tenir informació de la seqüència immediatament anterior i posterior a les nostres regions, aquestes es van ampliar amb un *offset* de 50 bases a cada extrem. D'aquesta manera es pot simular el *tiling* de les sondes de captura i testejar aquelles que hibridin parcialment a les regions d'interès. Un cop generades les sondes sintètiques, l'*script* les mapeja amb el *software* d'alineament GEM3 (436), amb condicions molt restrictives de mapeig, sense permetre cap *missmatch* a la seqüència. Es tria aquest alineador per la gestió de les seqüències *multimap* que duu a terme: les alinea a tots els llocs on podrien hibridar, sense

importar quants *hits* trobi al llarg del genoma. Aquest comportament és diferent al de BWA-MEM (42) (utilitzat per mapejar les mostres de la cohort), ja que de totes les possibles localitzacions que presenta la sonda, aquest en tria una a l'atzar. Seguidament, es filtra l'arxiu SAM generat a partir de l'etiqueta de bits associada a l'alineament, seleccionant aquelles que informin de que la seqüència té homologia amb més d'una localització genòmica. A partir d'aquests alineaments es genera un arxiu en el que es relaciona la coordinada d'origen de cada sonda amb les coordenades de la localització on ha mapejat. Finalment, aquest arxiu s'intersecta amb les coordenades de les regions d'interès originals per cada disseny i s'extreu el percentatge de *multimap* de cada regió.

A la Taula 4-2 es resumeixen les regions *multimap* identificades pels diferents panells. Algunes de les regions són compartides en tots els dissenys, com ara: la regió compresa des de l'exó 13 al 38 de *MYH6*, amb una homologia de seqüència significativa amb la regió de *MYH7* des de l'exó 14 fins al 39. Aquests dos gens són paràlegs –gens relacionats per processos de duplicació dins d'un mateix genoma– i no suposen gaires problemes de cares a l'anàlisi dels resultats. S'ha de tenir especial cura en la validació de les variants, assegurant que realment estiguin localitzades a l'exó del gen corresponent.

Els exons 153-157 de *TTN*, tenen una homologia del 100%. La *TTN* és la proteïna més llarga codificada pel genoma humà, i és de naturalesa modular, amb repeticions. Per altra banda, la seqüència codificant completa d'*HCN2*, *SDHA* o els exons 9-24 de *NF1* s'inclouen en els dissenys de 118 i/o 147 gens. Aquests 3 gens presenten 4, 18 i 13 pseudogens, respectivament, repartits al llarg del genoma. La seqüenciació de les regions d'aquests gens no és òptima, i les variants que puguin detectar-s'hi són pràcticament impossibles de validar, donat que no es pot saber a quina de les múltiples regions es localitza, realment. Per aquest motiu són candidates a no ser reeditades en versions futures del panell.

Taula 4-2 | Regions amb homologia de seqüència.

Panell	Exó	GC (%)	Multimap (%)	Regió multimap
Tots	NM_133378_158; <i>TTN</i>	35	13	NM_133378_155; <i>TTN</i>
Tots	NM_133378_157; <i>TTN</i>	41	100	NM_133378_154; <i>TTN</i>
Tots	NM_133378_155; <i>TTN</i>	38	100	NM_133378_(155/158); <i>TTN</i>
Tots	NM_133378_154; <i>TTN</i>	41	100	NM_133378_(154/157); <i>TTN</i>
Tots	NM_133378_153; <i>TTN</i>	47	100	NM_133378_153; <i>TTN</i>
147	NM_001743_3; <i>CALM2</i>	36	1	Altres regions
Tots	NM_198056_28; <i>SCN5A</i>	58	2	Altres regions
85, 147	NM_006514_27; <i>SCN10A</i>	50	2	NM_198056_28; <i>SCN5A</i>
85, 147	NM_006514_25; <i>SCN10A</i>	54	1	Altres regions
147	NM_003060_1; <i>SLC22A5</i>	71	11	Altres regions
147	NM_004168_2; <i>SDHA</i>	43	56	Altres regions
147	NM_004168_3; <i>SDHA</i>	51	60	Altres regions
147	NM_004168_4; <i>SDHA</i>	56	29	Altres regions
147	NM_004168_5; <i>SDHA</i>	50	12	Altres regions

147	NM_004168_6;SDHA	37	35	Altres regions
147	NM_004168_7;SDHA	57	5	Altres regions
147	NM_004168_8;SDHA	49	35	Altres regions
147	NM_004168_9;SDHA	58	20	Altres regions
147	NM_004168_10;SDHA	55	6	Altres regions
147	NM_004168_11;SDHA	40	3	Altres regions
147	NM_004168_12;SDHA	49	50	Altres regions
147	NM_004168_13;SDHA	63	37	Altres regions
147	NM_004168_14;SDHA	54	81	Altres regions
147	NM_004168_15;SDHA	41	8	Altres regions
118, 147	NM_000165_2;GJA1	48	2	Altres regions
85, 147	NM_001458_47;FLNC	64	39	Altres regions
85, 147	NM_001458_48;FLNC	57	33	Altres regions
147	NM_002834_10;PTPN11	38	4	Altres regions
Tots	NM_001129827_46;CACNA1C	58	44	Altres regions
Tots	NM_001129827_47;CACNA1C	57	70	Altres regions
Tots	NM_004572_6;PKP2	53	13	Altres regions
Tots	NM_004572_3;PKP2	59	3	Altres regions
Tots	NM_002471_38;MYH6	60	10	NM_000257_39;MYH7
Tots	NM_002471_36;MYH6	63	54	NM_000257_37;MYH7
Tots	NM_002471_34;MYH6	63	34	NM_000257_35;MYH7
Tots	NM_002471_33;MYH6	63	49	NM_000257_34;MYH7
Tots	NM_002471_31;MYH6	61	31	NM_000257_32;MYH7
Tots	NM_002471_30;MYH6	60	20	NM_000257_31;MYH7
Tots	NM_002471_29;MYH6	66	36	NM_000257_30;MYH7
Tots	NM_002471_26;MYH6	66	91	NM_000257_27;MYH7
Tots	NM_002471_21;MYH6	57	1	NM_000257_22;MYH7
Tots	NM_002471_19;MYH6	55	52	NM_000257_20;MYH7
Tots	NM_002471_18;MYH6	62	36	NM_000257_19;MYH7
Tots	NM_002471_14;MYH6	55	73	NM_000257_15;MYH7
Tots	NM_002471_13;MYH6	55	9	NM_000257_14;MYH7
Tots	NM_000257_39;MYH7	63	10	NM_002471_38;MYH6
Tots	NM_000257_37;MYH7	62	54	NM_002471_36;MYH6
Tots	NM_000257_35;MYH7	63	34	NM_002471_34;MYH6
Tots	NM_000257_34;MYH7	63	49	NM_002471_33;MYH6
Tots	NM_000257_32;MYH7	61	31	NM_002471_31;MYH6
Tots	NM_000257_31;MYH7	63	20	NM_002471_30;MYH6
Tots	NM_000257_30;MYH7	64	36	NM_002471_29;MYH6
Tots	NM_000257_27;MYH7	65	91	NM_002471_26;MYH6
Tots	NM_000257_22;MYH7	57	1	NM_002471_21;MYH6
Tots	NM_000257_20;MYH7	50	52	NM_002471_19;MYH6
Tots	NM_000257_19;MYH7	60	36	NM_002471_18;MYH6
Tots	NM_000257_15;MYH7	54	72	NM_002471_14;MYH6

				NM_000257_15;MYH7
Tots	NM_000257_14;MYH7	56	17	NM_002471_13;MYH6
147	NM_002755_7;MAP2K1	57	20	Altres regions
147	NM_002755_11;MAP2K1	50	11	Altres regions
118, 147	NM_021098_25;CACNA1H	69	3	NM_021096_25;CACNA1I
118, 147	NM_001040114_14;MYH11	61	3	Altres regions
147	NM_001042492_9;NF1	34	7	Altres regions
147	NM_001042492_10;NF1	35	10	Altres regions
147	NM_001042492_13;NF1	32	51	Altres regions
147	NM_001042492_15;NF1	29	39	Altres regions
147	NM_001042492_18;NF1	45	1	Altres regions
147	NM_001042492_20;NF1	36	16	Altres regions
147	NM_001042492_21;NF1	42	31	Altres regions
147	NM_001042492_22;NF1	36	8	Altres regions
147	NM_001042492_23;NF1	34	24	Altres regions
147	NM_001042492_24;NF1	35	53	NM_001042492_24;NF1
78,85,118,147	NM_024422_15;DSC2	46	1	Altres regions
147	NM_030662_6;MAP2K2	68	4	Altres regions
118, 147	NM_001194_1;HCN2	84	5	Altres regions
118, 147	NM_001194_2;HCN2	62	16	Altres regions
118, 147	NM_001194_4;HCN2	66	22	Altres regions
118, 147	NM_001194_5;HCN2	67	15	Altres regions
118, 147	NM_001194_6;HCN2	67	7	Altres regions
118, 147	NM_001194_8;HCN2	77	24	NM_001194_8;HCN2
118, 147	NM_021096_25;CACNA1I	63	3	NM_021098_25;CACNA1H

4.2 – Desenvolupament d'un algoritme informàtic per la detecció de CNVs

Les limitacions inherents de les tecnologies de seqüenciació de fragments curts han dificultat durant un cert temps la detecció de CNVs i, per tant, la detecció d'un gran nombre de variants potencialment causants de malaltia.

A l'inici d'aquesta tesi hi havien publicats diversos *softwares* de detecció de CNVs per dades de captura, tant d'exoma (274,275,281,282) com per dades provinents de panells de regions preseleccionades (280). Cap d'ells resultava una eina òptima pel cribratge de pacients en un context clínic, en el que cometre errors de tipus II (falsos negatius) pot arribar a tenir implicacions en la integritat del pacient i/o dels seus familiars. En aquest context també és desitjable cometre el menor nombre possible d'errors de tipus I (falsos positius), ja que qualsevol CNV potencial ha de ser validada per una tècnica *gold standard* abans de ser reportada al facultatiu. Les validacions han de ser assumibles pel laboratori de diagnosi genètica.

Entre els *softwares* disponibles es trobava el desenvolupat per Sathirapongsasuti et al., l'ExomeCNV (275). En aquest es posa en pràctica l'algoritme de segmentació binària circular desenvolupat per Olshen et al. (458), que subdivideix el genoma en segments que presenten característiques similars en quant a cobertura i freqüències d'heterozigositat dels polimorfismes seqüenciats. Aquest mètode incentiva la detecció de les CNVs de llargada superior, mentre que les petites són discriminades en favor d'un suposat guany d'especificitat. Al seu torn, el *pipeline* computacional desenvolupat per Krumm et al. (CoNIFER –*Copy Number Inference From Exome Reads*–), era capaç de detectar aquelles CNVs en les que com a mínim hi haguessin 3 exons implicats (274). S'obviaven, doncs, les CNVs més petites. Els autors van estimar una sensibilitat del 76% i una especificitat del 94%.

Wu et al., van realitzar un cribratge de CNVs, mitjançant un mètode propi basat en inferència bayesiana, a partir de dades d'exomes seqüenciats en el projecte *1000 Genomes* (282). Tot i que van detectar tant delecions com duplicacions, van centrar-se en la discussió de les primeres, al ser les variants per les que millors resultats havien obtingut. Els mateixos autors reconeixien que calia una millora de la tecnologia de seqüenciació, per tal d'obtenir millors exomes, de cobertures més elevades i homogènies, que facilitessin la detecció de variants estructurals desequilibrades. Per altra banda, Plagnol et al., van desenvolupar un paquet d'R (459) anomenat ExomeDepth (281). El programa ajustava les dades de cobertura a una distribució beta-binomial i utilitzava un model ocult de Markov per detectar quines de les regions tenien més probabilitats de ser les afectades per una CNV en funció de la cobertura observada. L'ExomeDepth era capaç de detectar delecions d'1 o 2 exons (281), però quan va fer-se públic, de la mateixa manera que els altres mètodes citats anteriorment, el paquet estava explícitament destinat a l'anàlisi de dades provinents d'exomes.

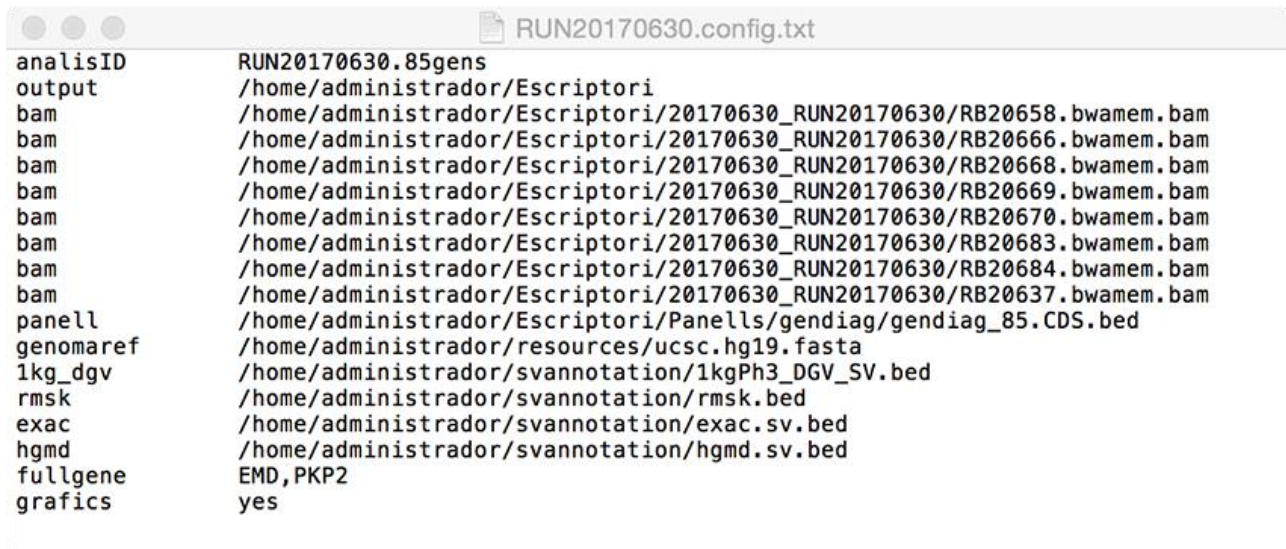
L'únic *software* dissenyat explícitament per dades de captura de regions preseleccionades era el CONTRA, desenvolupat per Li et al. (280). L'algoritme presentava una sensibilitat notable en quant a la detecció de variants de fins a 200 pb en cobertures del 50x. Tot i així, la validació del programa es va fer, en bona part, en base a la detecció de CNVs simulades. La simulació de variants és una pràctica en la que no sempre és senzill reflectir la complexitat real de les dades. No obstant, CONTRA és un dels *softwares* amb els que s'ha comparat l'algoritme desenvolupat en aquest apartat.

Així doncs, per maximitzar la utilitat de les dades de seqüenciació, era d'especial interès el desenvolupament d'un mètode de detecció de CNVs robust, sensible, específic i fàcil d'implementar en la rutina diària d'un laboratori de diagnosi genètica. L'algoritme es va dissenyar a mida per les mostres obtingudes a partir dels dissenys optimitzats de sondes.

4.2.1 – Disseny i implementació de l'algoritme

Al no disposar de dades contínues, sinó acotades a les coordenades de les regions del disseny, els punts de trencament de les potencials variants estructurals són molt difícils de detectar. Per aquest motiu, per l'anàlisi es descarten els mètodes basats en la detecció de clústers de parells anòmals de seqüències o la detecció de *soft-clipping* en seqüències solapants els punts de trencament. En canvi, l'aproximació utilitzada en el disseny de l'algoritme ha estat la detecció de canvis significatius en la cobertura de les regions seqüenciades. L'algoritme compensa els biaixos de les cobertures i les compara amb una referència dinàmica generada amb les mateixes mostres de l'anàlisi, tractant-les com pseudocontrols. D'aquesta comparació s'infereixen els números de còpia per regió.

Per la seva fàcil integració en pipelines d'anàlisi, l'algoritme capta arxius d'entrada amb formats estàndard: arxius BAM de mostres alineades i BED que continguin les coordenades de les regions de captura. És necessari també l'arxiu amb el genoma de referència que s'utilitzi, en format FASTA, i els arxius d'anotació desitjats (també en format BED). Opcionalment poden carregar-se arxius fastq; en tal cas, en la fase de preprocessament seran alineats. Els arxius resultants són informes en format VCF, XLSX i PDF i, opcionalment, gràfiques de les variants detectades en format PNG. El programa s'ha dissenyat per funcionar en un entorn Unix / Linux o Mac OS. Ha estat implementat en Perl v.5.14.2 i utilitza R v.3.2.5 i Gnuplot v.4.6 per tasques puntuals. S'executa mitjançant un arxiu de configuració (Figura 4-7) en el que apareixen els noms i les rutes dels arxius involucrats a l'anàlisi.



```

RUN20170630.config.txt
analísID      RUN20170630.85gens
output       /home/administrador/Escriptori
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20658.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20666.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20668.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20669.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20670.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20683.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20684.bwamem.bam
bam          /home/administrador/Escriptori/20170630_RUN20170630/RB20637.bwamem.bam
panell       /home/administrador/Escriptori/Panells/gendiag/gendiag_85.CDS.bed
genomaref    /home/administrador/resources/ucsc.hg19.fasta
1kg_dgv      /home/administrador/svannotation/1kgPh3_DGV_SV.bed
rmsk         /home/administrador/svannotation/rmsk.bed
exac         /home/administrador/svannotation/exac.sv.bed
hgmd         /home/administrador/svannotation/hgmd.sv.bed
fullgene     EMD,PKP2
grafics      yes

```

Figura 4-7 | Exemple d'arxiu de configuració necessari per l'execució de l'algoritme de detecció de CNVs.

És imprescindible que siguin incloses en una mateixa anàlisi aquelles mostres que s'hagin preparat seguint un mateix protocol de preparació o que hagin estat capturades amb un mateix disseny de sondes. Tot i que entre els panells hi ha solapament de regions, el número de sondes dissenyades varia entre ells. Per tant, no s'espera que les cobertures assolides a nivell de regió siguin comparables entre mostres tant diferents. El millor conjunt de mostres per l'anàlisi (en les que s'obtenen una millor correlació de cobertures i, per tant, menys soroll als resultats) és aquell format per mostres que han estat preparades conjuntament i que han estat seqüenciades en el mateix *run*.

L'algoritme divideix el flux de treball en 7 etapes principals (Figura 4-8): **(i)** l'etapa de preprocessament, en la que es generen els arxius de cobertures i les mètriques de control per cada mostra; **(ii)** la normalització de cobertures en funció de la mida de la llibreria i del contingut GC de les regions; **(iii)** el càlcul de ràtios per regió mitjançant la generació d'una referència dinàmica amb la que comparar les cobertures normalitzades de cada mostra; **(iv)** l'estimació del número de còpia associat a cada regió; **(v)** l'anotació de les senyals detectades mitjançant la informació procedent de diversos repositoris públics; **(vi)** el càlcul d'un score de fiabilitat per cadascuna de les senyals detectades **(vii)** i el resum i l'exportació dels resultats de l'anàlisi.

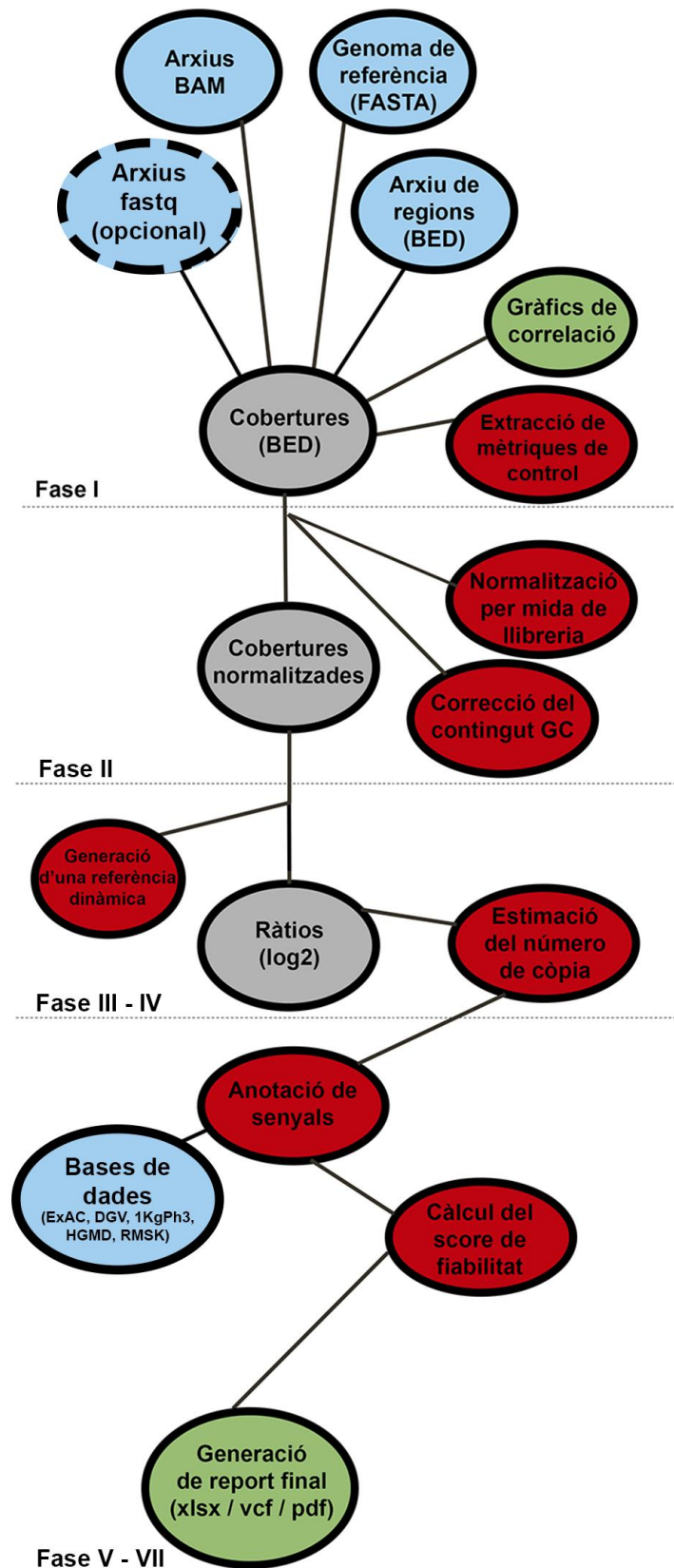


Figura 4-8 | Diagrama de flux de l'algoritme. En blau es representen els arxius de partida, en gris els arxius intermedis i en verd els arxius resultants de l'anàlisi. Els processos es representen en vermell.

La primera tasca que es duu a terme a l'etapa de preprocessament és l'extracció dels histogrames de cobertura de cada BAM mitjançant el software BEDtools v2.23.0 (460). Aquest pas acostuma a ser el coll d'ampolla del processament informàtic de mostres provinents de seqüenciació d'alt rendiment, motiu pel que és paral·lelitzat per reduir el temps d'anàlisi. A partir dels histogrames es calcula la mediana de cobertura per regió de cada mostra. L'algoritme optimitza així el temps d'anàlisi a l'etapa de normalització de cobertures, fent un primer escaneig de les regions en funció d'aquesta mesura central.

Seguidament es procedeix a correlacionar les cobertures dels cromosomes no sexuals de cada mostra contra la mediana de les mateixes calculades a partir de la resta de mostres. Si les regions s'han enriquit de manera homogènia, s'esperen obtenir uns coeficients de correlació (r) superiors a 0.97. Pel contrari, si la mostra estava parcialment degradada o si s'ha comès algun biaix important durant el procés de preparació de llibreries o de seqüenciació, els coeficients resultaran inferiors i les mostres poc comparables (Figura 4-9). Aquest paràmetre dona una previsió de com de sorollós serà l'anàlisi. L'algoritme no descarta de l'anàlisi cap mostra pel fet de presentar un baix coeficient de correlació. Això és degut a que, en l'hipotètic cas de que una mostra contingui una variant estructural que abasti un número d'exons important, aquesta mostra presentarà un valor r inferior a l'esperat. Les mostres amb baixos coeficients de correlació i que, de manera evident no sigui a causa de la presència de variants estructurals, van descartar-se com a candidates per l'anàlisi de CNVs (i, per tant, no es troben incloses a la cohort d'estudi).

Finalment, s'extreuen diverses mètriques de control per mostra que serviran tant per la normalització de les cobertures a la següent etapa, com per corroborar quines mostres no són aptes per l'anàlisi. Aquestes són: el número total de seqüències per mostra i el percentatge d'enriquiment en les regions d'interès; la mediana, el primer i el tercer quartil de cobertura, la mida mitjana i la desviació estàndard dels inserts i el contingut GC de cada una de les regions analitzades (les de mida inferior a 100 bases van ampliar-se amb 50 bases a cada extrem). Per fer d'aquest percentatge una mesura fiable, en cas de que alguna regió fos de mida inferior a 100 parells de base, aquesta es va ampliar amb 50 bases a cada extrem. Amb els percentatges de GC es generen finestres de cobertura a cada 5%. Aquestes són poblades amb les dades de cobertura de les regions que presentin un GC comprès entre els marges de la finestra. La primera finestra és la de 35% de GC; aquesta inclou les cobertures de les regions de les finestres de 20%, 25% i 30%, ja que al haver-hi un número tan baix de regions amb aquest GC, les finestres queden poc poblades i la cobertura que s'acumula és molt variable, generant artefactes durant la normalització. Això mateix també passa amb la finestra de 65%, que inclou les regions amb un GC del 70%, 75% i 80%.

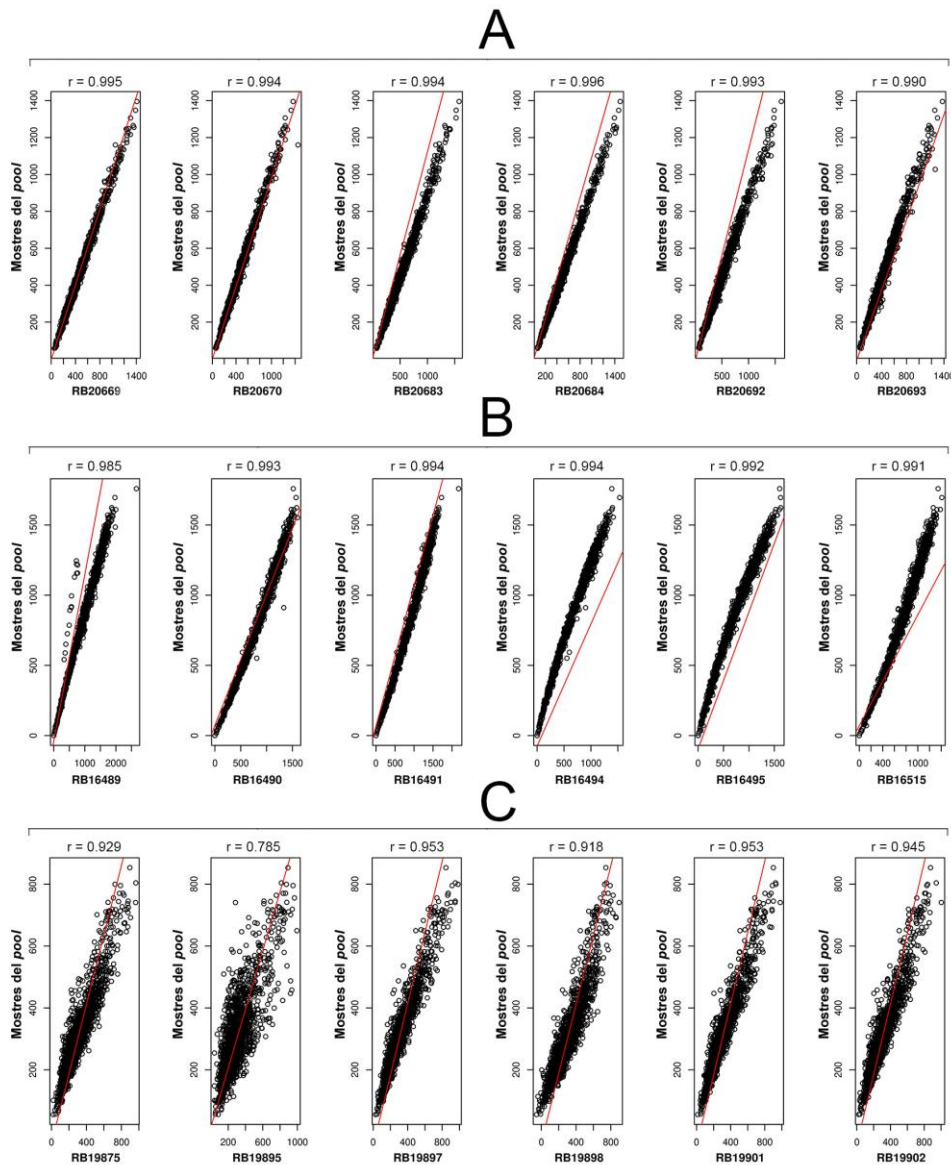


Figura 4-9 | Correlacions de cobertures entre mostres analitzades. **A)** Run de 6 mostres amb bons coeficients de correlació; **B)** Run de 6 mostres; la primera mostra del grup té una duplicació de 18 exons que fa disminuir el seu coeficient de correlació; **C)** Run de 6 mostres amb diferències d'enriament importants. No s'aconsella la seva anàlisi.

II – Normalització de cobertures

A l'etapa de normalització de cobertures es pretén corregir els biaixos causats durant el llarg procés de preparació i seqüenciació. Primerament, a partir de les cobertures extremes del cromosoma X es deriva el sexe de cada pacient. Quan aquest és masculí, les cobertures del cromosoma X són multiplicades per 2. Aleshores, la quantitat de DNA seqüenciat per regió s'equilibra en funció de la mediana general de cobertura de cada una de les mostres (Eq. 1):

$$N_{rm} = \frac{\mu_{rm}}{\mu_m} \quad (Eq. 1)$$

, on N_{rm} és la cobertura equilibrada de la regió r per la mostra m ; μ_{rm} és la mediana de cobertura a la regió r per la mostra m i μ_m és la mediana global de cobertura de la mostra m .

Tot seguit, les cobertures normalitzades per regió són corregides pel contingut GC. Com ja s'ha comentat, les regions de DNA amb continguts GC extrems són menys accessibles durant la hibridació de sondes i menys susceptibles a ser amplificades durant la preparació de llibreries (455). Aquesta afinitat diferencial d'hibridació es corregeix mitjançant la mediana de la cobertura acumulada a la respectiva finestra de GC corresponent a cada regió (calculada prèviament a la fase de preprocessament –Eq. 2–):

$$C_{rm} = \frac{N_{rm}}{\mu_{GCm}} \quad (Eq. 2)$$

, on C_{rm} és la cobertura corregida pel contingut GC a la regió r per la mostra m ; i μ_{GCm} és la mediana de la cobertura acumulada de la finestra de GC a la que pertany la regió r de la mostra m .

Teòricament, la cobertura assolida en funció del contingut GC teòricament ha de seguir una tendència unimodal amb un pic màxim en el rang del 45 al 55% de GC (Figura 4-10/B). Aquesta tendència ha estat prèviament reportada per diversos grups que han desenvolupat altres algoritmes de detecció de variants estructurals desequilibrades, ja sigui amb dades provinents de panells o d'exoma (455,461). A les dades generades amb els nostres dissenys de captura, aquesta tendència no pot apreciar-se. En canvi, i agafant com a referència les cobertures del panell de 85 gens –a mode d'exemple, per ser les cobertures més homogènies de tots els dissenys– es veu un pic màxim entorn al 35% de GC. Aquest és seguit d'una davallada i d'un segon pic relatiu entorn el 55-60% (Figura 4-10/A). Per tant, es pot comprovar com l'optimització del disseny de sondes ha fet variar la tendència esperada, assegurant un major guany de cobertures en aquests dos extrems de la distribució.

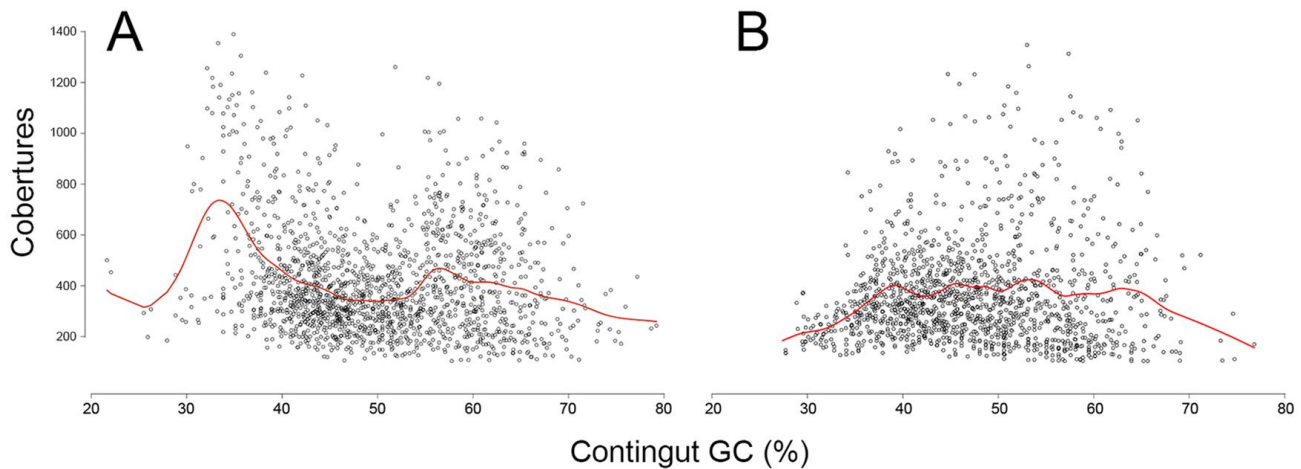


Figura 4-10 | Gràfics de cobertures en funció del contingut GC. La línia vermella reflexa la tendència de la cobertura. **A)** Cobertures medianes per regió d'un *run* realitzat amb el panell de 85 gens. **B)** Cobertures medianes per regió procedents de mostres d'un projecte no relacionat. El disseny de sondes per aquest panell no va ser optimitzat

III – Càlcul de ràtios

Les ràtios es generen a partir de les cobertures normalitzades de l'etapa anterior. Per poder-les calcular, l'algoritme genera una referència dinàmica a partir de les cobertures normalitzades per regió del conjunt de mostres analitzades (Eq. 3):

$$Rra = \frac{Cra}{\mu_r[\lambda - Cra]} \quad (Eq. 3)$$

, on Rra és la ràtio calculada a la regió r per la mostra a ; λ és el conjunt format per les cobertures normalitzades de totes les mostres $[Cra \dots Crm]$ i $\mu_r[\lambda - Cra]$ és la mediana de les cobertures normalitzades pel conjunt λ a la regió r (exceptuant la mostra a , per la qual es calcula la ràtio $-Cra-$).

Aquesta referència dinàmica (perquè varia en funció de la mostra per la que es calcula la ràtio) captura la variació tècnica de la plataforma. Per aquest motiu, com més extens sigui el *pool* de mostres que s'analitza, més representativa serà aquesta referència del material seqüenciat. Si les cobertures de les llibreries són consistents i comparables (comprovació realitzada a l'etapa de preprocessament, amb les correlacions), la variància de les cobertures per regió és inversament proporcional al número de mostres utilitzades per la construcció de la referència (Figura 4-11). Per tant, quantes més mostres s'analitzin alhora, més robust serà el càlcul de ràtios.

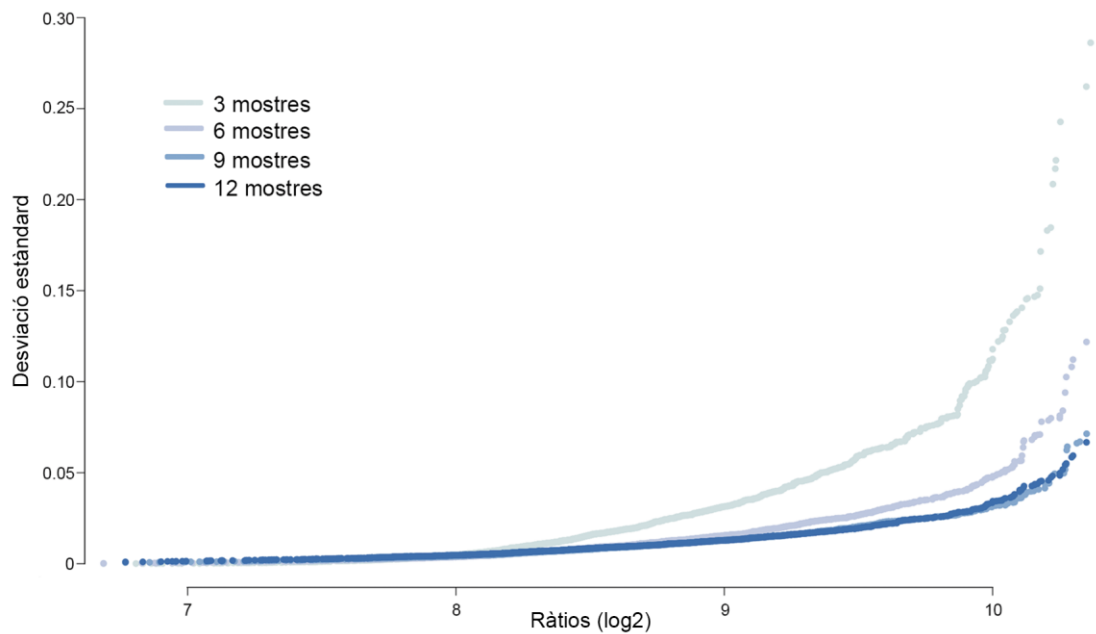


Figura 4-11 | Desviació de les ràtios per regió al llarg de la cobertura (\log_2) en funció del número de mostres incloses a l'anàlisi.

Tècnicament, el número mínim de mostres que poden analitzar-se són 3. El número mínim recomanat per anàlisi és de 6 mostres, ja que es veu una disminució significativa de la desviació en les ràtios. Tot i així, quantes més mostres s'incloguin a l'anàlisi, més variabilitat típica serà contemplada per la referència fent-la més robusta i l'anàlisi més fiable.

Un dels supòsits de partida és que la presència de CNVs a les regions codificants són events relativament poc freqüents i que, a més, són variants genètiques negativament seleccionades (90–92). Aquesta referència compleix, doncs, la condició de pseudocontrol, ja que difícilment en una anàlisi es detectarà la mateixa CNV per dos o més individus no relacionats. Per tant, en la mesura del possible s'ha d'evitar que formin part de la referència individus emparentats. D'aquesta manera s'aconsegueix que la referència no reflecteixi la variabilitat en el número de còpia de la regió afectada per una hipotètica variant estructural. Si això passés, l'efecte de la CNV en la cobertura quedaria diluït, afectant al valor esperat de la ràtio i podent arribar a emascarar la presència de la variant estructural. És per aquest motiu que en la generació de ràtios s'utilitza la mediana de les cobertures normalitzades de la regió, en comptes de la mitjana. El càlcul és més robust, ja que la mediana es veu menys afectada en el supòsit de que es detecti més d'una cobertura anormalment alta o baixa per una mateixa regió.

Les ràtios es representen en escala logarítmica (\log_2) per dos motius. El primer és per una qüestió d'escalat de les dades. Transformant logarítmicament les dades es redueixen els valors de les ràtios, provocant que els valors *outlier* destaquin sobre la resta. A més, la distribució de les ràtios resulta més Gaussiana i, per tant, més subjecta a les normes estadístiques d'aquesta distribució. El segon motiu

és per una qüestió de costum i estètica. Usualment, les anàlisis de CNVs a partir de dades d'*arrays* s'han representat sempre amb logaritmes en base 2.

A la Figura 4-12 es representa la progressiva disminució de la desviació de les ràtios a mesura que s'avança en el procés de normalització de cobertures.

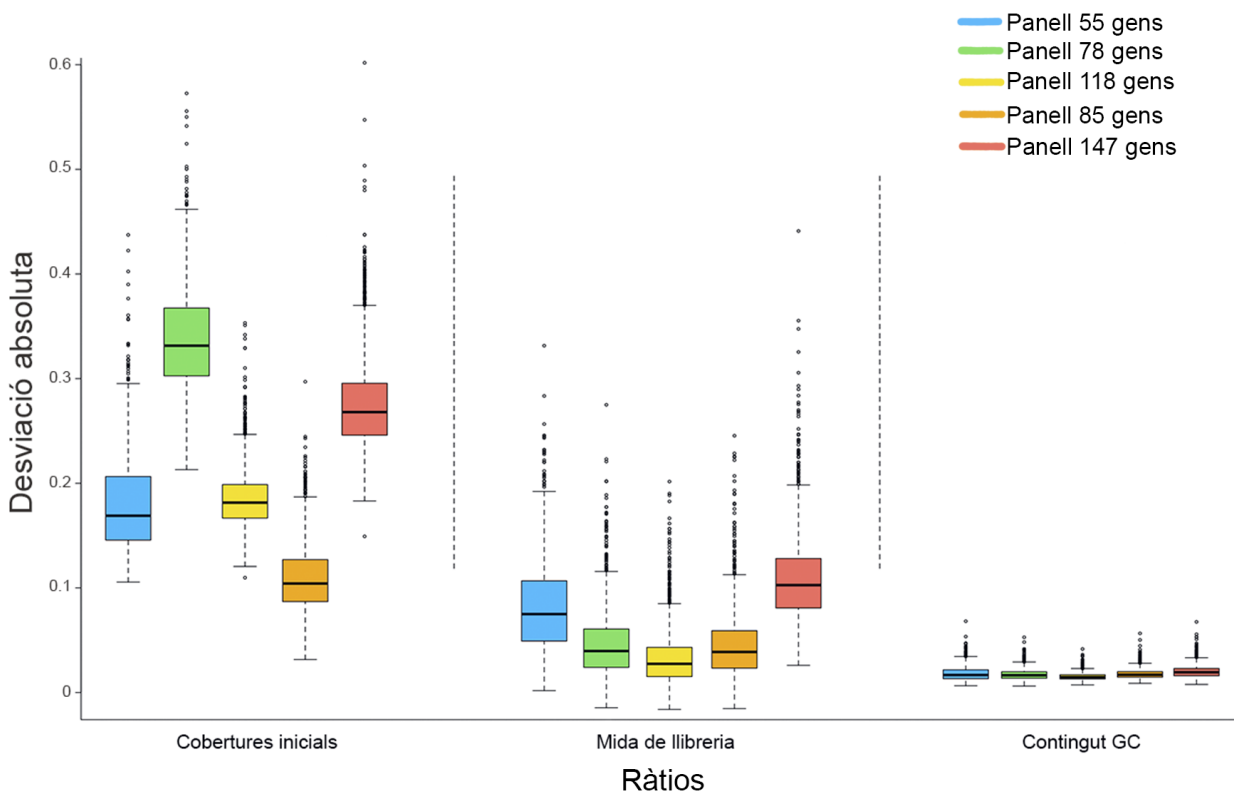


Figura 4-12 | Diagrames de caixa en els que es representa la desviació absoluta de les ràtios (log2) per regió a les diferents etapes de normalització de cobertures. Cada color correspon a un panell de gens diferent.

IV – Estimació del número de còpia

Les ràtios són considerades com *outliers* (i, per tant, susceptibles de ser considerades com CNVs) quan superen els llindars associats a números de còpia anòmals. Aquests són ≤ -2 , -0.8 , 0.45 i 0.8 pels següents números de còpia: 0 (l'equivalent a una deleció homozigota), 1 (deleció heterozigota – exemples a la Figura 4-15–), 2 (sense alteració), 3 (duplicació heterozigota) i 4 (duplicació homozigota). A més, però, per ser considerada com a tal, la ràtio ha de superar els llindars d'una finestra de variància construïda en base a la dispersió de cada regió. Un cop es comprova que un 97% de les ràtios (ja sigui per regió o les ràtios globals) es distribueixen de manera normal (Figura 4-13), la finestra de variància es fixa en la mitjana de les ràtios per aquella regió amb una dispersió afegida de 3 desviacions estàndard

(μ ràtio ± 3 SD). Les senyals que la superin seran considerades com CNVs potencials i passaran a la fase d'anotació.

La detecció de CNVs en mosaic és una limitació de l'algoritme, ja que aquest es centra en l'anàlisi de dades provinents de mètodes de captura. La variabilitat a la que estan sotmeses aquest tipus de dades juntament amb la dificultat de detectar l'aportació de DNA amb presència d'alteració en percentatges canvians de cèl·lules afectades fa que no sigui la millor opció per la detecció d'aquest tipus de variants. Tot i així, si el percentatge de duplicació o delecio és elevat, l'algoritme és capaç de detectar-la. Aquest és el cas observat en una pacient de SBr, que presenta una senyal sostinguda de ràtios anormalment elevades (tot i que la majoria no significatives) als exons 15 – 28 de *SCN5A* (Figura 4-14/A). La validació per triplicat mitjançant MLPA (Figura 4-14/B) va donar resultats similars, confirmant que es tractava d'una duplicació en mosaic (386).

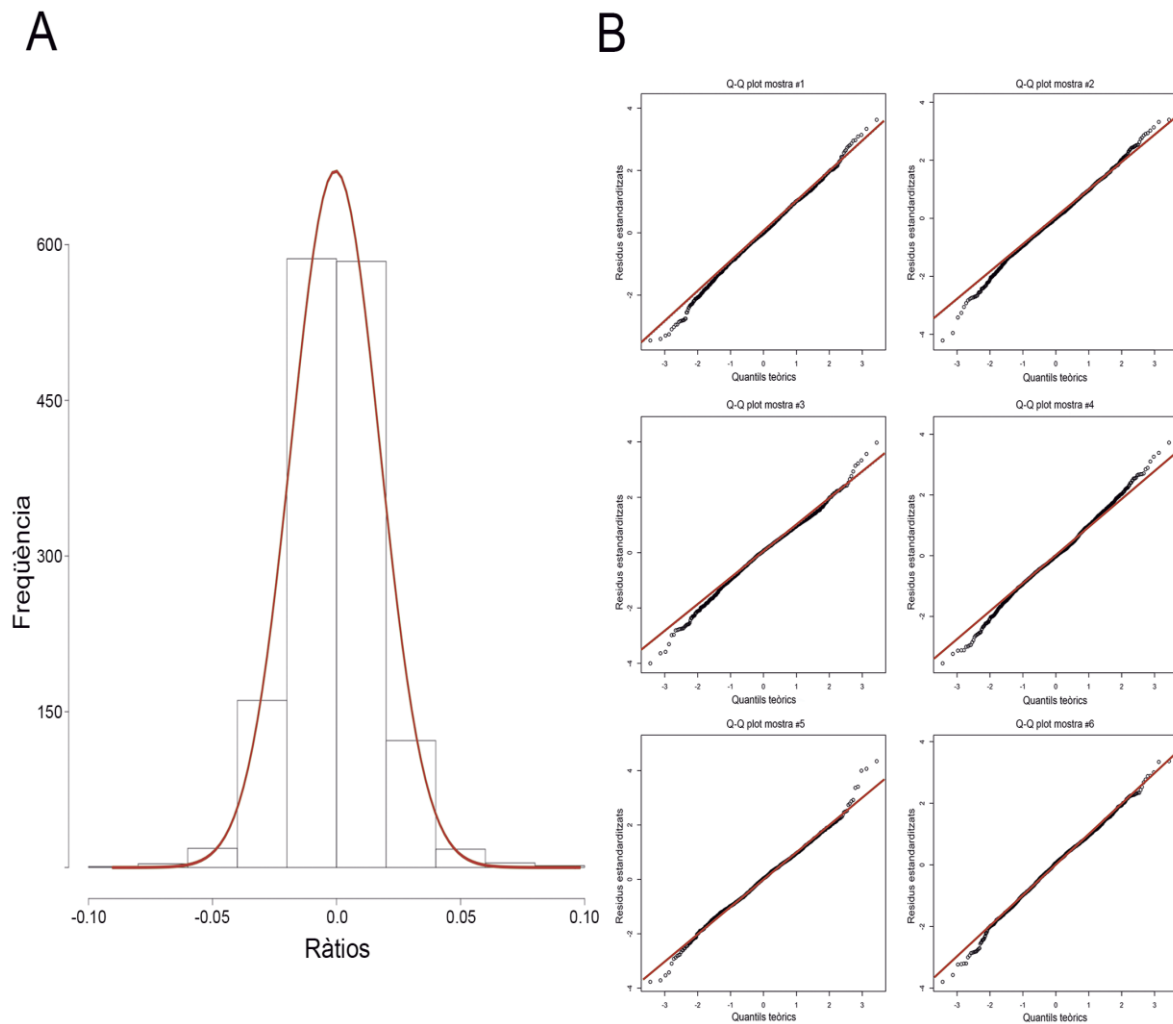


Figura 4-13 | A) Distribució normal de les ràtios generades en una anàlisi de 6 mostres. **B)** Gràfics Q-Q de normalitat per les 6 mostres incloses a l'anàlisi.

L'etapa d'anotació proveeix informació per cada potencial variant estructural identificada. S'interroguen diversos arxius procedents de bases de dades públiques, com la fase III del *1000 Genomes Project* (319), la DGV (421), ExAC (462) i l'HGMD (463). Amb aquesta informació es pot saber si les variants han estat prèviament descrites i es pot conèixer la seva freqüència poblacional estimada i, per tant, si poden considerar-se o no polimorfismes. També es pot conèixer la seva associació amb malalties, ja sigui per estudis poblacionals i/o funcionals, o si pel contrari també han estat detectades en pacients control sans. La informació dels gens associats a efectes de pèrdua de funció i de triplosensitivitat s'extreu de ClinGen (464).

A banda, les senyals són contrastades amb la informació procedent de *RepeatMasker* (465), per tal de conèixer la presència propera a la regió afectada de repeticions simples (d'1 a 5 bases), repeticions en tàndem, duplicacions segmentàries o repeticions intercalades, com els pseudogens processats, els SINEs i/o LINEs, transposons de DNA i retrotransposons provinents de seqüències de retrovirus.

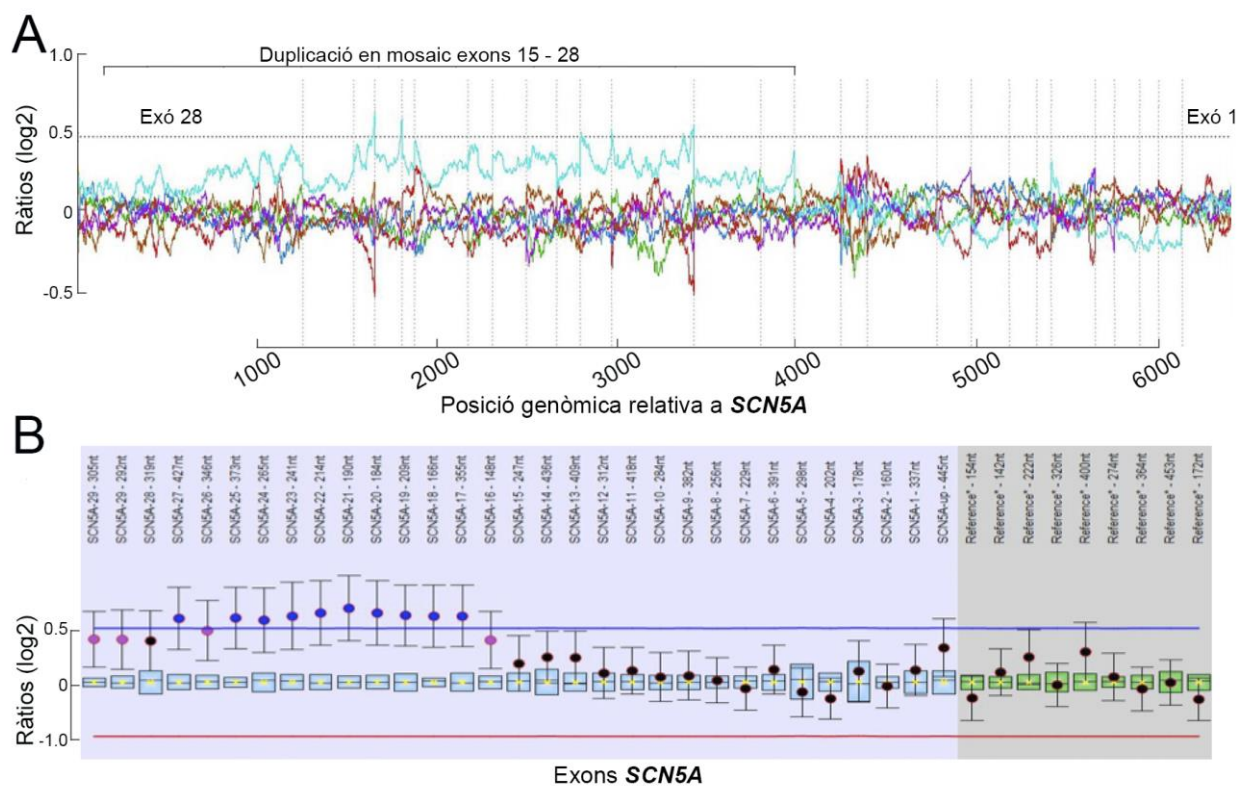


Figura 4-14 | Mosaic detectat al gen *SCN5A*, abastant els exons 15 - 28. **A)** Gràfic generat per l'algoritme en el que s'aprecia com les ràtios no arriben al límit de detecció de duplicació heterozigota; **B)** Resultats de MLPA per la validació del mosaic.

VI –Avaluació de les senyals mitjançant un score de fiabilitat

L'últim pas de l'anàlisi consisteix en la generació d'un *score* de qualitat per cada potencial CNV detectada. Aquest *score* té en compte les característiques de la mostra (prèviament calculades a l'etapa de preprocessament), el comportament de la senyal (es recuperen les cobertures de la regió a nivell de nucleòtid; les noves ràtios són recalculades per poder avaluar la qualitat de la senyal amb un nivell de precisió superior) i les característiques locals de les regions afectades.

Totes les senyals comencen amb una puntuació de 100 a partir de la qual es van sumant o restant punts en funció de la naturalesa de la mostra i de la regió. Els paràmetres que es tenen en compte són: l'enriquiment de la mostra (descompta si és marcadament inferior a les companyes d'anàlisi); un *call rate* a 30x inferior al 99%; el coeficient de correlació de la mostra; el contingut GC de la regió (que descompta si és considerat extrem –inferior a 35% o superior al 70%–; la cobertura mediana de la regió, i també si alguna de les bases es troba per sota del llindar mínim de cobertura permès de 50x. La resta de factors considerats per l'*score* no tenen tan a veure amb la mostra en sí, sinó en el comportament de les ràtios calculades a nivell de base. El que es té en compte és: la variància de les ràtios de les altres mostres en la regió analitzada, per conèixer si de manera general la regió ja és variable; si les ràtios de la mostra problema es mantenen per sobre del llindar de detecció de canvi en el número de còpia; com de sostingudes són aquestes ràtios, descomptant punts a l'*score* sempre i quan es detectin pics locals inesperats –ja que podrien ser símptoma d'un enriquiment poc homogeni de la regió–; la proximitat amb els exons adjacents, ja que podria ser que una regió molt propera a una altra estigués recollint cobertura residual d'aquesta i mostrar una duplicació que, en realitat, no existeix (461). Es puntua positivament que la senyal s'estengui als exons adjacents, evitant la sospita d'artefacte. També s'executa un *variant call* a nivell de regió per detectar SNVs amb freqüències de l'al·lel alternatiu anòmales que recolzin la presència d'una variant estructural. Per exemple, si un SNV heterozigot és present en un al·lel duplicat, la freqüència esperada serà pròxima al 66%, mentre que si és a l'al·lel delecionat, aquesta es trobarà al voltant del 33%. En el cas de deleció s'espera sempre una freqüència al·lèlica del 100%.

VII – Exportació dels resultats

Els resultats de l'anàlisi són resumits i exportats en diversos formats, en funció de la seva finalitat. A nivell visual es generen gràfiques de les CNVs detectades, tant a nivell local (exó o exons afectats, en particular) com del gen sencer on es localitza la variant. Això pot resultar útil per contextualitzar la CNV i ajudar en la seva interpretació i classificació. Aquestes gràfiques, juntament amb els resultats anotats i tabulats per cada mostra, són resumits en un arxiu PDF on també s'hi inclouen les mètriques de control de les mostres analitzades i les gràfiques de correlació.

Els resultats també s'exporten en format XLSX i VCF pel seu emmagatzematge en una base de dades pròpia i per possibles usos futurs de les dades.

4.2.2 – Validació i avaluació comparativa de l'algoritme

Per tal d'avaluar la sensibilitat i l'especificitat de l'algoritme es van analitzar 108 mostres provinents d'un banc de DNA de pacients d'hipercolesterolèmia familiar. Les mostres ja havien estat caracteritzades per MLPA, així que es coneixien les variants estructurals desequilibrades que presentaven al gen codificant pel receptor de la lipoproteïna de baixa densitat –*LDLR*– (Figura 4-15). Les llibreries de DNA genòmic van ser preparades als laboratoris de Gendiag S.L per tècnics de la mateixa empresa. Allà també van ser seqüenciades les mostres (amb un MiSeq) i analitzades amb l'algoritme.

Es van detectar un total de 16 portadors de CNVs (resultats resumits a la Taula 4-3), el que equival al 100% de variants identificables (especificitat del 100% –Eq. 4–). També es van detectar 9 falsos positius, principalment causats per 2 mostres que van resultar amb un enriquiment inferior a l'esperat. Això es tradueix en una especificitat del 91% –Eq. 5–.

$$\text{Sensibilitat} = \frac{PR}{(PR + FN)} * 100 = \frac{16}{(16 + 0)} * 100 = 100\% \quad (\text{Eq. 4})$$

$$\text{Especificitat} = \frac{NR}{(NR + FP)} * 100 = \frac{92}{(92 + 9)} * 100 = 91\% \quad (\text{Eq. 5})$$

on PR = Positius Reals; FN = Falsos Negatius; NR = Negatius Reals.

Com que el disseny de sondes inclou les regions UTR, en alguns dels pacients es va poder caracteritzar millor la CNV després de l'anàlisi. A les mostres VAL_27, VAL_40, VAL_55 i VAL_102 es va poder detectar l'afectació de les regions 3'UTR (prèviament no detectades amb MLPA). També, en el cas de VAL_62, la CNV constava inicialment com la duplicació heterozigota dels exons 7 – 10, però l'algoritme va detectar que aquesta implicava també l'exó número 11, que va revalidar-se mitjançant MLPA.

Taula 4-3 | CNVs detectades en la validació de l'algoritme amb mostres de pacients d'hipercolesterolèmia familiar.

Mostra	Gen / Isoforma	Regió afectada	Tipus de CNV*
VAL_55	LDLR / NM_000527	Exons 3 – 18 +3UTR	DEL
VAL_62		Exons 7 – 11	DUP
VAL_40		Exons 16 – 18 +3UTR	DEL
VAL_32		Promotor + 5UTR + Exons 1–2	DEL
VAL_99		Exons 9 – 12	DEL
VAL_5		Exons 3 – 6	DEL
VAL_46		Exons 3 – 6	DEL
VAL_95		Exons 3 – 6	DEL
VAL_102		Exons 17, 18 +3UTR	DEL
VAL_27		Exons 17, 18 +3UTR	DEL
VAL_98		Exons 8 – 10	DEL
VAL_67		Exons 4 – 6	DEL
VAL_38		Exons 13 i 14	DEL
VAL_71		Exons 11 i 12	DEL
VAL_61		Exó 5	DEL
VAL_81		Exó 16	DEL

*Totes les senyals detectades són en heterozigosi.

Per tal de comparar els resultats obtinguts per l'algoritme contra altres softwares de detecció de CNVs actuals, les mostres d'hipercolesterolèmia es van analitzar amb el software CNVKIT v.0.8.6 (461) i CONTRA v.2.0.8 (280). El resultat de la comparació es presenta a la Figura 4-16. Els resultats més similars són els assolits per l'algoritme desenvolupat en aquesta tesi i CNVKIT, tant a nivell d'exactitud (99.9% en els dos casos, mentre que l'obtinguda amb CONTRA és inferior (99.6%) –Eq. 6–), com a nivell de sensibilitat (100% en els dos casos, mentre que CONTRA és del 87.3%). El nostre algoritme demostra una precisió (Eq.7) una mica superior a CNVKIT i a CONTRA (85.9%, 83.3% i 75%, respectivament).

$$\text{Exactitud} = \frac{(NR + PR)}{TOTAL} * 100 \quad (\text{Eq. 6})$$

$$\text{Precisió} = \frac{PR}{(PR + FP)} * 100 \quad (\text{Eq. 7})$$

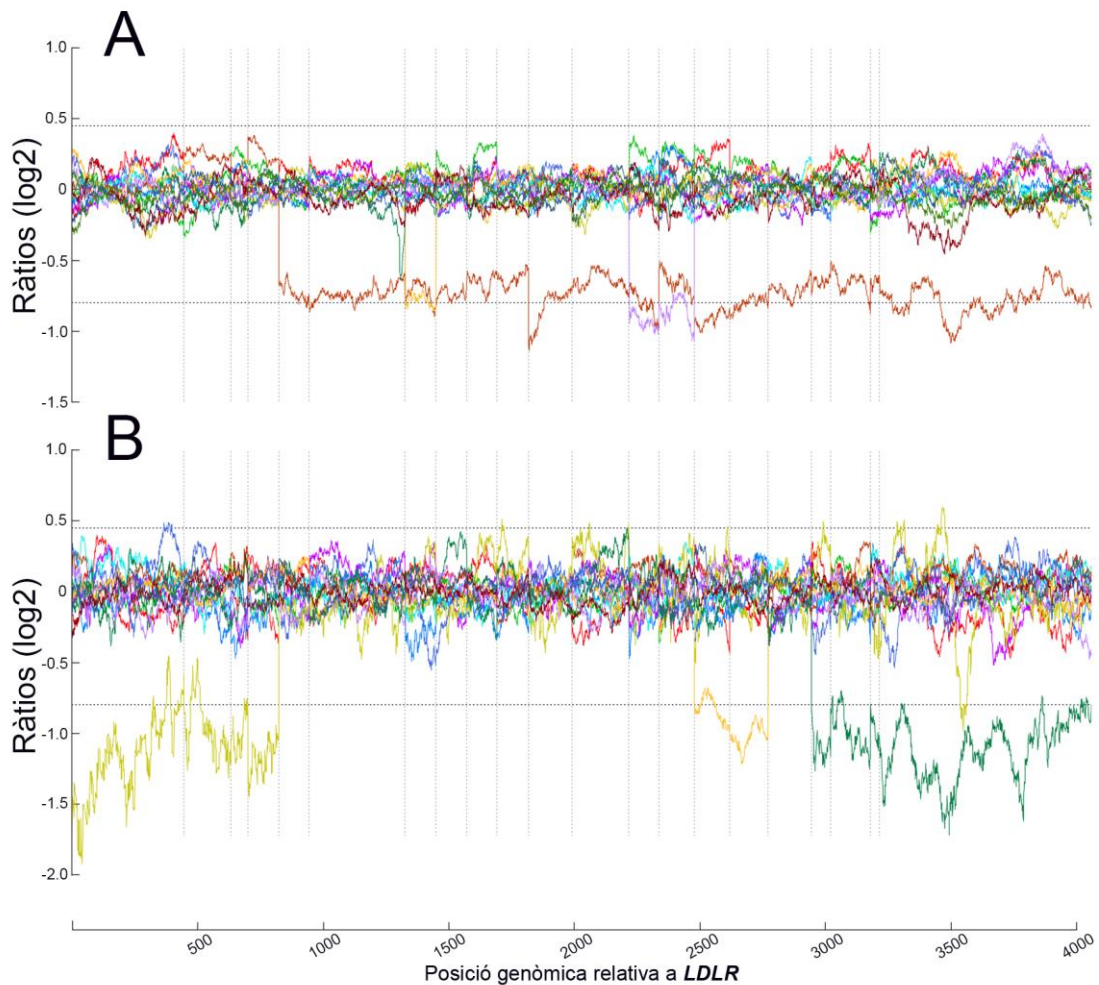


Figura 4-15 | Exemples de CNVs trobades al gen *LDLR* en pacients d'hipercolesterolèmia familiar. A) *Run* en el que s'identifica les delecions dels exons 3-18 + 3'UTR (marró, mostra VAL_55), de l'exó 5 (groc, mostra VAL_61) i dels exons 11 i 12 (lila, mostra VAL_71). B) *Run* en el que s'identifica les delecions de promotor + 5'UTR + exons 1-2 (mostassa, mostra VAL_32), la delecio dels exons 13-14 (groc, mostra VAL_38) i la dels exons 16-18 + 3'UTR (verd fosc, mostra VAL_40).

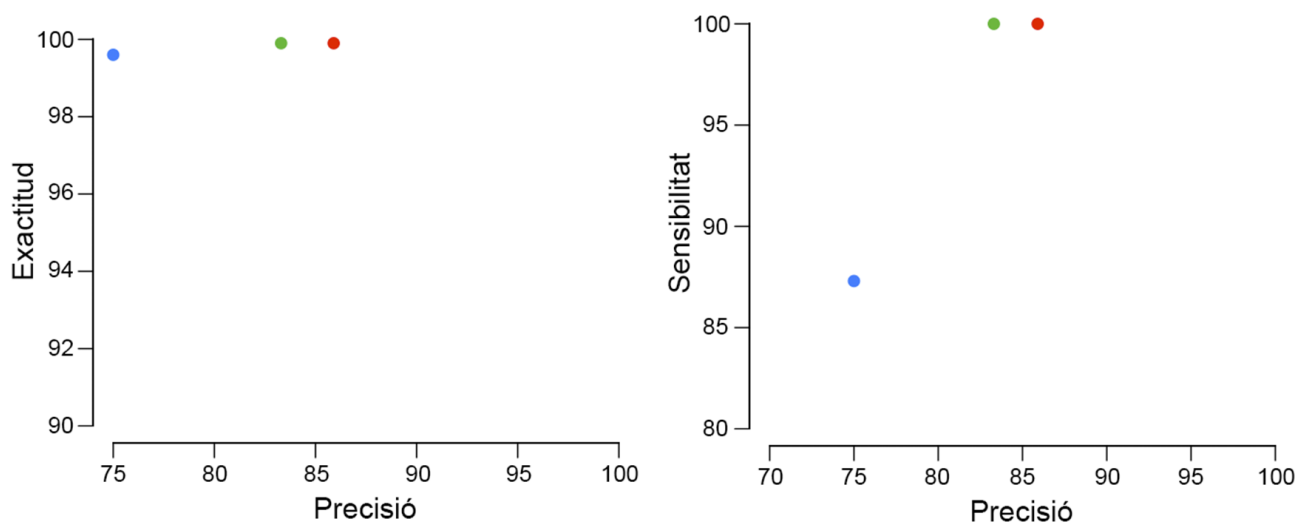


Figura 4-16 | Comparació de l'exactitud i de la sensibilitat de l'algoritme (vermell) contra el software CNVkit v.0.8.6 (verd) i el software CONTRA v.2.0.8 (blau).

4.3 – Resultats del cribratge genètic

Del total de 91 senyals detectades en els 2073 pacients estudiats, 48 van ser confirmades per MLPA o qPCR en 47 pacients diferents. Les altres 43 senyals van resultar ser falsos positius, el que equival a una taxa de descobriment de falsos positius del 47%. Per tant, la taxa de detecció global de CNVs per la cohort és del 2.3% (48/2073). La raó per la que es va detectar una freqüència relativament elevada de falsos positius va ser per evitar en la mesura del possible cometre errors de tipus II (falsos negatius). S'ha de tenir en compte que la seqüenciació d'alt rendiment de seqüències curtes té les seves limitacions, i que el cribratge de les mostres es va realitzar amb finalitats diagnòstiques. Algunes de les senyals en les que la validació va fracassar resultaven sospitoses i presentaven uns *scores* de fiabilitat baixos. Moltes d'elles eren regions amb continguts GC extrems i/o massa allunyades d'altres regions capturades com per provar estratègies de detecció de punts de trencament amb tècniques accessòries. No obstant s'ha de relativitzar el pes d'aquests 43 falsos positius. Amb l'experiència apresada, ara només són detectats quan hi han hagut problemes de preparació de llibreries, o bé en casos concrets en els que el DNA de partida era parcialment degradat. A més, la validació d'aquest tipus de senyals resulta ràpida i econòmica de dur a terme.

Es detecten 22 delecions i 26 duplicacions, totes en heterozigosi o hemizigosi. D'acord amb el nostre criteri de classificació, 16 CNVs van ser considerades VP, 10 VPP, 20 VSI i 2 VPB (Taula 4-4). De les 48 CNVs detectades, 5 havien estat prèviament reportades i 2 eren freqüents en població general. Això vol dir que 41 de les CNVs detectades no havien estat mai abans informades. Aquest fet pot explicar-se pel número reduït d'estudis dedicats al cribratge d'aquests gens per aquest tipus de variants, i per la baixa resolució de les tècniques utilitzades (com a mínim, fins fa pocs anys) pel genotipat de variants estructurals del *1000 Genomes Project*.

Únicament les freqüències d'identificació de CNVs en pacients de miocardiopaties són comparables amb les d'estudis previs, ja que són els més complets a nivell de diversitat de gens analitzats i quantitat de pacients inclosos a l'estudi. Tot i que les taxes de detecció de CNVs semblen ser més elevades en la nostra cohort, les úniques diferències significatives són entre el subgrup de pacients diagnosticats amb MCD (337,338). Això pot ser degut a les característiques de la cohort d'algun dels dos grups, o a que el mètode de detecció de CNVs del que es van servir Ceyhan-Birsoy. et al. era menys sensible que l'utilitzat en aquesta tesi.

Taula 4-4 | Resum de les CNVs identificades classificades per malaltia.

Diagnosi	CNVs	VP / VPP	VSI	VPB / VB
MCH	1.4% (9/645)	5/9	4/9	0/9
MCD	5.8% (9/154)	5/9	3/9	1/9
MCA	4.8% (6/124)	6/7	1/7	0/7
NCVE	3.1% (1/32)	0/1	1/1	0/1
MCR	-	-	-	-
SM i DATA	5.6% (2/36)	2/2	0/2	0/2
SBr	1.6% (3/193)	1/3	2/3	0/3
SQTL	4.1% (6/145)	5/6	1/6	0/6
SQTC	0.0% (0/4)	-	-	-
TVPC	0.0% (0/21)	-	-	-
FA	6.7% (1/15)	0/1	1/1	0/1
MSI	1.4% (9/651)	2/9	6/9	1/9
MSI - MSL	2.0% (1/51)	0/1	1/1	0/1
MSI - MSIU	0.0% (0/2)	-	-	-

Per cadascun dels 2073 integrants de la cohort es va realitzar un cribratge exhaustiu de SNVs i *indels*. No obstant, com l'objectiu principal d'aquesta tesi és el d'explorar la incidència de CNVs en pacients de MSI i de malalties relacionades, tan sols es reporten les SNVs i els *indels* detectats en els portadors de CNVs. La informació detallada de les variants detectades es resumeix a la Taula 4-5. Resulta interessant comprovar que totes les CNVs classificades com VP o VPP (26 de 47 portadors, un 55'3% dels casos) van ser detectades en portadors sense cap altra variant puntual considerada com una possible responsable del fenotip observat en el pacient, amb l'única excepció de P2, un pacient de MCH que presenta un genotip complex (i inèdit, a jutjar per la bibliografia) amb la deleció de l'exó 27 de *MYBPC3* –considerada com VP– i una VPP detectada en el mateix gen. Encara que la taxa de detecció d'aquestes variants sigui reduïda –cosa que no és estranya, tenint en compte que són variants negativament seleccionades a les regions codificants del genoma (90–92)–, sembla que en un percentatge important dels casos són la causa més plausible de la malaltia. En els 21 portadors restants (44,7% dels casos), les CNVs van ser classificades com VSI o VPB. En 5 casos, a banda de la CNV va detectar-se també una VP o VPP puntual que es considera la causa més probable de la malaltia del pacient. Cal tenir en compte, però, que per poder establir l'associació de les variants amb el fenotip de manera fiable es necessita tota la informació (clínica, familiar i provinent d'estudis funcionals) possible que ajudi a traçar aquesta relació. És possible que els genotips complexos en els que s'identifiquen tant CNVs classificades com VSI com variants puntuals (independentment de la seva classificació) siguin la causa de la malaltia. Tant les variants puntuals com les estructurals poden jugar el paper de moduladors que, en conjunt, poden provocar la desregulació fisiològica del pacient que el porti a desenvolupar la malaltia. Altre cop, per poder arribar a conclusions d'aquest tipus és necessari ampliar la investigació tant en la consulta del metge com al laboratori.

Per altra banda, s'han detectat 8 CNVs de manera recurrent (Taula 4-4). Aquestes són: la deleció dels exons 7 i 8 a *KCNQ1*; la duplicació dels exons 8 – 10 a *PKP2*; la deleció dels exons 21 – 23 a *DSP*; la duplicació de *KCNE1* i *KCNE2*; la deleció de la regió codificant de *PLN*; la duplicació dels exons 2 – 11 a *CASQ2* i, finalment, la duplicació de *TRDN*. Aquestes variants estructurals recurrents poden ser el resultat d'un efecte fundador, o degudes a regions genòmiques flanquejants a la reorganització genètica especialment riques en elements genètics mòbils i/o seqüències de DNA de baixa complexitat, elements genètics que tendeixen a provocar inestabilitat genòmica i a promoure l'aparició de reorganitzacions en la seqüència de DNA (158).

A continuació es detalla, per cada un dels tres grups principals de la cohort, les variants genètiques detectades en els portadors de CNVs.

4.3.1 – CNVs identificades a la cohort de miocardiopaties

I – Pacients diagnosticats amb MCH

La investigació genètica en aquest grup de pacients va revelar que 9 dels 645 eren portadors d'una CNV (1.4%). Cinc d'aquestes van ser localitzades en gens associats amb la malaltia (3 variants a *MYBPC3* i 2 a *PLN*), i classificades com VP. Les altres 4 van ser considerades VSI per exclusió segons els nostres criteris de classificació.

Dos dels 3 portadors de CNVs a *MYBPC3* presentaven un fenotip molt sever de la malaltia i eren portadors de DAI (Desfibril·lador Automàtic Implantable). A P1, home de 60 anys, se li va identificar una deleció que abastava la regió compresa entre els exons 4 i 12 (ambdós inclosos), juntament amb 3 variants puntuals a *TTN* (1 VSI i una VPB a *TTN* i 1 VSI a *NEBL*). L'estudi de cosegregació familiar, que es va realitzar a causa de la mort del pacient, va resultar en la identificació de la CNV en una de les filles, també afectada de MCH. Al segon cas, P2 és una dona de 45 anys de la que no es disposa informació clínica. Se li detecta la deleció de l'exó 27 de *MYBPC3* i 4 variants puntuals: 1 VPP prèviament reportada a *MYBPC3* (p.Val771Met); i 3VSI (a *MYBPC3*, *TTN* i *DSP*). Fins on sabem, mai abans s'havia reportat un genotip tant complex com aquest en un pacient de MCH. Un estudi previ en 113 pacients de MCH, dissenyat per identificar variants estructurals a *MYBPC3* en individus portadors d'una VP puntual al mateix gen no va detectar cap CNV (334). Per tant, el nostre cribratge demostra que, inclús en pacients amb VP en gens sarcomèrics, el cribratge per CNVs pot continuar aportant informació valuosa. Les dues CNVs van poder-se caracteritzar de manera precisa (Figura 4-17/A-B) mitjançant seqüenciació Sanger (c.406+69_1091-1154del5654 i c.2737+148_2905+40del727insG, respectivament). El tercer cas és el de P3, un home de 72 anys amb hipertròfia severa a nivell mig-ventricular, hipertròfia asimètrica de predomini septal i ECG (electrocardiograma) suggestiu de MCH. Se li va detectar la duplicació intragènica dels exons 9 – 29 de *MYBPC3* i 3 VSI puntuals (1 *TTN* i 2 *TRPM4*). En els 3 casos, les CNVs van considerar-se VP ja que es considera que poden provocar la pèrdua de funció de la proteïna,

interferint en l'ensamblatge o en el correcte funcionament del sarcòmer. A més, les variants radicals a *MYBPC3* són una causa ben coneguda de MCH (308).

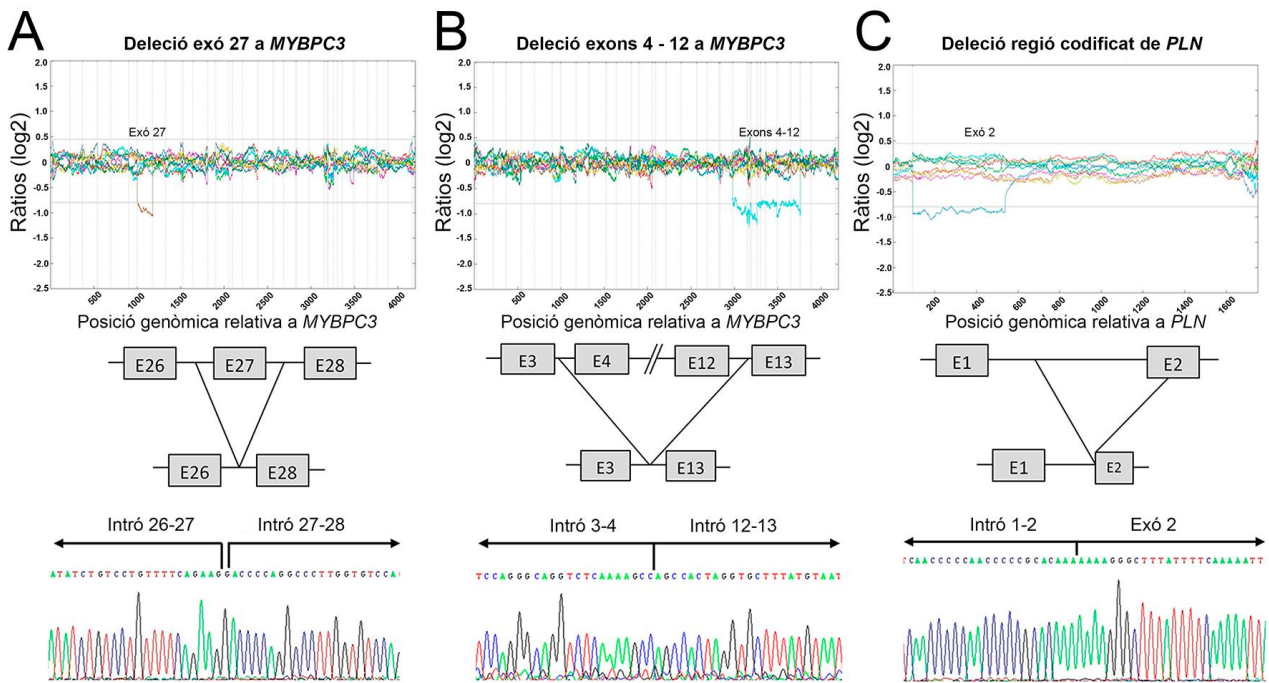


Figura 4-17 | Caracterització de 3 CNVs. Representació dels punts de trencament i la caracterització precisa per seqüenciació Sanger. **A)** La deleció de l'exó 27 de *MYBPC3* del pacient P2; **B)** La deleció de l'exó 4 al 12 a *MYBPC3* del pacient P1. **C)** La deleció de la regió codificant de *PLN*, del pacient P4. Adaptada de (466).

Taula 4-5 | Resum de les dades clíniques i els resultats genètics dels pacients portadors de CNVs. Totes les variants genètiques van ser detectades en heterozigosi o hemizigosi.

Diagnosi	Gènere	Edat	Diagnòstic	Gen	Exons	Tipus	Classificació	Criteris de classificació ^{1,4}	Altres variants genètiques
P1	H	60	MCH	MYBPC3 (NM_000256)	4 – 12	DEL	VP	PdF	TTN (NM_133378) c.77716C>T (p.Arg25906Cys), VSI TTN (NM_133378) c.6163G>A (p.Glu2051Lys), VPB NEBL (NM_006393) c.326T>C (p.Ile109Thr), VSI
P2	D	45	MCH	MYBPC3 (NM_000256)	27	DEL	VP	PdF	TTN (NM_133378) c.34552T>A (p.Ser11518Thr), VSI MYBPC3 (NM_000256) c.2311G>A (p.Val171Met), VPP MYBPC3 (NM_000256) c.1564G>A (p.Ala522Thr), VSI DSP (NM_004415) c.3862A>C (p.Lys1288Gln), VSI TTN (NM_133378) c.13451-9T>C (Variant de <i>splicing</i>), VSI
P3	H	72	MCH	MYBPC3 (NM_000256)	9 – 29	DUP	VP	PdF	TRPM4 (NM_017636) c.1294G>A (p.Ala432Thr), VSI TRPM4 (NM_017636) c.1744G>A (p.Gly582Ser), VSI
P4	D	44	MCH	PLN (NM_002667)	2	DEL	VP	PdF	-
P5	D	73	MCH	PLN (NM_002667)	2 ²	DEL	VP	PdF	-
P6	H	61	MCH	FKTN (NM_001079802)	3, 4	DUP	VSI	Exclusió	MYBPC3 (NM_000256) c.1624G>C (p.Glu542Gln), VP LAMP2 (NM_002294) c.755T>G (p.Ile252Ser), VPB ACTN2 (NM_001103) c.1984C>T (p.Arg662Trp), VSI FLNC (NM_001458) c.3757G>A (p.Val1253Ile), VSI HCN4 (NM_005477) c.2399G>A (p.Arg800His), VSI
P7	H	78	MCH	DSP (NM_004415)	Gen sencer	DUP	VSI	Exclusió	CRYAB (NM_001885) c.152C>T (p.Pro51Leu), VSI
P8	H	66	MCH	SCN4B (NM_174934)	5	DEL	VSI	Exclusió	-
P9	H	70	MCH	ABCC9 (NM_005691)	28	DEL	VSI	Exclusió	-
P10	H	61	MCD	DSP (NM_004415)	21 – 23	DEL	VP	PdF	DSC2 (NM_024422) c.928C>G (p.Gln310Glu), VSI TTN (NM_133378) c.62042T>G (p.Val20681Gly), VSI
P11	D	64	MCD	DSP (NM_004415)	21 – 23	DEL	VP	PdF	ANK2 (NM_001148) c.6176C>T (p.Thr2059Met), VSI
P12	H	44	MCD	DMD (NM_004006)	48, 49	DEL	VP	PdF	DSC2 (NM_024422) c.1436G>T (p.Arg479Leu), VSI

/resultats i discussió

P	D	60	MCD	DMD (NM_004006)	45 – 62	DUP	VP	Previament publicada (467) ⁵	MYPBC3 (NM_000256) c.3535G>A (p.Glu1179Lys), VSI
P13	D	60	MCD	DMD (NM_004006)	45 – 62	DUP	VP	Previament publicada (467) ⁵	MYPBC3 (NM_000256) c.3535G>A (p.Glu1179Lys), VSI
P14	D	27	MCD	LMNA (NM_170707)	5 – 10	DUP	VPP	DUP en gen associat	-
P15	H	8	MCD	ACTC1 (NM_005159)	2 – 7 ²	DUP	VSI	Exclusió	TTN (NM_133378) c.70426A>G (p.Ile23476Val), VSI TTN (NM_133378) c.41307T>G (p.Cys13769Trp), VSI TP63 (NM_003722) c.1372T>C (p.Ser458Pro), VSI MYH6 (NM_002471) c.1702C>T (p.Arg568Cys), VSI DMD (NM_004006) c.1688G>A (p.Arg563His), VSI TAZ (NM_000116) c.548T>A (p.Val183Glu), VPP TTN (NM_133378) c.81985C>T (p.L27329F), VSI TTN (NM_133378) c.60745C>T (p.R20249*), VPP
P16	H	66	MCD	CACNA1C (NM_000719)	46 – 48	DUP	VSI	Exclusió	LMNA (NM_170707) c.1541G>A (p.Trp514Ter), VP
P17	H	64	MCD	KCNE1 (NM_000219) KCNE2 (NM_172201)	2 – 4 ² Gen sencer	DUP	VPB	>10 DGV	CACNB2 (NM_201590) c.1540G>C (p.Val514Leu), VSI CTNNA3 (NM_013266) c.580-8C>T (variant de splicing), VSI VCL (NM_014000) c.2905G>A (p.Ala969Thr), VSI TTN (NM_133378) c.12389G>A (p.Cys4130Tyr), VSI TTN (NM_133378) c.22762A>G (p.Ile7588Val), VPB TTN (NM_133378) c.16609G>A (p.Glu5537Lys), VSI TTN (NM_133378) c.43291_43293dupGAT (p.Asp14431dup), VSI TTN (NM_133378) c.10754A>C (p.Gln3585Pro), VSI DSP (NM_004415) c.2720G>A (p.Arg907His), VSI TTN (NM_133378) c.13332G>C (p.Gln4444His), VSI
P18	H	48	MCD	MYH11 (NM_002474)	Gen sencer	DUP	VSI	Exclusió	TTN (NM_133378) c.66285G>T (p.Trp22095Cys), VSI TTN (NM_133378) c.49738A>G (p.Met16580Val), VSI TTN (NM_133378) c.83060_83062delGAG (p.Gly27687del), VSI
P19	H	42	CA	PKP2 (NM_004572)	8 – 10	DUP	VPP	DUP en gen associat	TGFBR2 (NM_003242) c.671G>A (p.Arg224His), VSI
P20	H	58	CA	PKP2 (NM_004572)	8 – 10	DUP	VPP	DUP en gen associat	
P21	H	48	CA	PKP2 (NM_004572)	8 – 10	DUP	VPP	DUP en gen associat	
P22	H	29	CA	PKP2 (NM_004572)	1	DEL	VP	PdF	
P23	H	45	CA	PKP2 (NM_004572)	1	DEL	VP	PdF	
P24	H	30	CA	DSP (NM_004415)	9 – 24	DEL	VP	PdF	

/resultats i discussió

P25	H	39	CA	MYOZZ (NM_016599)	2 – 6 ²	DEL	VSI	Exclusió	MYH7 (NM_000257) c.115G>A (p.Val39Met), VSI PKP2 (NM_004572) c.259G>C (p.Val87Leu), VSI RYR2 (NM_001035) c.649A>G (p.Ile217Val), VSI
P26	H	44	NCVE	TTN (NM_133378)	45 – 275	DUP	VSI	Exclusió	ANK2 (NM_001148) c.9842A>G (p.Gln3281Arg), VSI
P27	D	48	SBr	SCN5A (NM_198056)	15 – 28	DUP	VPP	Mosaic DUP en gen associat	-
P28	H	58	SBr	TRDN (NM_006073)	Gen sencer	DUP	VSI	Exclusió	SLMAP (NM_007159) c.2170A>T (p.Ser724Cys), VSI
P29	H	52	SBr	CASQ2 (NM_001232)	2 – 11	DUP	VSI	Exclusió	CASQ2 (NM_001232) c.749G>A (p.Arg250His), VSI ³
P30	D	39	SQTL	KCNQ1 (NM_000218)	7, 8	DEL	VP	Prèviament publicada (393,394)	DSP (NM_004415) c.529TT>C (p.Leu1764Pro), VSI TTN (NM_133378) c.32880A>C (p.Glu10960Asp), VSI MYH6 (NM_002471) c.1763A>C (p.Asp588Ala), VPB
P31	D	17	SQTL	KCNQ1 (NM_000218)	7, 8	DEL	VP	Prèviament publicada (393,394)	TTN (NM_133378) c.62185A>G (p.Arg20729Gly), VPB
P32	D	42	SQTL	KCNQ1 (NM_000218)	7, 8	DEL	VP	Prèviament publicada (393,394)	KCNQ1 (NM_000218) c.1896A>C (p.Arg632Ser), VSI
P33	D	22	SQTL	KCNH2 (NM_000238)	1 – 14	DEL	VP	Prèviament publicada (393)	KCNE1 (NM_000219) c.277G>A (p.Ala93Thr), VSI AKAP9 (NM_005751) c.4707T>G (p.Ile1569Met), VSI
P34	D	37	SQTL	KCNE1 (NM_000219)	3, 4	DEL	VPP	DEL en gen associat	CACNA1C (NM_001129827) c.3983G>T (p.Cys1328Phe), VSI
P35	D	5	SQTL	NEXN (NM_144573)	2 – 13 ²	DUP	VSI	Exclusió	KCNQ1 (NM_000218) c.1486_1487delCT (p.L496Afs*19), VP
P36	H	50	FA	KCNJ5 (NM_000890)	2 – 3 ²	DUP	VSI	Exclusió	ANK2 (NM_001148) c.1903A>G (p.Ile635Val), VSI TTN (NM_133378) c.81610G>A (p.Glu27204Lys), VSI TTN (NM_133378) c.48551C>T (p.Pro16184Leu), VSI
P37	H	42	MSI	EMD (NM_000117)	2 – 5	DUP	VPP	DUP en gen associat	TTN (NM_133378) c.46610G>A (p.Arg15537His), VSI
P38	H	22	MSI	TNNI3 (NM_000363)	4,5,8	DUP	VPP	DUP en gen associat	-
P39	H	48	MSI	TRDN (NM_006073)	Gen sencer	DUP	VSI	Exclusió	SLMAP (NM_007159) c.198+5T>C (variant de splicing), VSI
P40	H	28	MSI	CASQ2 (NM_001232)	2 – 11	DUP	VSI	Exclusió	TTN (NM_133378) c.58483G>C (p.Val19495Leu), VSI TTN (NM_133378) c.64051A>G (p.Met21351Val), VSI

P41	H	18	MSI	CTNMA3 (NM_013266)	12, 13	DEL	VSI	Exclusió	HCN2 (NM_001194) c.2404C>T (p.Pro802Ser), VSI
									GAA (NM_000152)
P42	D	<1	MSI - MSL	RANGRF (NM_016492)	Gen sencer	DUP	VSI	Exclusió	-
P43	H	48	MSI	PDLIM3 (NM_014476)	Gen sencer	DEL	VSI	Exclusió	MYPN (NM_032578) c.59A>G (p.Tyr20Cys), VSI
P44	D	38	MSI	TAZ (NM_000116) EMD (NM_000117)	Gen sencer	DUP	VSI	Exclusió	PDLIM3 (NM_014476) c.812C>T (p.Thr271Met), VSI
P45	H	49	MSI	KCNF1 (NM_000219) KCNE2 (NM_172201)	2 - 4 ² Gen sencer	DUP	VPB	>10 DGV	-
P46	H	17	SM i DATA	FBN1 (NM_000138)	45 - 65	DUP	VPP	DUP en gen associat	KCNH2 (NM_000238) c.2707G>A (p.903G>R), VSI TTN (NM_133378) c.2611G>T (p.871V>L), VSI TTN (NM_133378) c.35986T>A (p.11996S>T), VPB
P47	D	8	SM i DATA	FBN1 (NM_000138)	55	DEL	VPP	DEL en gen associat	TGFBRT (NM_004612) c.46G>A (p.Val116Met), VSI

1. "Exclusió" vol dir que la CNV no compleix els criteris per ser considerada VP, VPP, VB o VPB.
2. Aquesta deleció/duplicació pot ser del gen sencer (les regions no codificants no eren incloses al panell de gens utilitzat per la seqüenciació).
3. SNV present a l'al·lel duplicat.
4. PdF significa "Pèrdua de Funció".
5. Aquesta duplicació ha estat prèviament identificada en un pacient masculí amb Distrofia Muscular de Duchenne (467). S'ha reportat, però, que els portadors femenins d'alteracions al gen *DMD* poden exhibir exclusivament un fenotip de MCD (468).

En quant als dos pacients portadors de CNVs a *PLN*, en els 2 casos la deleció involucrava la regió codificant sencera del gen. P4, una dona de 44 anys, presenta MCH de predomini septal no obstructiva. El seu pare pateix una miocardiopatia, però és pendent d'estudi. El seu oncle patern va morir de MSL. Té un fill amb miocardiopatia congènita complexa en seguiment. Se li detecta una deleció de 7936 pb a *PLN*, que inclou una porció del primer intró i de la part no codificant del segon exó (Figura 4-17/C). La deleció va ser identificada amb el panell de 55 gens, que incloïa les regions UTR del gen. Els punts de trencament de la variant estructural van poder-se confirmar per Sanger [c.1-7587_159+190del7936]. L'altra deleció, detectada a P5, una dona de 73 anys, va ser detectada amb el panell de 78 gens i, per tant, no es disposa d'informació en regions no codificants allunyades de l'exó. Tot i així, es va poder comprovar que la deleció és diferent a la de P4, donat que, després de fer Sanger amb els encebadors utilitzats per la caracterització dels punts de trencament de P4, no es va amplificar cap fragment. Les delecions a *PLN* són probablement associades a la malaltia, ja que els portadors només presenten una còpia funcional del gen, i la reducció de l'expressió de *PLN* (deguda a VP *nonsense* o localitzades al promotor) ha estat prèviament associada amb el desenvolupament de MCH (181,469). A més, per cap dels dos es va identificar cap altra SNV o *indel* de rellevància que pogués explicar el fenotip. La informació relativa als pacients P1, P2, P4 i P5 va publicar-se recentment pel nostre grup (Annex 3) (466).

Al grup restant de portadors de CNVs, aquestes s'han detectat en gens no associats a MCH. Totes les CNVs van ser classificades com VSI. El grup el formen els pacients: P6, home de 61 anys en el que s'identifica la duplicació dels exons 3 i 4 de *FKTN* –gen associat amb diversos tipus de distròfia muscular (470)– i la possible causa del fenotip, una VP puntual a *MYBPC3* (p.Glu542Gln); P7, home de 78 anys amb la duplicació de *DSP* (gen desmosomal) i 3 VSI a *ACTN2*, *FLNC* i *HCN4*; el pacient P8, home de 66 anys en el que es detecta la deleció de l'exó 5 de *SCN4B* (subunitat del canal de sodi, associada amb SQTL) i 1 VSI a *CRYAB*; i el pacient P9, amb la deleció de l'exó 28 d'*ABCC9*, gen relacionat amb FA, hipertricosi, osteocondrodisplàsia i cardiomegàlia.

II – Pacients diagnosticats amb MCD

Els resultats de l'anàlisi genètica dels 154 pacients de MCD de la cohort van permetre identificar 9 portadors de CNVs, l'equivalent al 5.8% del grup. Es van detectar 6 variants en gens associats a la malaltia (2 variants a *DSP*, 2 variants a *DMD*, 1 a *LMNA* i 1 a *ACTC1*). Les 3 CNVs restants van ser detectades en altres gens, en principi no associats a MCD, i considerades VSI.

Entre els 2 portadors de la deleció dels exons 21 – 23 a *DSP* trobem el pacient P10, un home de 61 anys diagnosticat de MCD idiopàtica amb disfunció greu i portador de DAI. Es va investigar la presència de la CNV entre familiars no afectats i es van obtenir resultats negatius. P10 és també portador de dues VSI puntuals a *DSC2* i *TTN*. L'altre pacient, P11, una dona de 64 anys, és portadora de DAI i rep descàrregues amb freqüència. Té història de mort sobtada familiar: una germana

afectada i portadora de DAI; la tia paterna va morir sobtadament als 47 anys i la filla d'una altra tia paterna sofreix una MSC als 15 anys. A la pacient se li identifica també una VSI a *ANK2*. Per l'estudi de cosegregació familiar tan sols es disposa del DNA d'un familiar no afecte, negatiu per la deleció. En els dos casos, la deleció és considerada com una VP per causar la pèrdua de funció de la proteïna codificada. La Desmoplaquina és la proteïna del citoesquelet encarregada d'ancorar els filaments intermedis a les plaques desmosomals. És un obligat constituent dels desmosomes funcionals i mutacions en aquest gen s'han associat a diverses miocardiopaties, entre elles la MCD i la MCA.

Pels 2 portadors de CNVs a *DMD* no es disposa de cap informació clínica a banda del diagnòstic de MCD. P12 és un home de 44 anys amb la deleció dels exons 48 i 49. A més, és portador d'una VSI puntual a *DSC2*. El segon, P13, és una dona de 60 anys amb una duplicació que abasta els exons 45 – 62 de *DMD*. La CNV havia estat prèviament identificada en un pacient masculí amb Distròfia Muscular de Duchenne (467). S'ha reportat, però, que els portadors femenins d'alteracions a *DMD* (gen localitzat al cromosoma X) poden exhibir un fenotip exclusiu de MCD (468). La pacient també és portadora d'una VSI puntual a *MYBPC3*.

El cas de P14 és interessant, ja que es disposa de molta informació clínica i de suficient informació familiar com per poder establir un bon estudi de cosegregació (Figura 4-18). P14 és una dona de 27 anys que es queixa de palpitations. A l'ecocardiograma se li detecta una funció del ventricle esquerre normal i sense dilatació, però una disfunció sistòlica de tipus I inusual per la seva edat. El seu pare, de 64 anys i també diagnosticat de MCD, presenta un fenotip molt sever i és portador de DAI. L'oncle patern, de 59 anys, també afecte i portador de DAI (amb més de 18 descàrregues fins a data d'avui), té tres filles. Entre aquestes, una dona de 34 anys diagnosticada amb MCD, palpitations i un lleuger arrodoniment del ventricle esquerre. La pacient presenta un deteriorament moderat de la funció sistòlica i episodis puntuals de fibril·lació ventricular, motiu pel que es procedeix a la implantació d'un DAI. Amb el cribratge genètic se li detecta la duplicació dels exons 5 – 10 a *LMNA*. La CNV és classificada com VPP, ja que les mutacions en aquest gen han estat associades amb MCD, per considerar-se probable la disrupció de l'embolcall nuclear dels miòcits cardíacs, deguda o bé a una baixa expressió de la proteïna, o bé a una pèrdua de funció de part de la proteïna codificada (345). L'estudi familiar conclou la cosegregació de la duplicació entre els afectats. S'inicia el seguiment clínic dels fills de la pacient, al ser també portadors de la variant.

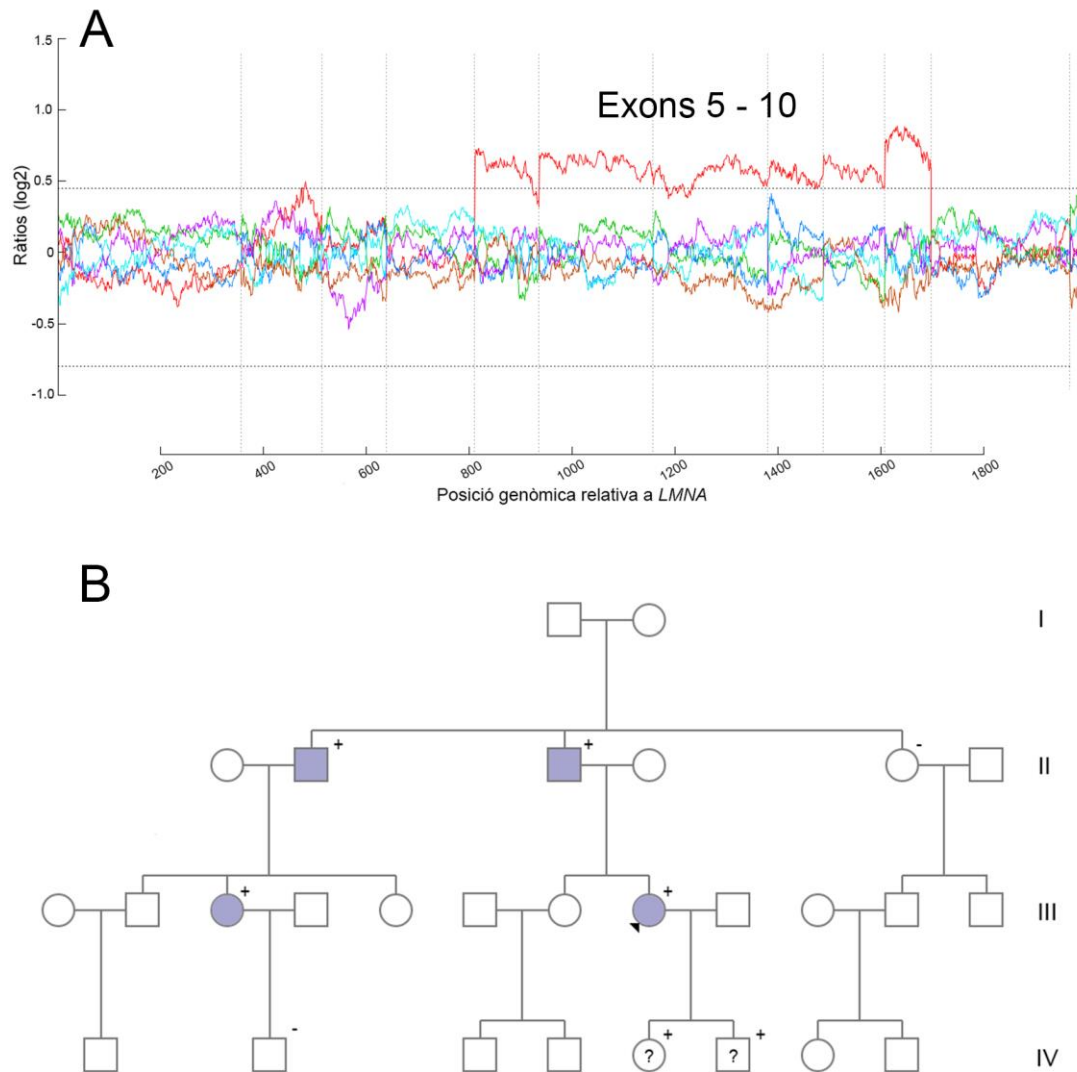


Figura 4-18 | Duplicació dels exons 5 – 10 a *LMNA* **(A)** i pedigrí **(B)** del pacient P14. El pacient s’assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol “+” indica que l’individu és portador de la variant, i el “-”, que no ho és portador.

Els pacients P15, P16 i P17 van ser identificats com portadors de duplicacions classificades com VSI. Al primer, un nen de 8 anys, se li detecta la duplicació de la regió codificant d’*ACTC1*, codificant per una proteïna de la família de les actines, les constituents majoritàries de l’aparell contràctil del múscul llis (470). La duplicació no pot caracteritzar-se amb precisió, però probablement s’estengui més enllà dels límits del gen, motiu pel que va classificar-se com VSI tot i l’associació d’*ACTC1* a MCD. Al pacient se li detecta també, a banda de 5 VSI puntuals en diversos gens, una VPP a *TAZ* (p.Val183Glu). Al seu torn, a P16, un home de 66 anys, se li detecta la duplicació dels exons 46 – 48 de *CACNA1C*. La disrupció d’alguns canals iònics es troba entre els mecanismes responsables d’un fenotip de MCD. Molts canals iònics interactuen amb el sarcolemma i les proteïnes sarcomèriques. Aquestes relacions entre proteïnes són les que acaben marcant el ritme del cor miopàtic (343). Tot i així, la duplicació és considerada VSI perquè el gen no ha estat mai abans associat a MCD. A més, al

pacient se li detecta també un codó *stop* prematur a *TTN* (p.R20249*), classificat com VPP i com la causa més plausible del fenotip. A P17, un home de 64 anys, se li detecta la duplicació de *KCNE1* i *KCNE2* (gens adjacents). La CNV és classificada com VPB per ser detectada en més de 10 individus a DGV. Per aquest pacient es reporta un codó *stop* prematur a *LMNA* (p.Trp514*) classificat VP i considerat com la causant més plausible del fenotip.

L'últim portador de CNV identificat en aquest subgrup és P18, un home de 48 anys al que se li detecta la duplicació de *MYH11*, juntament amb 4 altres VSI puntuals. La CNV no cosegrega amb la família, no es detecta en alguns dels individus afectats i és present en altres sans (al moment de l'anàlisi). Per aquest motiu, i tot i que la família continuï en seguiment, la duplicació va considerar-se VSI.

III – Pacients diagnosticats amb MCA

L'anàlisi genètica dels 124 pacients de MCA resulta en la detecció de 6 portadors de CNVs (4.8% del grup). D'aquests, 5 es troben en gens associats amb la malaltia (*PKP2* i *DSP*) i són classificats com VP o VPP.

La duplicació dels exons 8 – 10 de *PKP2* (Figura 4-19/A) va detectar-se en tres pacients. El primer, P19, és un home de 42 anys, portador de DAI i sense antecedents familiars de MS. Presenta una evolució de 2 a 3 anys de palpitations desencadenades per l'esforç físic, juntament amb taquicàrdies ventriculars que reverteixen farmacològicament. L'ecocardiograma és suggestiu de displàsia i la RMC (Ressonància Magnètica Cardíaca) mostra aprimament i irregularitats a la paret del ventricle dret, particularment al tracte de sortida. No poden obtenir-se imatges clares d'infiltració adiposa, tot i que s'intueixen. Al pacient se li detecten també dues variants puntuals a *TTN*, una VSI i una VPB. El segon pacient, P20, és un home de 58 anys sense antecedents familiars de miocardiopatia. També és portador de 2 VSI puntuals, a *TTN* i 1 a *DSP*. La RMC mostra la dilatació del ventricle dret, amb una disfunció sistòlica global. La paret del ventricle és aprimada, trabeculada i irregular. Se li distingeixen zones de discinèsia a l'àpex i al tracte d'entrada i sortida del ventricle dret. Es detecten petites àrees aneurísmiques i infiltració adiposa poc clara, però aparent. Tant per P19 com per P20 no es pot dur a terme l'estudi de cosegregació familiar per falta de dades i DNA dels familiars. En canvi, per P21 (portador de la mateixa duplicació i d'una VSI puntual a *TTN*) es disposa de familiars, però l'única informació clínica que s'obté és que el pacient compleix els criteris *Task Force* per la diagnosi de MCA. P21 és un home de 48 anys amb 5 germans i un fill, en els quals es desconeix la possible afectació. Tres dels germans són portadors de la duplicació (encara que no de la variant puntual) i el fill comparteix tant la variant estructural com la puntual, motiu pel que es posen tots sota seguiment clínic. Les duplicacions es classifiquen com VPP.

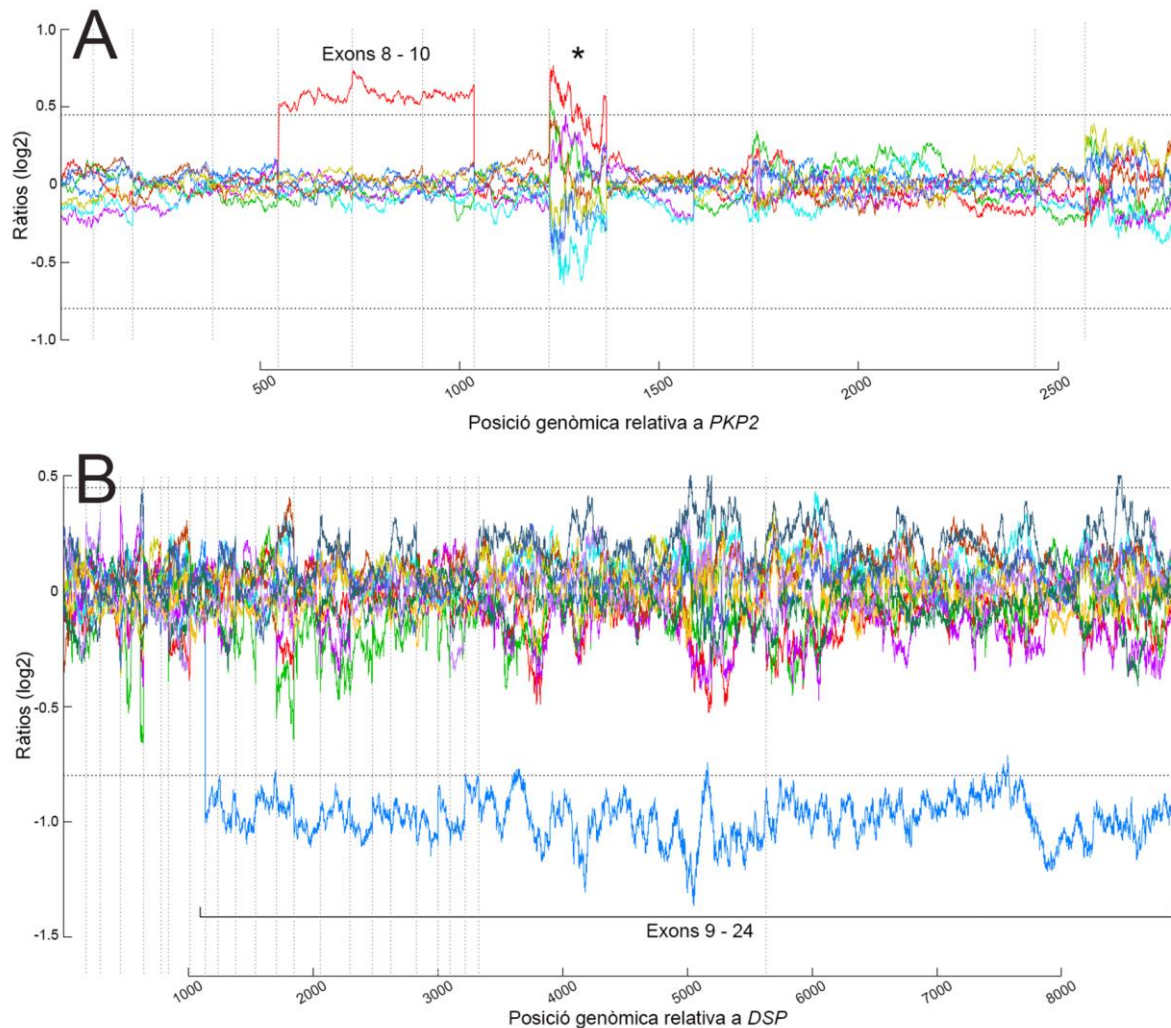


Figura 4-19 | (A) Duplicació dels exons 8 – 10 a *PKP2*, detectada als pacients P19, P20 i P21. L'asterisc marca l'exó 6 de *PKP2*, una regió amb marcada variabilitat en l'homogeneïtat de la seqüenciació, tal i com queda resumit a la Taula 4-2; **(B)** Deleció dels exons 9 – 24 a *DSP*, detectada al pacient P24.

Continuant amb les CNVs que afecten a *PKP2*, es detecten dos portadors de la deleció de l'exó 1. El primer, P22, és un home de 29 anys, amb deteriorament de la capacitat funcional, però sense episodis de síncope o palpitations. Mostra un ECG suggestiu de MCA i la RMC revela un àpex cardíac més trabeculat del que es considera normal. Per aquests motius se li diagnostica una MCA incipient. Existeix història familiar de MS: el pare va morir als 38 anys, havent patit episodis de síncope (sempre després de l'exercici) i amb un patró d'ECG alterat que mai va investigar-se amb profunditat. El seu bessó es sotmet a les proves clíniques i no mostra evidències d'afectació, però no es disposa de la informació genètica. L'altra germana, una dona de 27 anys, presenta un ECG similar al de P22 i resulta portadora de la deleció, motiu pel que se la posa sota seguiment clínic. El segon,

P23, és un home de 48 anys sense informació clínica ni familiar disponible, en el que també es detecten 2 VSI puntuals a *TTN*. En els dos casos, les delecions es classifiquen com VP.

Al pacient P24, de 28 anys d'edat i portador de DAI, se li detecta la delecio dels exons 9 – 24 a *DSP* (Figura 4-19/B) i una VSI puntual a *TGFBR2*. El pacient presenta palpitations després de practicar esforç físic, taquicàrdia ventricular monotònica sostinguda incessant i una disfunció marcada de ventricle esquerre. La mare del pacient, una dona de 56 anys, també és portadora de la delecio. Pateix disfunció ventricular lleu, cardiomegàlia, i es troba sota seguiment clínic. La delecio és classificada com a VP.

Les variants detectades en els pacients P19 – P21 són localitzades en gens desmosomals associats a MCA. Les alteracions en aquests gens poden alterar l'expressió o la funció de les proteïnes codificades, interferint en el correcte ensamblatge del desmosoma. Les delecions es classifiquen com VP per la hipotètica pèrdua de funció que poden provocar a la proteïna, mentre que les duplicacions es classifiquen com VPP.

L'últim portador de CNV del grup és P25, un home de 39 anys amb la delecio de la regió codificant de *MYOZ2* (i probablement del gen sencer), a més de 3 VSI puntuals a *MYH7*, *PKP2* i *RYR2*. *MYOZ2* codifica per una proteïna de la família de les miozenines. Aquestes juguen un rol important en la modulació de la senyalització de la Calcineurina, una fosfatasa involucrada en la transducció de la senyal dependent de calci en diversos tipus cel·lulars. A més, es creu que poden jugar un paper en la miofibril·logènesi. Al no disposar-se d'informació clínica ni familiar, i al no estar el gen directament associat amb la MCA, la delecio és classificada com VSI per exclusió.

IV – Pacients diagnosticats amb NCVE

L'únic pacient de NCVE portador d'una CNV detectat entre els 32 pacients estudiats (3.1% del grup) és P26, un home de 44 anys del que no es té informació familiar. Presenta un fenotip de no-compactació amb disfunció lleugera del ventricle esquerre i moderada del dret. En aquest pacient s'ha detectat la variant estructural més extensa de tota la cohort (Figura 4-20). Consisteix en una duplicació de 176,8 Kb que abasta la regió compresa entre els exons 45 – 275 de *TTN* (ambdós inclosos). Al pacient també se li detecta una VSI puntual a *ANK2*. El gen *TTN* codifica per la Titina, la proteïna humana més llarga. Consisteix en 364 exons que, en totes les seves variants de *splicing*, codifiquen per entre 5000 i 34000 aminoàcids. Degut a la seva extraordinària mida i a la ubiqüitat de variants al llarg del genoma, aproximadament un 3% dels individus són portadors de variants a *TTN*. Els defectes en aquest gen han estat associats a miocardiopaties com MCH i MCD. En aquest cas, la duplicació és classificada com VSI perquè, fins al moment, *TTN* no ha estat associat a NCVE i tampoc es té accés a més informació amb la que continuar investigant aquesta mutació. No obstant, s'ha reportat que la disrupció de l'associació entre la Titina i la Teletonina (codificada pel gen *TCAP*) altera

la contractilitat del miocardi (471). A més, Hastings et al. reporten una variant missense a *TTN* que cosegrega perfectament en una família de tres generacions d'individus diagnosticats amb NCVE (472). Tot això són motius suficients com per considerar aquesta variant (tot i la seva classificació) com la causa genètica més probable de la malaltia.

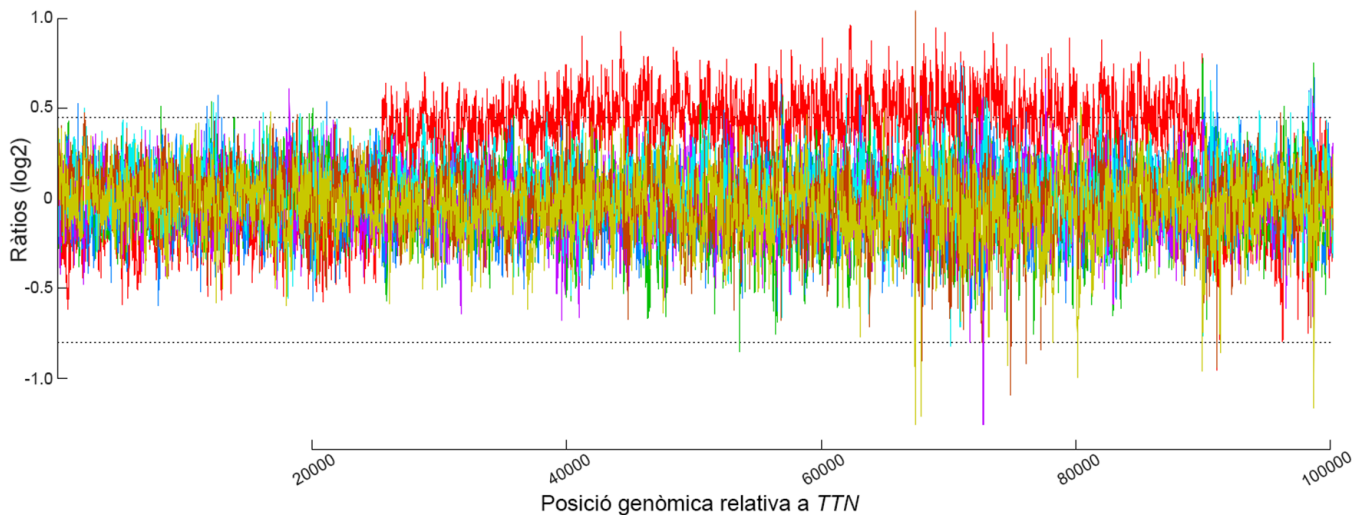


Figura 4-20 | Duplicació de 176,8 Kb (exons 45 – 275) a *TTN*, detectada en el pacient P26.

El primer estudi robust, centrat en la identificació de CNVs en un número considerable de gens associats a MCH va ser el publicat al 2015 per Lopes et al. (337). En aquest es reporta el cribratge de 19 gens associats a MCH en un grup de 505 pacients, detectant-ne 4 (un 0.8% de la cohort d'estudi): 1 deleció a *MYBPC3*; 1 deleció a *PDLIM3*; 1 duplicació a *TNNT2*; i una duplicació a *LMNA*. Les dues delecions van ser classificades com VPP, mentre que les duplicacions van ser catalogades com VSI. Recentment, Ceyhan-Birsoy et al. (338) van realitzar el cribratge de 18 gens associats a MCH (i 46 altres que teòricament haurien de cobrir l'espectre de gens associats a les altres miocardiopaties) en una cohort de 708 pacients. Van detectar-ne 4 (un 0.56% de la cohort d'estudi): 1 duplicació a *MYOZ2*; 1 deleció a *MYBPC3*; la duplicació del gen *NEXN*; i la duplicació dels gens *GLA*, *LAMP2*, *EMD* i *TAZ* en un pacient amb trisomia del cromosoma X. Només van considerar com una VP la deleció a *MYBPC3*.

La freqüència de detecció de CNVs en pacients de miocardiopatia a la nostra cohort (1.4% en MCH, 4.8% en MCA i 3.1% en LVNC) no és significativament superior a la reportada per Lopes et al. (0.8% en MCH) (337) i Ceyhan-Birsoy et al. (0.56% en MCH, 1% en MCA i 1.9% en LVNC) (338), a excepció del subgrup de pacients de MCD (5.8% contra 0.6%; $p=0.000036$). Aquesta diferència es deu o bé a les característiques de la cohort de MCD o bé a que el mètode de detecció de CNVs utilitzat per Ceyhan-Birsoy et al. no era tant sensible com el nostre.

4.3.2 – CNVs identificades a la cohort de canalopaties

I – Pacients diagnosticats amb SBr

El cribratge genètic del grup de 194 pacients de SBr va resultar en la identificació de 3 portadors de CNVs (1.55% del grup). El primer cas, P27, és una dona de 48 anys diagnosticada de SBr després d'un episodi sincopal. L'ECG va mostrar un patró de Brugada de tipus I a V1 i V2 i el test de flecainida va resultar positiu, motiu pel que se li va implantar un DAI. L'única variant rellevant detectada en la pacient va ser la duplicació en estat de mosaic que abasta des de l'exó 15 al 28 de *SCN5A* (Figura 4-14). L'avaluació clínica i genètica dels familiars (els pares, el fill i la filla) va revelar que no presentaven símptomes de SBr i que, a més, no eren portadors de la duplicació. La CNV va ser classificada com a VPP, ja que *SCN5A* és el gen principal associat a SBr. El cas va ser publicat de manera detallada pel nostre grup (Annex 5) (386).

El segon cas és el de P28, un home de 58 anys i amb història de MS familiar: l'avi patern mor amb 64 anys i el pare als 57. El pacient va ser diagnosticat després d'un ECG suggestiu i d'un test d'ajmalina positiu. Se li va implantar un DAI fa 8 anys i mai ha rebut cap descàrrega del dispositiu. En aquest pacient es va detectar la duplicació de *TRDN* i 1 VSI puntual a *SLMAP*. No es va poder realitzar l'estudi de cosegregació familiar. El tercer i últim portador és P29, un home de 52 anys del que no es disposa ni d'informació clínica ni familiar. Se li va detectar la duplicació de la regió compresa entre els exons 2 – 11 de *CASQ2* (ambdós inclosos) i una VSI puntual al mateix gen, a l'al·lel duplicat (en funció de la freqüència de l'al·lel alternatiu observada). Les duplicacions observades a P28 i P29 van ser classificades com VSI, ja que els gens *TRDN* i *CASQ2* no s'han associat a SBr. A més, són casos clínics en els que falta informació important –la cosegregació familiar, per exemple– per poder reforçar la sospita d'associació de la variant amb el fenotip. Tot i així resulta interessant haver detectat les CNVs en dos gens en els que els defectes genètics s'associen amb la TVPC. Les proteïnes codificades per *TRDN* i *CASQ2* estan relacionades en major o menor mesura amb la regulació del calci intracel·lular de les cèl·lules musculars cardíques. És àmpliament conegut que la desregulació d'aquesta via facilita l'aparició d'arrítmies (406). Tot i així, manquen estudis que recolzin un possible solapament de l'arquitectura genètica de la SBr i la TVPC.

II – Pacients diagnosticats amb SQTL

L'estudi genètic dels 145 pacients diagnosticats amb SQTL va donar peu a la identificació de 6 portadors de CNVs (4.1% del grup). Cinc de les variants van detectar-se en diferents subunitats de canals de potassi i van ser classificades com VP o VPP.

La deleció dels exons 7 i 8 de *KCNQ1*, prèviament reportada i associada a SQTL per Barc et al. (393) va ser detectada en 3 pacients diferents. Al primer, P30, una dona de 39 anys, se li va detectar també 1 VPB a *MYH6* i 2 VSI a *DSP* i *TTN*. El seu pare havia estat diagnosticat de SQTL i va patir una MS no recuperada. El segon portador és P31, una noia de 17 anys per la que l'ECG basal mostra un QTc (interval QT corregit) de 500 ms sense estar sota efecte de cap medicació i sense patir cap alteració iònica que pogués explicar la prolongació del QT. L'estudi familiar va revelar l'afectació de la mare i del germà –de 10 anys–, que també van resultar ser portadors de la deleció. La informació detallada del cas va ser publicada pel nostre grup (394). El tercer cas és el de P32, una dona de 42 anys en la que també s'identifica 1 VSI puntual a *KCNQ1*. El pare i el germà són també pacients de SQTL, al igual que la filla, que mor als 10 dies de vida. L'estudi familiar va demostrar que tots els afectats eren també portadors de la deleció, mentre que als sans no es detectava (Figura 4-21/B). Pels 3 casos, la CNV va classificar-se com una VP.

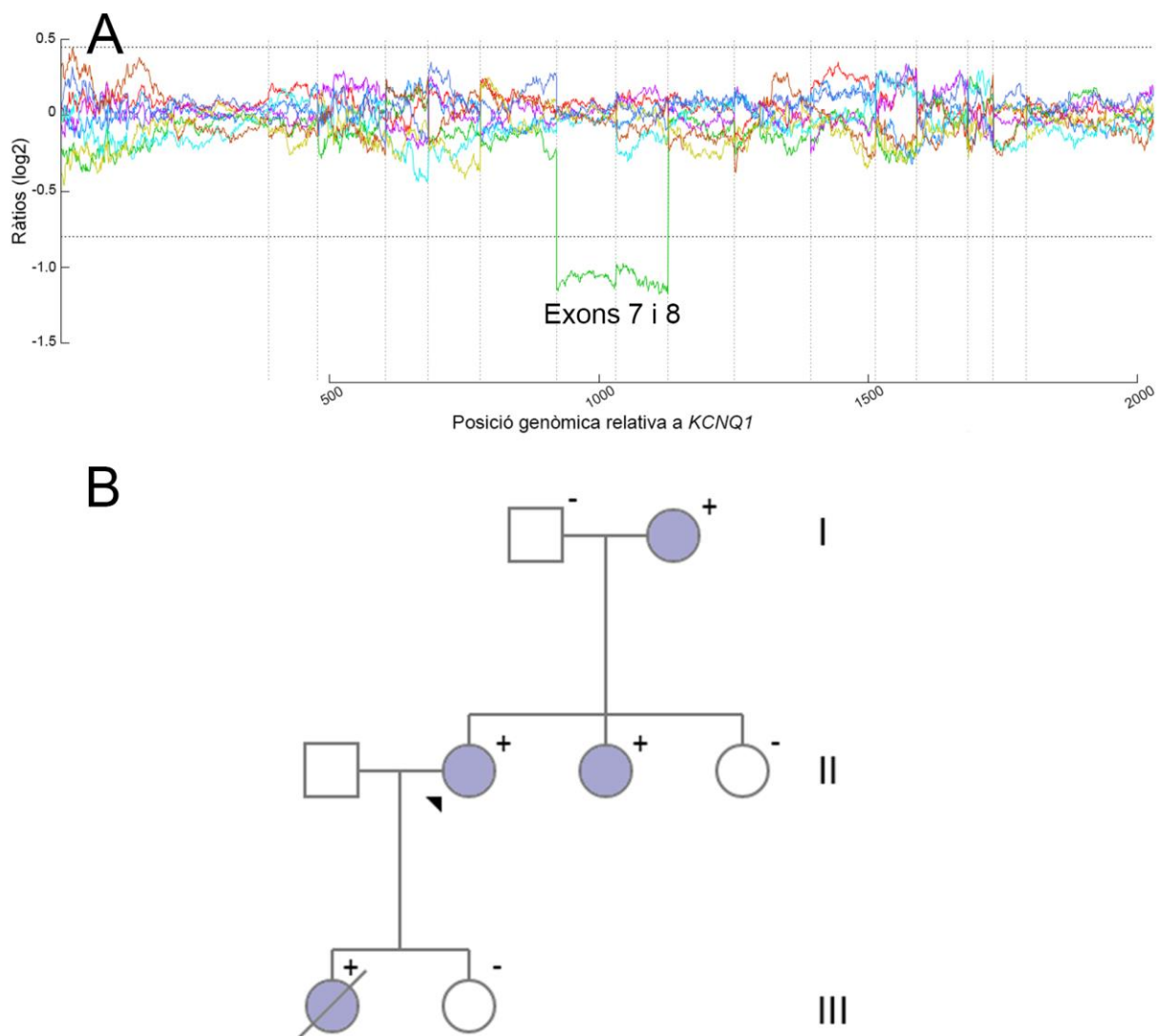


Figura 4-21 | (A) Deleció dels exons 7 i 8 a *KCNQ1* detectada als pacients P30, P31 i P32; **(B)** Pedigrí del pacient P32. El pacient s'assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol "+" indica que l'individu és portador de la variant, i el "-", que no ho és portador.

A P33, una dona de 22 anys, se li va detectar la deleció dels exons 1 – 14 a *KCNH2* i 2 VSI puntuals a *KCNE1* i *AKAP9*. No es disposa ni d'informació clínica ni familiar, però la variant havia estat prèviament associada a SQT (393) i va ser considerada VP. A P34, una dona de 37 anys, es detecta la deleció dels exons 3 i 4 de *KCNE1* (deleció que inclou tota la regió codificant del gen) i una VSI puntual a *CACNA1C*. La deleció va detectar-se en alguns familiars de la pacient, però tampoc es té informació clínica que recolzi la segregació. La CNV va ser considerada VPP, ja que *KCNE1* és un gen associat amb la SQT. Finalment, a la pacient P35, una nena de 5 anys, se li detecta la duplicació de la regió codificant de *NEXN*. Aquest gen codifica per una proteïna filamentosa d'unió a actina que intervé en l'adhesió i la migració cel·lular. Les mutacions reportades per *NEXN* s'associen a MCH i MCD. Per aquest motiu, la CNV és considerada com una VSI. No obstant, a P35 se li detecta també un *frameshift* a *KCNQ1* (p.Leu496Alafs*19) classificat com a VP i que podria explicar el fenotip de la pacient.

III – Pacients diagnosticats amb FA

Entre els 15 pacients de FA inclosos a la cohort s'ha detectat un únic portador de CNV (6.7%). P36 és un home de 50 anys amb FA persistent diagnosticada als 34 anys. Se li han realitzat diversos intents de cardioversió elèctrica i ha estat en tractament amb flecainida, sense èxit. Als 10 anys de la diagnosi se li realitza una RMC en la que es troba una petita escara que fa sospitar de miocardiopatia primària. L'ecocardiograma revela que la fracció d'ejecció del ventricle esquerre és del 35%. Se li diagnostica una MCD amb disfunció ventricular moderada, sense insuficiència cardíaca. En aquest pacient es detecta la duplicació del gen *KCNJ5*, juntament amb 3 VSI puntuals, 2 a *TTN* i 1 *ANK2*. El pacient té un germà bessó diagnosticat amb FA als 35 anys. És pendent de visita i es desconeix si té disfunció ventricular i si és portador de les variants detectades en el seu germà. Tot i que bona part de les subunitats dels canals de potassi s'hagin associat a FA, no és el cas per *KCNJ5*. Per aquest motiu, la variant va classificar-se com VSI. Tot i així, seria interessant conèixer la implicació en el fenotip de la duplicació i de la VSI a *ANK2* de manera conjunta. La variant puntual, tot i ser considerada VSI, no ha estat prèviament reportada però els predictors *in silico* prediuen un efecte deleteri.

Els estudis previs publicats de detecció de CNVs en pacients de canalopaties són molt escassos. Per la SBr, tan sols hi ha l'estudi publicat pel nostre grup (386), que inclou el cribratge de *SCN5A* juntament amb altres gens minoritaris entre pacients d'aquesta malaltia. A banda d'aquest, Eastaugh et al. van reportar una deleció intragènica a *SCN5A* en un pacient de SBr, classificada com VP (385). En relació amb la SQT, les petites sèries publicades (393,394) consisteixen en l'estudi de 2 a 5 gens associats amb la malaltia (*KCNQ1*, *KCNH2*, *KCNE1*, *KCNE2* i *SCN5A*), amb una taxa de detecció de CNVs entre el 2 i l'11.5%. La nostra taxa de detecció de CNVs en aquesta porció de la cohort és del

4.1% (6/145 pacients), i 5 de les 6 CNVs detectades van ser-ho en aquests 5 gens. Tot i que els nostres resultats siguin compatibles amb les freqüències prèviament reportades, s'ha de tenir en compte el baix número de pacients analitzats als estudis previs, fet que provoca que no siguin percentatges comparables entre ells. Per la SQTC no s'ha reportat mai cap CNV, i per la TVPC s'han identificat diverses CNVs a *RYR2* (371,372). Fins aquesta tesi no s'havia investigat mai la presència de CNVs en altres gens relacionats amb les malalties. La nostra freqüència de CNVs per aquestes dues fraccions de la cohort és del 0%, però sense dubte això és degut al baix número de pacients inclosos a la cohort, concretament 4 i 21, respectivament.

4.3.3 – CNVs identificades a la cohort de MSI

La cohort de 704 casos de MSI inclou 51 casos de MSL i 2 de MSIU. En total van identificar-se 8 portadors de CNV entre els casos de MSI (1.14% del grup) i 1 entre els casos de MSL (2% del grup).

El primer cas és el de P37, un home de 42 anys en el que es detecta la duplicació dels exons 2 – 5 al gen *EMD*, i 1 VSI puntual a *TTN*. Aquest és un cas similar al de P38, un home de 22 anys en el que es detecta una reordenació genòmica complexa a *TNNI3*, en la que els exons 4,5 i 8 són duplicats. En els 2 casos, les CNVs van ser classificades com VPP, ja que defectes en aquests dos gens han estat relacionats, a banda de amb diverses miocardiopaties, amb FA i problemes de conducció cardíacs (473,474).

Dos dels casos més interessants són els de P39 i P40 (47,XYX). En aquests dos homes de 48 i 28 anys, respectivament, es detecten dues duplicacions prèviament detectades en altres pacients de la cohort. La primera (P39, mateixa que a P27 –pacient de SBr–), és la duplicació del gen *TRDN*; la segona (P40, mateixa que a P28 –també pacient de SBr–) és la duplicació dels exons 2 – 11 de *CASQ2*. De la mateixa manera que passava amb els pacients de SBr, no es disposa ni d'informació clínica ni de familiars. Les duplicacions no havien estat reportades prèviament i, per tant, no són associades a cap fenotip. Tot i classificar-se com a VSI, s'acumula evidència a favor de que aquestes CNVs tinguin un rol arritmogènic, ja que es detecten en pacients amb un cor estructuralment normal en els que el primer símptoma de la malaltia és una MS d'origen suposadament cardíac.

El cas de P41 és el d'un noi de 17 anys, portador de DAI, que pateix una aturada cardiorespiratòria durant un partit de futbol. El pacient es pot recuperar i durant el trajecte a l'hospital pateix una altra aturada. En cap moment se li registra una arrítmia, convulsions febrils ni epilèpsia, com tampoc síncope ni pèrdua de coneixement. Presenta un QTc dins de la normalitat i se li practica un test de flecainida que resulta dubtós i una estudi electrofisiològic que no indueix a arrítmies. A la família hi ha història de MS: l'àvia materna va morir als 33 anys, a l'anar a dormir perquè es trobava malament. No se li va practicar autòpsia. La mateixa dona va tenir un avortament als 8 mesos de gestació. Un dels germans de l'àvia va patir una MS als 17 anys mentre nedava. No se li va practicar

l'autòpsia i se li va comunicar a la família que havia estat un "tall de digestió". Un cosí del pacient per part materna és diagnosticat amb epilèpsia. La mare refereix un episodi de síncope mentre anava en autobús. Es considera de dubtosa associació al succeir en un moment de molt d'estrès per la mort de la seva àvia uns dies abans. A P41 se li va detectar la deleció dels exons 12 i 13 de *CTNNA3*, juntament amb la deleció del gen *GAA* (Figura 4-22/A-B). És l'únic cas registrat en tota la cohort en el que es detecten dues CNVs independents a les regions estudiades. També van detectar-se 2 VSI puntuals a *MYPN* i a *HCN2*. L'estudi de cosegregació realitzat al germà i als 2 pares va revelar que la mare era portadora de les dues deleccions i la VSI a *MYPN*, i el germà de la deleció a *GAA* (Figura 4-22/C). *CTNNA3* codifica per una proteïna de la família de les α -catenines que juga un paper important en les unions cèl·lula-cèl·lula entre els miòcits. Els defectes en aquest gen s'associen amb MCA. Al seu torn, *GAA* codifica per l' α -glucosidasa lisosomal, encarregada de la degradació del glicogen cap a glucosa. Les alteracions a *GAA* s'associen amb la malaltia de Pompe, caracteritzada per errors al metabolisme del glicogen. Depenent de la mutació, existirà una deficiència total o parcial de l'activitat de l'enzim a totes les cèl·lules de l'organisme, que pot tenir conseqüències sobre diferents teixits. L'efecte més notable té lloc a les cèl·lules musculars, on s'acumula una gran quantitat de glicogen residual que és absorbit pels lisosomes. L'emmagatzematge extens de glicogen pot interferir amb la funció cel·lular, causant danys a les cèl·lules. Tenint en compte l'haploinsuficiència de *GAA* i el genotip complex que presenta, sembla lícit associar l'efecte combinat de les variants amb el fenotip. Tot i així, cenyint-nos als criteris de classificació establerts, les CNVs són classificades com a VSI. No obstant, cal tenir en compte que el pacient pot ser encara massa jove com per mostrar defectes cardíacs estructurals.

Pels 4 casos restants, no es disposa d'informació clínica ni familiar i les CNVs identificades són classificades com VSI. P42 és l'únic cas de portador de CNV en el grup de MSL. És una nena que mor abans del primer any de vida. Se li detecta la duplicació del gen *RANGRF*, associat amb SBr. P43 és el cas d'un home de 48 anys en el que es detecta la duplicació de *PDLIM3* i una VSI puntual a *MYPN*. Els defectes, també estructurals, en aquest gen han estat associats amb MCH i MCD (337). Seria necessària més informació clínica i familiar per poder establir una associació directa entre la CNV i el fenotip, però tot i així, aquesta és una candidata plausible a ser la causa genètica del fenotip. A P44, una dona de 38 anys, se li detecta la duplicació dels gens adjacents *TAZ* i *EMD*, juntament amb 1 VSI puntual a *PDLIM3*. Per últim, la duplicació dels gens adjacents *KCNE1* i *KCNE2* a P45 va ser considerada com una VPB al veure's en més de 10 individus a DGV.

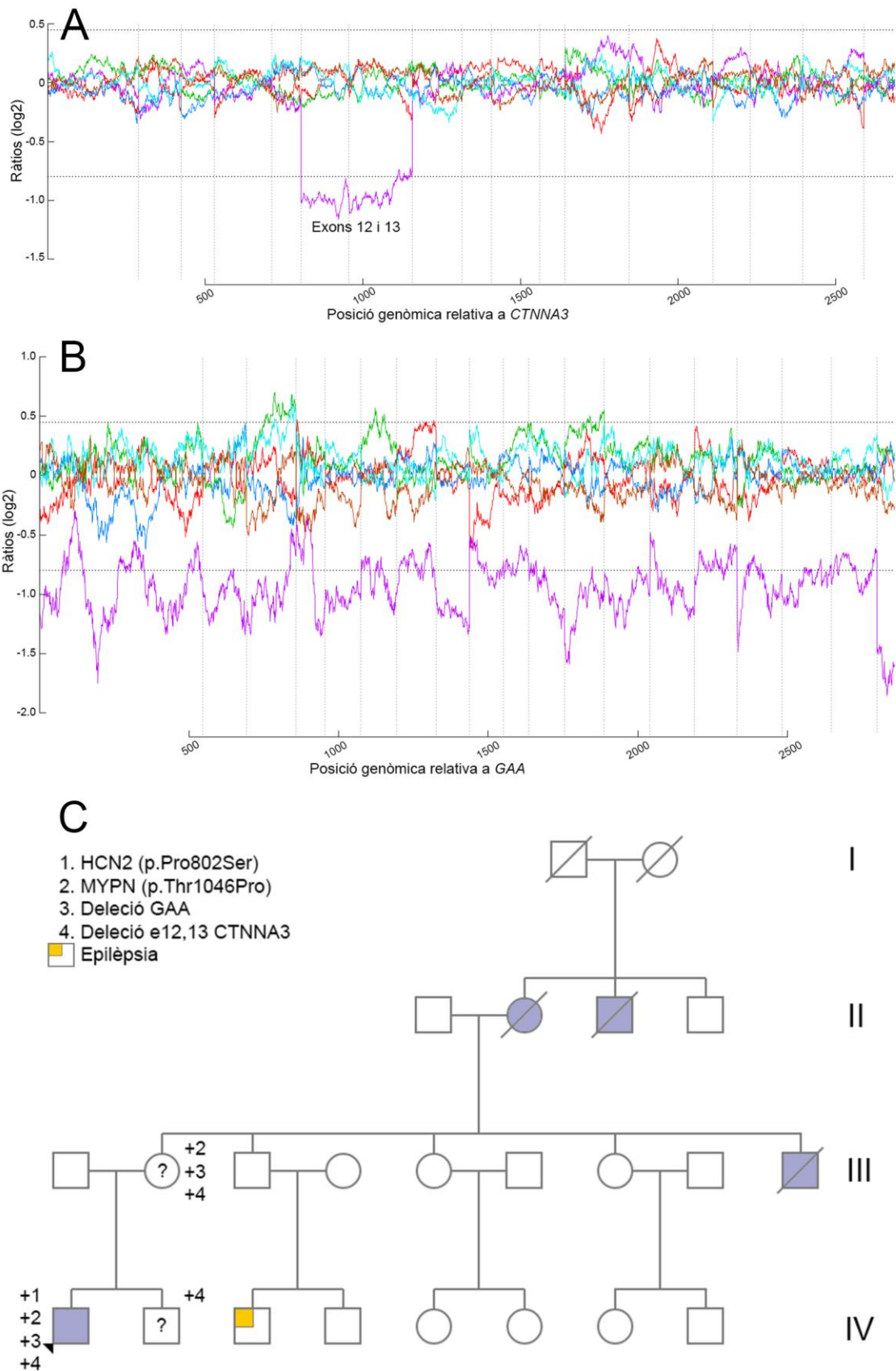


Figura 4-22 | Deleció dels exons 12 i 13 a *CTNNA3* **(A)** i del gen *GAA* **(B)** detectada al pacient P41; **(C)** Pedigrí del pacient P41. El pacient s'assenyala amb una fletxa. Els afectats són els individus acolorits. El símbol "+" indica que l'individu és portador de la variant.

Malauradament, degut a la naturalesa dels casos de MSI, per la gran majoria de portadors no es disposa d'informació clínica i/o familiar. Sovint la MS esdevé com el primer símptoma de la malaltia, abans de qualsevol diagnòstic, fet que complica la investigació. Els resultats obtinguts per la cohort de MSI no poden ser comparats amb cap altre estudi previ, ja que aquest treball constitueix el primer cribratge de CNVs en una cohort extensa de casos d'aquest tipus. Únicament hi ha una publicació prèvia en la que s'investiga la incidència de CNVs en casos de MSI mitjançant aCGH (475). En 27 casos es van detectar 3 CNVs. No obstant, aquestes eren grans (superiors a 240 Kb, involucrant diversos gens) i almenys una podria estar associada a un fenotip sindròmic.

4.3.4 – CNVs identificades als pacients de SM i DATA

En els 36 pacients de SM i DATA van detectar-se 2 portadors de CNVs a *FBN1* (5.6%). Les variants estructurals són detectades amb una freqüència relativament elevada en aquests pacients i s'associen a la malaltia (419). Per aquest motiu, les dues van considerar-se com VPP.

El primer cas és el de P46, un noi de 17 anys en el que es detecta la duplicació dels exons 45 – 65, juntament amb 2 VSI a *TTN* i *KCNH2* i una VPB a *TTN*. El segon cas és el de P47, una nena de 8 anys en la que es detecta la deleció de l'exó 55 i una VSI puntual a *TGFBR2*.

V. Discussió general

Més de 4 milions de persones moren anualment, arreu del món, a causa de la MSC (290). Considerant a banda la malaltia coronària, que és la causa més comuna de MSC en adults (en un 75-80% dels casos), les malalties cardíques arritmogèniques suposen la principal causa de MSC entre la població menor de 35 anys. En aquest col·lectiu, el 10-15% de casos de MSC són deguts a miocardiopaties, com la MCH, la MCD, la MCA, la NCVE i, en menor mesura, la MCR. El 5-10% de casos restants són els causats per anomalies elèctriques en absència de defectes estructurals del cor. Entre aquestes malalties, conegudes com canalopaties (per estar relacionades amb l'afectació dels canals iònics de les cèl·lules cardíques), trobem majoritàriament la SQTL, la SQTC, la TVPC, la SBr i la FA (290,293).

Les malalties cardiovasculars arritmogèniques d'etiologia genètica són considerades d'herència mendeliana. Tot i així, l'evidència acumulada durant l'última dècada suggereix patrons d'herència complexos per algunes d'elles (300,301). Aquestes malalties comparteixen certes característiques, com l'heterogeneïtat genètica i fenotípica (81), l'expressivitat variable i la penetrància incompleta. En alguns casos, el solapament fenotípic i genètic entre els diferents desordres dificulta la diagnosi i el consell genètic (354,356). Per aquest motiu, la identificació d'una mutació causal en un pacient és crucial per la confirmació del diagnòstic en els casos límit, pel maneig precoç dels familiars en situació de risc i per evitar el seguiment innecessari de familiars no portadors.

L'arribada de les tecnologies de seqüenciació d'alt rendiment a la pràctica clínica ha millorat indiscutiblement el camp de la diagnosi genètica. Al reduir el temps, la mà d'obra i el cost dels reactius destinats a la seqüenciació, es pot generar una quantitat d'informació extraordinària per cada pacient de forma ràpida i eficient. Però, per altra banda, la informació clínica associada a la majoria de variants identificades és escassa. Si s'hagués de comprovar l'impacte fenotípic de totes les variants trobades en un pacient, la diagnosi genètica seria inviable: la inversió econòmica i de temps seria molt elevada i el sistema, en general, insostenible. És per aquest motiu que existeixen unes directrius clíniques internacionals que recomanen quines regions del genoma seqüenciar en funció de la diagnosi del pacient i com classificar les variants genètiques en base a la seva rellevància (422). Així doncs, les guies clíniques recomanen la seqüenciació dels gens en els que es registren freqüències superiors de detecció de variants patogèniques (encara que hi hagin altres gens associats a la malaltia). Aquest conservadorisme de les recomanacions internacionals es deu a la necessitat d'optimitzar la inversió destinada a la detecció de les variants i a la seva transformació en un resultat útil i interpretable per un cardiòleg i/o forense. És desitjable, doncs, trobar l'equilibri entre una diagnosi clínica concisa i informativa i l'avenç de la recerca.

Tot i la millora en el camp de la diagnosi genètica, propiciada pel desenvolupament de les tecnologies de seqüenciació d'alt rendiment, el percentatge de casos que resulten sense una causa genètica un cop realitzat el cribratge és encara elevat. Les proporcions de casos diagnosticats i amb VP detectades en gens associats amb miocardiopaties es troben al voltant del 70-80% per la MCH, del 50-60% per la MCD i la MCA, i en un 35% i 75% per la NCVE (en funció de si el pacient és un adult

o un nen, respectivament) (309–312). Per les canalopaties, aquests percentatges són aproximadament del 25-35% per la SBr, del 75-85% per la SCTL, del 50-60% per la SCTL i d'un 65% per la TVPC (313–316). És possible que les causes genètiques dels casos no resolts després del cribratge siguin VPs en gens no associats fins al moment amb la patologia, que es trobin en regions reguladores no incloses en els cribratges genètics (317), o que la causa del fenotip siguin alteracions epigenètiques (318).

Un altre tipus de variants potencialment patogèniques i responsables del fenotip d'aquests pacients són les variants estructurals. Les CNVs (de l'anglès *Copy Number Variants*) són reorganitzacions genòmiques desequilibrades superiors als 50 pb (421). Durant l'última dècada, la detecció de CNVs s'ha reivindicat com una estratègia particularment útil en el descobriment de *loci* i gens associats a desordres complexos i fenotípicament heterogenis, com la discapacitat intel·lectual i les anomalies congènites (269,476). No obstant, en el camp de la cardiologia, el cribratge d'aquestes variants estructurals s'ha relegat a les malformacions cardíques congènites (477). Els cribratges en pacients de malalties cardíques arritmogèniques que inclouen la detecció de CNVs conjuntament amb les ja comunes SNVs i *indels* és molt poc habitual. Tot i que s'han reportat evidències que relacionen aquest tipus de variants amb les malalties associades a la MSC (revisió exhaustiva a l'apartat 1.4 d'aquesta tesi), no existeixen unes directrius fermes que recomanin la inclusió de les CNVs en el cribratge rutinari d'aquests pacients. Fins i tot el *Broad Institute* (Cambridge, USA) ofereix la detecció de CNVs a la cartera de serveis genòmics (478), però aquesta no és específicament recomanada a les guies clíniques internacionals (422). La prevalença de les CNVs, l'associació amb el fenotip del pacient i, per tant, la incorporació en el context de la diagnosi genètica és una qüestió que tot just ara comença a explorar-se a un nivell més sistemàtic i exhaustiu, en grans cohorts de pacients, i considerant una ampla varietat de gens associats i candidats (337,338,466).

Aquesta manca d'informació en quant a la prevalença de les CNVs en la MSC i en les malalties associades va posar de relleu, a l'inici d'aquesta tesi, la necessitat de desenvolupar un mètode de detecció de CNVs a partir de dades provinents de panells de captura de regions genòmiques associades a la MSC. Al tenir com a objectiu l'anàlisi de mostres en un context clínic, aquest mètode havia de ser sensible i precís per evitar els falsos negatius, però també específic i robust, per obtenir després de cada anàlisi una quantitat assumible de senyals per validar. Les dificultats tècniques que s'haurien d'afrontar durant el desenvolupament del mètode són les inherents a la tecnologia de seqüenciació de fragments curts. La uniformitat de les cobertures a les regions seqüenciades és sensible al disseny de les sondes de captura i als biaixos en l'enriquiment, causats tant per la tecnologia com per les propietats intrínseques de les seqüències de les regions d'interès – per exemple el contingut GC o la complexitat i/o repetitivitat d'aquestes—. Per altra banda, la manca de controls positius (i sobretot negatius) amb els que comparar les mostres problema dificulta la detecció de les CNVs. Per tant, l'estratègia d'anàlisi va centrar-se en el desenvolupament d'un algoritme que processés el resultat de la seqüenciació a partir d'uns dissenys de sondes optimitzats. Aquest algoritme hauria d'avaluar la idoneïtat de les mostres de l'anàlisi, corregir –si fos possible– els

biaixos causats pel protocol de preparació de llibreries genòmiques (o per la mateixa plataforma de seqüenciació) i realitzar la comparació de cobertures sota el supòsit de que la freqüència de CNVs en regions codificants és baixa (90–92), podent considerar així el conjunt de mostres analitzades com pseudocontrols. El processament de la informació associada a cada potencial CNV hauria de permetre un bon filtratge qualitatiu, que resultés en un conjunt final de senyals d'alta fiabilitat. Un cop validades mitjançant qPCR o MLPA, la informació obtinguda amb l'anàlisi genètica es relaciona amb les dades clíniques i familiars del pacient per poder extreure una conclusió que afavoreixi la diagnosi.

Amb el desenvolupament del mètode de detecció, en aquesta tesi s'ha realitzat el cribratge de CNVs més extens mai publicat per pacients de MSC i de malalties associades, juntament amb una cohort de casos de MSI. Les 2073 mostres s'han recol·lectat durant els últims 5 anys i provenen d'un total de 15 centres hospitalaris de l'estat espanyol. En el cribratge s'han inclòs tant els gens associats amb major prevalença com una àmplia selecció de gens minoritaris i candidats.

Per tal de determinar si el mètode de detecció de CNVs és robust i fiable i, per tant, adequat per l'anàlisi de mostres clíniques, cal avaluar cada un dels components que el constitueixen.

Els nostres dissenys de sondes s'optimitzen per capturar de manera acceptable aquelles regions difícils de seqüenciar –ja sigui per la baixa complexitat de la seqüència, l'homologia amb altres regions del genoma o per presentar continguts de GC extrems–. Les mostres seqüenciades amb aquests dissenys exhibeixen una gran homogeneïtat de cobertura al llarg de més d'un 99% de les regions capturades, així com una elevada cobertura global per mostra (els diferents paràmetres són exposats detalladament a l'apartat 4.1). Aquests són uns resultats molt notables, tenint en consideració que les mostres són generades amb una tecnologia amb limitacions evidents i seqüenciades en una plataforma de capacitat molt moderada, com és el MiSeq –un seqüenciador de sobretaula de segona generació–. Les regions que, tot i l'optimització, es capturen de manera irregular (Taula 4-2) formen part del percentatge de regions genòmiques difícils (i algunes impossibles) de seqüenciar de manera acceptable amb la tecnologia de seqüenciació de fragments curts.

Amb la qualitat de les mostres seqüenciades, l'algoritme de detecció de CNVs desenvolupat en aquesta tesi demostra una alta sensibilitat –incloent la detecció d'alteracions d'un únic exó– i un número assumible de falsos positius. Això resulta de particular importància, ja que aquest tipus de senyals acostumen a ser discriminades, sobretot a les anàlisis de dades provinents d'exoma (337,479). A l'etapa de validació (apartat 4.2.2), l'algoritme va demostrar una sensibilitat del 100% i una especificitat del 91%. En comparar els resultats de l'algoritme amb els de CNVKIT v.0.8.6 (461) i de CONTRA v.2.0.8 (280) –dos *softwares* de detecció de CNVs en dades provinents de mètodes de captura– el nostre algoritme va demostrar uns percentatges d'exactitud (99.9%, 99.9% i 99.6%, respectivament) i de sensibilitat (100%, 100% i 87.3%, respectivament) comparables a CNVKIT i

superiors a CONTRA. A nivell de precisió, el nostre algoritme va demostrar ser una mica superior als altres dos (85.9%, 83.3% i 75%, respectivament).

Per tant, el nostre mètode demostra una alta fiabilitat per la detecció de CNVs en mostres generades amb panells de seqüenciació per captura. Això el converteix en una eina candidata per ser implementada en les anàlisis rutinàries de mostres clíniques amb finalitats diagnòstiques i de suport a la clínica, independentment de les malalties a les que estiguin associades les regions seqüenciades.

El cribratge resulta en la detecció de 47 portadors de 48 CNVs (22 delecions i 26 duplicacions). Això equival a una taxa de detecció global del 2.3%. En un primer moment, aquesta taxa pot semblar modesta, però no és d'estranyar, donat que les CNVs són variants negativament seleccionades a les regions codificants del genoma humà (90,92). La taxa de detecció pels diferents grups de pacients és variable. Oscil·la entre el 0% detectat en pacients de SQTC, TVPC o Mort Sobtada Intrauterina –MSIU– (probablement pel número limitat de pacients i casos en els que s'ha pogut dur a terme el cribratge) i el 6.7% detectat en pacients de FA o el 5.8% en pacients de MCD (Taula 4-5). Tot i que les dades sobre la prevalença real de CNVs en pacients de malalties associades a la MSC és escàs, els nostres resultats són compatibles amb els estudis publicats fins al moment. L'única malaltia per la que aparentment hem detectat un número de portadors superior a l'esperat és la MCD. La nostra taxa és significativament superior a la publicada per Ceyhan-Birsoy et al. (5.8% contra 0.6%; $p=0.000036$), un dels dos cribratges de CNVs més exhaustius en pacients de miocardiopatia publicats fins a data d'avui (338). Considerem que aquesta diferència pot deure's tant a les característiques de la cohort d'estudi com al fet de que el mètode de detecció de CNVs utilitzat per Ceyhan-Birsoy et al. sigui menys sensible que el nostre (480). Tenint en compte els resultats de sensibilitat i de precisió obtinguts durant la validació del nostre mètode, aquesta sembla la hipòtesi més plausible.

En els casos en els que es detecta una CNV classificada com a VP o VPP (en 26 pacients, un 55.3% del total de portadors identificats), aquesta sembla ser la causa genètica més probable de la malaltia. No es detecten altres variants puntuals que puguin explicar el fenotip. Tant sols hi ha una excepció, la del pacient P2. En aquest s'identifica un genotip complex sense precedents en un pacient de MCH, amb la delecio intragènica de l'exó 27 de *MYBPC3* considerada VP i una VPP puntual al mateix gen. És un cas exemplar en el que es mostra la importància del cribratge rutinari de CNVs inclús en aquells pacients en els que s'hagi detectat prèviament una VP o VPP puntual. D'aquí es conclou que, encara que la taxa de detecció global de CNVs sigui relativament baixa, representa una fracció de la cohort que no pot menystenir-se sota cap concepte.

En quant a la classificació de les variants estructurals, les recomanacions internacionals actuals es centren en la interpretació de grans reorganitzacions cromosòmiques que, sovint, involucren diversos gens adjacents (269,481). No existeixen recomanacions detallades per la interpretació de CNVs intragèniques, tot i que sota el nostre punt de vista requereixen d'una consideració especial. Per

exemple, les duplicacions són considerades generalment menys deletèries que les delecions (269), però una duplicació intragènica pot ser igual de disruptiva que una deleción. En un intent d'aportar quelcom a la interpretació de les variants estructurals, en aquesta tesi s'han proposat diversos criteris per la seva classificació. La idea de fons és que si la pèrdua de funció d'un gen en particular és un mecanisme de patogenicitat conegut per la malaltia del pacient (precisament un dels criteris més determinants a l'hora de classificar les variants puntuals com VP, d'acord amb les recomanacions per la interpretació de variants (422)), qualsevol deleción o duplicació intragènica en tàndem –que no inclogui l'últim exó del gen– hauria de ser considerada com una VP o VPP. Aquesta CNV alterarà la codificació del gen, resultant en un transcrit aberrant, que pot ser degradat per *nonsense-mediated decay* (422), o ser traduït a una proteïna igualment aberrant que, o bé serà processada per la maquinària de degradació cel·lular (les dues possibilitats donarien lloc a una situació d'haploinsuficiència, al no disposar la cèl·lula del producte del gen d'aquest al·lel), o bé no podrà realitzar la seva funció de manera òptima. Aquesta declaració concerneix exclusivament a aquelles duplicacions en tàndem confirmades, però s'ha de tenir en compte que el 83% de les duplicacions es troben en tàndem i en orientació directa (482). Per altra banda, per millorar la classificació de les variants, és important disposar de la informació clínica i familiar dels portadors de CNVs. La cosegregació de la CNV en diversos membres afectats de la família recolza la patogenicitat de la variant, mentre que la no cosegregació dona peu a que aquesta sigui considerada benigna. De qualsevol manera, aquesta informació ajuda a reduir la proporció de VSI.

En una sisena part dels portadors de CNVs de la nostra cohort es detecta una duplicació intragènica, suposadament en tàndem. Segons les recomanacions internacionals per la classificació de variants, aquestes s'haurien classificat com VSI o VPB. En base a la nostra classificació (criteris descrits a l'apartat 3.7 d'aquesta tesi) i pels motius exposats anteriorment, les variants es classifiquen com VP o VPP. Entre aquests pacients trobem: el cas de P3, amb la duplicació dels exons 9 – 29 a *MYBPC3*; P14, amb la duplicació dels exons 5 – 10 a *LMNA*; el trio de pacients P19-P20-P21, amb la duplicació dels exons 8 – 10 a *PKP2*; la duplicació en mosaic detectada a *SCN5A* en la pacient P27; la duplicació dels exons 2 – 5 a *EMD* del pacient P37 i la reorganització complexa que resulta en la duplicació dels exons 4, 5 i 8 a *TNNI3* en el pacient P38 (en la que, gràcies a la localització propera dels exons, es poden validar els punts de trencament, confirmant així que és en tàndem). Tots aquests pacients presenten fenotips severos de la malaltia. A més, pels casos en els que es va poder realitzar la cosegregació familiar (del 100% en el cas de P14 i esclaridora en el cas de P27), la nostra classificació surt encara més reforçada.

La recerca clínica bàsica pot nodrir-se dels resultats obtinguts en un cribratge com el que s'ha realitzat en aquesta tesi. L'estudi funcional de les CNVs identificades –ja sigui en gens associats prèviament a la malaltia o no– pot resultar d'utilitat per entendre millor (o descobrir) les bases moleculars i els mecanismes fisiopatològics subjacents a la malaltia del pacient. En aquesta tesi trobem casos paradigmàtics d'aquesta situació, com els dels pacients de SBr (P28 i P29) i MSI (P39 i

P40) amb duplicacions en (o dels) gens associats amb la regulació dels nivells de calci intracel·lular a les cèl·lules musculars cardíques –i associats a TVPC–. Malauradament, la informació de la que s’ha disposat per aquests pacients ha estat escassa. Els estudis funcionals podrien aportar una mica més de llum a una possible associació molecular entre la SBr i la TVPC, de manera similar a la que existeix entre la SBr i la MCA, publicada recentment per Moncayo-Arlandi i Brugada (483). Un altre cas interessant és el de P27, amb la identificació d’una duplicació intragènica en mosaic detectada a *SCN5A* en una pacient de SBr. Els detalls del cas van ser publicats pel nostre grup a Mademont-Solet et al., el 2016 (Annex 4) (386). La CNV va ser classificada com VPP, ja que afectava al gen associat a SBr amb més prevalença, per ser *de novo* i no afectar a la descendència (sent el cas índex l’únic afectat de SBr a la família) i perquè les variants radicals són una causa coneguda de SBr, provocant la pèrdua de funció del canal de sodi cardíac. És un cas interessant, ja que mai abans s’havia reportat una variant causal de SBr en mosaic i, a més, és un cas poc freqüent de “mosaicisme desapareixent” (484). La informació recopilada suggereix que una porció dels casos esporàdics de SBr pot aparèixer a causa de variants patogèniques detectables en el cor però no en altres teixits. També és d’especial interès el cas de P41, un pacient de Mort Sobtada Recuperada –MSR– i història de MSI familiar no investigada, en el que es detecta un genotip complex amb dues delecions independents, la intragènica a *CTNNA3*, un gen estructural, i la deleció del gen *GAA*, relacionat amb el metabolisme del glicogen. En aquest sentit i a la llum d’investigacions recents (485), els protocols contemporanis d’edició genòmica (486,487), juntament amb la utilització de models funcionals més realistes, que tinguin en compte el *background* genètic de l’individu –com els que poden generar-se amb la reprogramació cel·lular (488)–, poden ser de gran ajuda per dur a terme aquests estudis funcionals.

Resumint, en aquesta tesi es presenten els resultats del cribratge de CNVs en la cohort més extensa de pacients de MSC, malalties associades i casos de MSI mai publicada. S’han analitzat les regions codificants d’una àmplia gamma de gens associats, minoritaris i candidats per aquests tipus de pacients. L’anàlisi s’ha dut a terme amb un mètode de detecció de CNVs robust i fiable, dissenyat per l’anàlisi de mostres de seqüenciació d’alt rendiment generades a partir de la captura de regions genòmiques amb rellevància clínica associada. Les variants detectades s’han classificat en funció d’uns criteris propis, inspirats en les recomanacions internacionals. Aquests tenen en consideració algunes reorganitzacions genòmiques que acostumen a discriminar-se (com les duplicacions intragèniques) i que poden aportar més informació de cares a la translació a la pràctica clínica.

Les CNVs són la causa genètica més probable de la malaltia d’una fracció modesta però significativa de la nostra cohort. Hem identificat com a portadors de CNVs un 2.3% dels pacients de la cohort (47 de 2073 pacients). En aquests, un 55.3% de les CNVs detectades han estat classificades com VP o VPP, en coherència amb la diagnosi del pacient. Tan sols en un cas es va detectar una variant puntual classificada com VP o VPP; per la resta, la CNV sembla la causa genètica més probable del fenotip. Un 40’4% de les CNVs s’han classificat com VSI; per aquests casos, els criteris de

classificació utilitzats no han pogut establir una possible relació genotip-fenotip amb el portador. En 4 d'aquests pacients (un 21% del grup) es detecten VP o VPP puntuals que podrien explicar el fenotip. Finalment, en un 4'3% dels portadors de CNVs (2/47) aquesta és classificada com VPB, i en un cas també s'hi identifica una VP puntual com la principal causa de la malaltia. Cal esmentar, però, que una porció de les CNVs classificades com VSI –degut a que el gen no ha estat prèviament associat amb la malaltia– poden ser, de fet, una VP o VPP. És àmpliament conegut que existeix molta variabilitat genètica i fenotípica entre les malalties associades a la MSC i el mateix gen (o inclús la mateixa variant genètica) pot associar-se amb diferents desordres (81,483). L'impacte que pot tenir una CNV en la correcta expressió o funcionalitat del producte d'un gen, independentment del tipus de CNV del que es tracti, fa que aquestes variants siguin rellevants a l'hora d'establir noves relacions genotip-fenotip.

En vistes dels resultats obtinguts, es recomana la inclusió del cribratge rutinari de CNVs en la diagnosi genètica de pacients de MSC o diagnosticats amb malalties associades, així com en casos de MSI de sospita arritmogènica. Amb les eines adequades, l'anàlisi és directe, sense ser necessari cap processament addicional de les mostres, a excepció de la validació de les senyals detectades mitjançant tècniques *gold standard* ràpides i econòmiques, com la qPCR o la MLPA.

El nostre estudi presenta certes limitacions tècniques i metodològiques. Aquestes es discuteixen a continuació.

A l'haver seqüenciat les mostres amb panells de captura, s'ha limitat molt la possibilitat de caracteritzar els punts de trencament de la gran majoria de CNVs detectades. De cares a qualsevol possible estudi funcional i en cas de tractar-se d'una duplicació intragènica, el primer pas hauria de ser la corroboració de que la CNV es troba en tàndem. A banda, a l'analitzar regions acotades en funció de la rellevància clínica coneguda en l'actualitat, és possible que s'estiguin discriminant altres regions on podria localitzar-se la causa genètica d'una fracció dels pacients (sobretot la d'aquells pels que no es té clara –o es descarta– la causalitat de les variants detectades). També és possible que les variants causals es trobin en altres gens no inclosos en les regions de captura, tot i que la diversitat de gens minoritaris i candidats analitzada ha estat generosa. Això pot qüestionar la utilitat dels panells de captura amb finalitats clíniques, però és important analitzar el panorama actual de tecnologies alternatives.

Per una banda l'usuari disposa d'una tecnologia d'*arrays* millorada, més econòmica que la preparació i seqüenciació de llibreries genòmiques (489) i amb una resolució que permet la investigació de CNVs a nivell exònic –sempre i quan es pagui un preu significativament superior en el disseny de sondes– (489,490). La limitació principal dels *arrays* és que no serveixen pel descobriment de noves variants puntuals, ja que els oligonucleòtids sempre són dissenyats en base a SNVs (rars o freqüents) prèviament reportades. Aquesta és una limitació important que relega la tècnica o bé al genotipat de grups molt acotats de pacients, en els que tant sols interessa la detecció de variants

considerades causals i prèviament reportades, o bé al cribratge de grans variants estructurals desequilibrades en pacients pels que aquestes variants siguin una causa genètica típica de la malaltia. Per altra banda no sembla que, en un context clínic, la WES (de l'anglès *Whole Exome Sequencing*) rutinària sigui la millor de les solucions. Fa una dècada que es seqüencien exomes a gran escala, i el coneixement que es té sobre les causes genètiques en regions codificants de gens no associats a la malaltia dels pacients és escàs. Aquest fet reflecteix la necessitat d'invertir més esforços i recursos en estudis funcionals de VSI, per tal d'avançar en el coneixement de les bases moleculars i dels mecanismes fisiopatològics de les malalties. Però, també pot ser un clar indicador de que les causes genètiques han de localitzar-se en regions no codificants, i que no per incloure un número superior de gens en el cribratge estarem més a prop d'identificar la causa genètica. Belkadi et al (491) publiquen evidències de que la WGS (de l'anglès *Whole Genome Sequencing*) és la millor estratègia per la detecció de variants d'exoma, molt més eficaç que la mateixa WES. A l'estudi es conclou que, de mitjana, amb la WGS es detecten 650 SNVs més en regions codificants que amb la WES, a més d'una taxa de detecció de falsos positius molt més reduïda (17% per WGS mentre que per WES és del 78%). Això és degut a que la WES encara és una tecnologia limitada per la captura i els biaixos que l'afecten. La WGS no basa l'estratègia de seqüenciació en sondes de captura, i s'especula que aquest sigui el motiu pel que el preu de la tècnica es reduirà significativament, i en pocs anys, en comparació amb la WES. Al seu torn, Meienberg et al. (492) comparen els resultats de la WES amb els d'un protocol de WGS lliure de PCR, sempre dins d'un context clínic. Les cobertures assolides amb la WGS lliure de PCR, al ser menys sensibles als biaixos típics de les tecnologies de captura i al disseny de sondes, van exhibir una qualitat i una homogeneïtat sense precedents. A diferència dels resultats obtinguts amb WES, es van poder cobrir totes les regions codificants al 100%. Aquest és un requisit indispensable per la seqüenciació amb finalitats clíniques.

Es conclou, doncs, que en un context clínic, el cribratge genètic mitjançant panells de captura de gens associats, minoritaris i candidats és una opció cost-efectiva (308), però temporal. L'optimització dels dissenys de captura i l'ús de *software* sensible i precís fan d'aquesta estratègia l'opció més adequada per aconseguir un cribratge concís i informatiu, que al mateix temps doni marge a la recerca. Tenint en compte la qualitat a la que poden arribar les dades de WGS (492) i la facilitat relativa que plantegen de cares a la identificació de variants estructurals (491), un cop superat l'obstacle econòmic no quedaran motius per no seleccionar aquesta tecnologia com l'estratègia més desitjable. Tot i que es generarà una gran quantitat de dades per pacient i que moltes variants seran classificades com VSI, els criteris diagnòstics són en millora i evolució constant, i les dades sempre poden tornar-se a analitzar.

La classificació de variants s'ha dut a terme en base a uns criteris teòrics. Els estudis funcionals escapen de l'àmbit d'un laboratori de diagnosi genètica, i la inversió de temps i diners dedicats a la realització d'aquests experiments hauria estat molt notable. No obstant, les noves tècniques d'edició genòmica (486,487), juntament amb els protocols de reprogramació cel·lular (488), plantegen un

escenari ideal en el que seran possibles els estudis funcionals a gran escala. En aquests es podrà comprovar l'efecte de variants de difícil estudi fins a dia d'avui, com les localitzades en regions no codificants, i sempre tenint en compte el *background* genètic de l'individu. Aquestes variants podrien estar desregulant l'expressió gènica, ja sigui de manera fina, afectant els promotors, els *enhancers* o els *silencers*, o de manera més generalista, alterant l'organització dels dominis topològics associats a cada gen. També és possible que l'efecte combinat de diverses variants, ja siguin CNVs o variants puntuals i amb independència de la seva classificació, siguin la causa de la malaltia. Tant unes com les altres poden jugar el paper de variant moduladora que, en conjunt, poden provocar la desregulació fisiològica del pacient i el desenvolupament de la malaltia. També es podran explorar altres models de diagnosi genètica, com per exemple la perspectiva multigènica, en la que és l'acumulació de variants de freqüència relativament baixa (i per tant no considerades rares) la que desregula l'homeòstasi de l'individu. Teòricament aquestes aproximacions permetran l'establiment de noves relacions genotip-fenotip que, idealment, donaran resposta a una porció dels casos negatius actuals.

Per acabar, per poder extreure conclusions definitives dels casos és imprescindible disposar de la major quantitat d'informació clínica i familiar possible. Els estudis de cosegregació familiar són una eina determinant a l'hora de traçar una relació causa-efecte entre la variant genètica i el fenotip observat. En la majoria dels casos, l'èxit de la investigació vindrà determinat pel grau de col·laboració que s'estableixi entre un equip multidisciplinari de metges, forenses i investigadors. Per molts dels casos inclosos en aquesta tesi no hi havia informació disponible del pacient a banda de l'estrictament necessària per assegurar la seva diagnosi. D'haver-ne disposat, és possible que moltes de les CNVs detectades s'haguessin acabat classificant de manera diferent, reduint la proporció de VSI. És de vital importància per l'avenç de la diagnosi genètica i pel *feedback* que aquesta pugui traslladar a la clínica la millora de la connectivitat i de la transferència d'informació entre l'hospital i el centre de recerca. Calen iniciatives desinteressades de col·laboració internacional –com les sorgides durant la cursa per la seqüenciació del genoma humà– per fer de la translació a la medicina (i a la societat, en general) una realitat.

VI. Conclusions

1. L'optimització dels dissenys de sondes de captura és un procés determinant a l'hora d'obtenir mostres de qualitat, lliures de biaixos i que permetin una anàlisi genètica exhaustiva en un context clínic.
2. L'algoritme de detecció de CNVs desenvolupat en aquesta tesi demostra ser una eina específica (99.9%) i sensible (100%), comparable i en ocasions superior a *softwares* similars. En termes de precisió (85.9%), el nostre algoritme resulta superior. Aquestes característiques el converteixen en una eina adequada per l'anàlisi de mostres clíniques, amb independència del diagnòstic del pacient.
3. Un 2.3% dels pacients de la cohort de MSC, malalties associades i MSI han estat identificats com portadors de CNVs. Aquesta és una fracció modesta però no negligible de la cohort, ja que en un 55'3% dels portadors la CNV és considerada la causa genètica més probable de la malaltia.
4. L'algoritme detecta CNVs d'un únic exó en 8 pacients. Entre aquests, la variant resulta ser la causa genètica més probable en 6 casos. Aquestes variants estructurals curtes són fàcilment discriminables per altres mètodes de detecció.
5. Les duplicacions intragèniques, generalment discriminades per les guies clíniques d'interpretació de variants, es detecten en 13 pacients (el 27'6% dels portadors de CNVs). Segons els nostres criteris de classificació, la CNV és la causa més probable de la malaltia en 11 dels 13 portadors. Es recomana, doncs, una classificació en coherència amb l'evidència clínica, genètica i familiar, en la que no es subestimi la capacitat disruptiva de cap alteració genètica.
6. En aquesta tesi s'ha identificat un genotip complex sense precedents bibliogràfics en un pacient de MCH. El pacient és portador d'una CNV i d'una variant puntual a *MYBPC3* (VP i VPP, respectivament). S'aconsella el cribratge de CNVs inclús en aquells pacients en els que prèviament s'hagi detectat una variant puntual patogènica, ja que la detecció de genotips complexos pot ajudar al facultatiu a prendre una decisió, sobretot en els casos límit.
7. La detecció de CNVs millora de manera evident la diagnosi genètica en un 55.3% dels portadors de la nostra cohort. Aquesta permet la identificació de familiars portadors asimptomàtics de la variant en possible situació de risc, i el seguiment dels familiars del cas en el que s'hagi detectat una CNV classificada com VSI.
8. La detecció sistemàtica de CNVs contribueix a l'establiment de noves relacions genotip-fenotip i a la resolució de dubtes sobre la causalitat de variants en pacients de MSC, malalties arritmogèniques associades i MSI. No obstant, la realització d'estudis funcionals i de cosegregació familiar són determinants a l'hora d'arribar al fons dels casos clínics i traslladar els resultats de la investigació genètica a la pràctica clínica.

VII. Bibliografia

1. Franklin R, Gosling R. Molecular configuration in sodium thymonucleate. *Nature*. 1953;171:740–1.
2. Watson J, Crick F. Molecular structure of nucleic acids. *Nature*. 1953;Nature(171):737–8.
3. Lehman I, Bessman M, Simms E, Kornberg A. Enzymatic Synthesis of Deoxyribonucleic Acid. *J Biol Chem*. 1958;233(2):163–70.
4. Bessman M, Lehman I, Simms E, Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. II. General properties of the reaction. *J Biol Chem [Internet]*. 1958;233(1):171–7.
5. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*. 1975;94(3):441–8.
6. Sanger F, Air G, Barrell BG, Brown NL, Coulson a R, Fiddes C a, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977;265(5596):687–95.
7. Maxam a M, Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A*. 1977;74(2):560–4.
8. Sanger F, Nicklen S, Coulson a R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A [Internet]*. 1977;74(12):5463–7.
9. Smith L, Fung S, Hunkapiller M, Hunkapiller T, Hood L. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res*. 1985;13(7):2399–412.
10. Smith L, Sanders J, Kaiser R, Hughes P, Dodd C, Connell C, et al. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986;321(6071):674–9.
11. Wada A, Yamamoto M, Soeda E. Automatic DNA sequencer: Computer-programmed microchemical manipulator for the Maxam-Gilbert sequencing method. *Rev Sci Instrum*. 1983;54(11):1569–72.
12. Saik RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science (80-)*. 1988;239:487–91.
13. Weier H, Gray J. A Programmable System to Perform the Polymerase Chain Reaction. *Dna*. 1988;7(6):441–7.
14. Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, et al. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res*. 1990;18(15):4417–21.
15. Huang X, Quesada M, Mathies R. DNA Sequencing Using Capillary Array Electrophoresis. *Anal Chem*. 1992;64:2149–54.
16. Goffeau a, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 Genes Old Questions and New Answers coding either RNA or protein products had The genome of the yeast *Saccharomyces cerevisiae* has been completely sequenced through an international effort involving *Schizosaccharomyces pombe* indicate that t. *Science (80-)*. 1995;274(546–6).
17. Ansorge W. Next-generation DNA sequencing techniques. *N Biotechnol*. 2009;25(4).
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature [Internet]*. 2001;409(6822):860–921.
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science (80-)*. 2001;291(5507):1304–51.
20. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268(1):78–94.
21. Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*. 2001;106(4):413–5.
22. Collins F, Lander E, Rogers J, Waterston R. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;432:931–45.
23. Shendure J, Hanlee J. Next-generation DNA

- sequencing. *Nat Biotechnol.* 2008;26(10):1135–45.
24. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). 2015. p. www.genome.gov/sequencingcostsdata.
 25. Methods N. Method of the Year. *Nat Methods.* 2008;5(1):2008.
 26. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, et al. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 2000;28(20):E87.
 27. Dressman D, Yan H, Traverso G, Kinzler K, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A.* 2003;100(15):8817–22.
 28. Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* [Internet]. 2005;437(7057):376–80.
 29. Ronaghi M, Uhlén M, Nyrén P. A Sequencing Method Based on Real-Time Pyrophosphate. *Science* (80-) [Internet]. 1998;281(5375):363–5.
 30. Hyman ED. A New Method of Sequencing DNA. *Anal Biochem.* 1988;174:423–36.
 31. Metzker M. Sequencing technologies — the next generation. *Nat Rev Genet* [Internet]. 2009;11(1):31–46.
 32. Nyrén, P. Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem.* 1985;509:504–9.
 33. Huse S, Huber J, Morrison H, Sogin M, Welch D. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 2007;8(7):1–9.
 34. Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene.* 1994;148(1):1–6.
 35. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA , a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* 2006;34(3):e22.
 36. Turcatti G, Romieu A, Fedurco M, Tairi A. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis y. *Nucleic Acids Res.* 2008;36(4):e25.
 37. Illumina. History of Illumina Sequencing [Internet]. 2017 [cited 2017 Apr 1].
 38. Bentley D, Balasubramanian S, Swerdlow H, Smith G, Milton J, Brown C, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;456(6):53–9.
 39. Dohm J, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 2008;36(16):e105.
 40. Shendure J, Porreca G, Reppas N, Lin X, John P, Rosenbaum A, et al. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* (80-). 2005;309:1728–32.
 41. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 2008;24(3):133–41.
 42. Li H, Durbin R. Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
 43. Zhou X, Ren L, Li Y, Zhang M, Yu Y, Yu J. The next-generation sequencing technology: A technology review and future perspective. *Sci China Life Sci.* 2010;53(1):44–57.
 44. Hong G. A method for sequencing single-stranded cloned DNA in both directions. *Biosci Rep.* 1981;1:243–52.
 45. Fullwood M, Wei C, Liu E. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome.* 2009;19:521–32.
 46. Chaisson M, Huddleston J, Dennis M,

- Sudmant P, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* [Internet]. 2015;1–11.
47. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* [Internet]. 2016;1–7.
48. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data set of 5 . 4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*. 2017;27:157–64.
49. Schadt EE, Turner S, Kasarskis A, Biosciences P, Road W, Park M. A window into third-generation sequencing. *Hum Mol Genet*. 2010;19(2):227–40.
50. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* (80-). 2009;323:133–8.
51. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biol*. 2013;14:405–9.
52. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of Next-Generation Sequencing Technologies. *Anal Chem*. 2011;83:4327–41.
53. Urban J, Bliss J, Lawrence C, Gerbi S. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. Prepr <http://biorxiv.org/content/early/2015/06/22/019281>. 2015;
54. Simpson J, Workman R, Zuzarte P, David M, Dursi L, Timp W. Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. Prepr <http://biorxiv.org/content/early/2016/04/04/047142>. 2016;
55. Ezkurdia I, Juan D, Rodriguez J, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet*. 2014;1–13.
56. Chi K. The dark side of the human genome. *Nature*. 2016;538:275–7.
57. Maston, G. A. et al. Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet* [Internet]. 2006;7(1):29–59.
58. Tyndall C, Mantia G La, Robert K. the polyoma virus genome between the replication origin and late protein coding sequences is required in cis for both early gene expression and viral DNA replication. *Nucleic acids ...* [Internet]. 1981;9(23):6231–50.
59. Brand AH, Breeden L, Abraham J, Sternglanz R, Nasmyth K. Characterization of a “silencer” in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell*. 1985;41(1):41–8.
60. Udvardy A, Maine E, Schedl P. The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *J Mol Biol*. 1985;185(2):341–58.
61. Li M, Marin-Muller C, Bharadwaj U, Chow KH, Yao Q, Chen C. MicroRNAs: control and loss of control in human physiology and disease. *World J Surg* [Internet]. 2009;33(4):667–684.
62. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*. 2008;134(1):25–36.
63. Stojic L, Niemczyk M, Orjalo A, Ito Y, Ruijter AEM, Uribe-Lewis S, et al. Transcriptional silencing of long noncoding RNA GNG12-AS1 uncouples its transcriptional and product-related functions. *Nat Commun* [Internet]. 2016;7:10406.
64. Patrushev L, Minkevich I. The Problem of the Eukaryotic Genome Size. *Biochemistry*. 2008;73(13):1519–52.
65. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Siew WC, et al. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res*. 2007;17(6):839–51.

66. Xiao-Jie L, Ai-Mei G, Li-Juan J, Jiang X. Pseudogene in cancer: real functions and promising signature. *Med Genet.* 2014;0:1–8.
67. Treangen T, Salzberg S. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev.* 2012;13:36–46.
68. Budworth H, McMurray C. A Brief History of Triplet Repeat Diseases Helen. *Methods Mol Biol.* 2013;1010:3–17.
69. Zhang F, Gu W, Hurles M, Lupski J. Copy Number Variation in Human Health, Disease, and Evolution. *Annu Rev Genomics Hum Genet.* 2009;10:451–81.
70. Lieber M. The mechanism of human nonhomologous DNA End joining. *J Biol Chem.* 2008;283(1):1–5.
71. Han J. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA.* 2010;1(15):1–12.
72. Häslér J, Samuelsson T, Strub K. Useful "junk": Alu RNAs in the human transcriptome. *Cell Mol Life Sci.* 2007;64(14):1793–800.
73. Ponicsan SL, Kugel JF, Goodrich JA. Genomic gems- SINE RNAs regulate mRNA production. 2011;20(2):149–55.
74. Walters RD, Kugel JF, Goodrich JA. InvAluable junk: The cellular impact and function of Alu and B2 RNAs. *IUBMB Life.* 2009;61(8):831–7.
75. Pace J, Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res.* 2007;17:422–32.
76. The_ENCODE_Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.
77. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A [Internet].* 2014;111(17):6131–8.
78. Ran F, Hsu P, Wright J, Agarwala V, Scott D, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc.* 2013;8(11):2281–308.
79. Lee D, Gorkin D, Baker M, Strober B, Asoni A, McCallion A, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet.* 2015;(June):1–9.
80. Fersht AR, Knill-Jones JW. DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine.purine, purine.pyrimidine, and pyrimidine.pyrimidine mismatches during DNA replication. *Proc Natl Acad Sci U S A [Internet].* 1981;78(7):4251–5.
81. Bezzina C, Veldkamp MW, van Den Berg MP, Postma a V, Rook MB, Viersma JW, et al. A single Na(+) channel mutation causing both long-QT and Brugada syndromes. *Circ Res.* 1999;85:1206–13.
82. Tachmazidou I, Dedoussis G, Southam L, Farmaki AE, Ritchie GR, Xifara DK, et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun.* 2013;4:2872.
83. Girirajan S, Campbell C, Eichler E. Human Copy Number Variation and Complex Genetic Disease. *Annu Rev Genet.* 2011;45:203–226.
84. Zhou T, Ko EA, Gu W, Lim I, Bang H, Ko JH. Non-Silent Story on Synonymous Sites in Voltage-Gated Ion Channel Genes. *PLoS One.* 2012;7(10):1–8.
85. Lupski J. Genomic rearrangements and sporadic disease. *Nat Genet [Internet].* 2007;39:S43–7.
86. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature [Internet].* 2015;526(7571):68–74.
87. Carlin J. Mutations Are the Raw Materials of Evolution [Internet]. *Nature Education Knowledge.* 2011 [cited 2017 May 4].
88. Stankiewicz P, Shaw CJ, Withers M, Stankiewicz P, Inoue K, Lupski JR. Serial segmental duplications during primate

- evolution result in complex human genome architecture Serial segmental duplications during primate evolution result in complex human genome architecture. 2004;(713):2209–20.
89. Bailey J a, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* [Internet]. 2006;7(7):552–64.
 90. Conrad D, Andrews T, Carter N, Hurles M, Pritchard J. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* [Internet]. 2006;38(1):75–81.
 91. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature* [Internet]. 2006;444(7118):444–54.
 92. Emerson J, Cardoso-Moreira M, Borevitz J, Long M. Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*. *Science* (80-) [Internet]. 2008;320(5883):1629–31.
 93. Singleton a B, Farrer M, Johnson J, Singleton a, Hague S, Kachergus J, et al. a-Synuclein Locus Triplication Causes Parkinson ' s Disease. *Science* (80-). 2003;302(October):841.
 94. Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerrière A, Vital A, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* [Internet]. 2006;38(1):24–6.
 95. Froyen G, Corbett M, Vandewalle J, Jarvela I, Lawrence O, Meldrum C, et al. Submicroscopic Duplications of the Hydroxysteroid Dehydrogenase HSD17B10 and the E3 Ubiquitin Ligase HUWE1 Are Associated with Mental Retardation. *Am J Hum Genet*. 2008;82(2):432–43.
 96. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese- C, Walsh T, et al. Strong Association of De Novo Copy Number Mutations with Autism. *Science* (80-). 2007;316(5823):445–9.
 97. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos J a, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet*. 2008;40(7):880–5.
 98. Flint J, Hill A V, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, et al. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* [Internet]. 1986;321(6072):744–50.
 99. Higgs DR, Old M, Hunt DM. Negro alpha-Thalassaemia is caused by deletion of a single alpha-globin gene. *Lancet*. 1979;1:272–6.
 100. Gonzalez E. The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. *Science* (80-) [Internet]. 2005;307(5714):1434–40.
 101. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Kamesh L, Heward JM, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet*. 2007;39(6):721–3.
 102. Le Maréchal C, Masson E, Chen J, Morel F, Ruszniewski P, Levy P, et al. Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet* [Internet]. 2006;38(12):1372–4.
 103. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet*. 2006;79(3):439–48.
 104. Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, et al. and primate evolution Gene copy number variation spanning 60 million years of human. *Genome Res* [Internet]. 2007;17(31):1266–77.
 105. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* [Internet]. 2007;39(10):1256–60.
 106. Hastings P, Lupski J, Rosenberg S, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009;10(8):551–64.

107. Lieber M, Ma Y, Pannicke U, Schwarz K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol.* 2003;4(9):712–20.
108. Kobayashi T, Horiuchi T, Tongaonkar P, Vu L, Nomura M. SIR2 regulates recombination between different rDNA repeats, but not recombination within individual rRNA genes in yeast. *Cell.* 2004;117(4):441–53.
109. Unal E, Arbel-Eden A, Sattler U, Shroff R, Lichten M, Haber JE, et al. DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol Cell.* 2004;16(6):991–1002.
110. Kaye J, Melo J, Cheung S, Vaze M, Haber J, Toczyski D. DNA Breaks Promote Genomic Instability by Impeding Proper Chromosome Segregation. *Curr Biol.* 2004;14:2096–106.
111. Soutoglou E, Dorn J, Sengupta K, Jasin M, Nussenzweig A, Ried T, et al. Positional stability of single double-strand breaks in mammalian cells. *Nat Cell Biol.* 2007;9(6):675–82.
112. Oh S, Lao J, Hwang P, Taylor A, Smith G, Hunter N. molecules. *Cell.* 2007;130(2):259–72.
113. Jain S, Sugawara N, Lydeard J, Vaze M, Gac NT Le, Haber JE. A recombination execution checkpoint regulates the choice of homologous recombination pathway during DNA double-strand break repair. *Genes Dev.* 2009;23(3):291–303.
114. Raghavan SC, Kirsch IR, Lieber MR. Analysis of the V(D)J Recombination Efficiency at Lymphoid Chromosomal Translocation Breakpoints. *J Biol Chem.* 2001;276(31):29126–33.
115. Schwarz K, Ma Y, Pannicke U, Lieber MR. Human severe combined immune deficiency and DNA repair. *BioEssays.* 2003;25(11):1061–70.
116. Berg J, Tymoczko J, Stryer L. *Biochemistry.* 7th editio. 2012.
117. Lydeard JR, Jain S, Yamaguchi M, Haber JE. Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature.* 2007;448(7155):820–3.
118. Smith CE, Llorente B, Symington LS. Template switching during break-induced replication. *Nature* [Internet]. 2007;447(7140):102–5.
119. Stankiewicz P, Lupski J. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 2002;18(2):74–82.
120. Lupski JR. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 1998;14(10):417–22.
121. Lupski JR, Stankiewicz P. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* 2005;1(6):0627–33.
122. Turner D, Miretti M, Rajan D, Fiegler H, Carter N, Martyn L, et al. The rates of de novo meiotic deletions and duplications causing several genomic disorders in the male germline. *Nat Genet.* 2008;40(1):90–5.
123. Flores M, Morales L, Gonzaga-Jauregui C, Domínguez R, Zepeda C, Yañez O, et al. Recurrent DNA inversion rearrangements in the human genome. *PNAS.* 2007;104(15).
124. Lam K-WG, Jeffreys AJ. Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proc Natl Acad Sci U S A.* 2006;103(24):8921–7.
125. Carvalho CMB, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, et al. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet.* 2009;18(12):2188–203.
126. Greenawalt DM, Cui X, Wu Y, Lin Y, Wang H, Luo M, et al. structure in a 2 . 5-Mb region on the long arm of chromosome 21 Strong correlation between meiotic crossovers and haplotype structure in a 2 . 5-Mb region on the long arm of chromosome 21. *Genome Res.* 2006;208–14.
127. Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R, Arnheim N. High-resolution recombination patterns in a region of

- human chromosome 21 measured by sperm typing. *PLoS Genet.* 2006;2(5):682–92.
128. Shaw CJ, Lupski JR. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet.* 2004;13 Spec No(1):R57–64.
129. Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am J Hum Genet.* 2006;79(5):890–902.
130. Myers SR, McCarroll S a. New insights into the biological basis of genomic disorders. *Nat Genet.* 2006;38(12):1363–4.
131. Raedt T De, Stephens M, Heyns I, Brems H, Thijs D, Messiaen L, et al. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. *Nat Genet.* 2006;38(12):1419–23.
132. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet [Internet].* 2008;40(9):1124–9.
133. Barbouti A, Stankiewicz P, Nusbaum C, Cuomo C, Cook A, Höglund M, et al. The breakpoint region of the most common isochromosome, i(17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. *Am J Hum Genet [Internet].* 2004;74(1):1–10.
134. Carvalho CMB, Lupski JR. Copy number variation at the breakpoint region of isochromosome 17q. *Genome Res.* 2008;18(11):1724–32.
135. Cuscó I, Corominas R, Bayés M, Flores R, Rivera-brugués N, Campuzano V, et al. susceptibility factor for the Williams-Beuren syndrome deletion Copy number variation at the 7q11 . 23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. *Genome Res.* 2008;683–94.
136. Lupski JR. Genome structural variation and sporadic disease traits. *Nat Genet.* 2006;38(9):974–6.
137. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer M a. L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A [Internet].* 2008;105(49):19366–71.
138. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, et al. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res.* 2003;13(12):2519–32.
139. Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, et al. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet.* 2007;3(10):1939–49.
140. Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet.* 2006;79(1):41–53.
141. Kim PM, Lam HYK, Urban a. E, Korbelt JO, Affourtit J, Grubert F, et al. Analysis of copy number variants and segmental duplication in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* 2008;18:1865–74.
142. Steinmann K, Cooper DN, Kluwe L, Chuzhanova N a, Senger C, Serra E, et al. Type 2 NF1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. *Am J Hum Genet [Internet].* 2007;81(6):1201–20.
143. She X, Cheng Z, Zöllner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. *Nat Genet [Internet].* 2008;40(7):909–14.
144. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature [Internet].* 2008;453(7191):56–64.
145. Korbelt JO, Urban AE, Affourtit JP, Godwin B,

- Grubert F, Simons JF, et al. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* (80-). 2007;318(5849):420–6.
146. Daley JM, Vander Laan RL, Suresh A, Wilson TE. DNA joint dependence of Pol X family polymerase action in nonhomologous end joining. *J Biol Chem*. 2005;280(32):29030–7.
147. McVey M, Lee S. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet*. 2008;24(11):529–38.
148. Tanaka H, Bergstrom D a, Yao M-C, Tapscott SJ. Large DNA palindromes as a common form of structural chromosome aberrations in human cancers. *Hum cell Off J Hum Cell Res Soc*. 2006;19(1):17–23.
149. Slack A, Thornton P, Magner D, Rosenberg S, Hastings P. On the mechanism of gene amplification induced under stress in *Escherichia coli*. *PLoS Genet*. 2006;2(4):385–98.
150. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet*. 2009;5(1).
151. Lieber MR. The mechanism of human nonhomologous DNA End joining. *J Biol Chem*. 2008;283(1):1–5.
152. Ma Y, Pannicke U, Schwarz K, Lieber MR. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell*. 2002;108(6):781–94.
153. Nick McElhinny SA, Ramsden DA. Sibling rivalry: Competition between Pol X family members in V(D)J recombination and general double strand break repair. *Immunol Rev*. 2004;200(D):156–64.
154. Wu P-Y, Frit P, Meesala S, Dauvillier S, Modesti M, Andres SN, et al. Structural and functional interaction between the human DNA repair proteins DNA ligase IV and XRCC4. *Mol Cell Biol* [Internet]. 2009;29(11):3163–72.
155. Lans H, Martejijn J, Vermeulen W. ATP-dependent chromatin remodeling in the DNA-damage response. *Epigenetics Chromatin*. 2012;5(4):1–14.
156. Daley J, Palmbos P, Wu D, Wilson T. Nonhomologous end joining in yeast. *Annu Rev Genet*. 2005;39:431–51.
157. Mimitou LS SE. DNA end resection: many nucleases make light work. *DNA Repair (Amst)*. 2009;8(9):983–95.
158. Toffolatti L, Cardazzo B, Nobile C, Danieli GA, Gualandi F, Muntoni F, et al. Investigating the mechanism of chromosomal deletion: Characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. *Genomics*. 2002;80(5):523–30.
159. Shaw CJ, Lupski JR. Non-recurrent 17p11.2 deletions are generated by homologous and non-homologous mechanisms. *Hum Genet*. 2005;116(1–2):1–7.
160. Stankiewicz P, Shaw CJ, Dapper JD, Wakui K, Shaffer LG, Withers M, et al. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am J Hum Genet* [Internet]. 2003;72(5):1101–16.
161. Espejel S, Franco S, Rodr??guez-Perales S, Bouffler SD, Cigudosa JC, Blasco MA. Mammalian Ku86 mediates chromosomal fusions and apoptosis caused by critically short telomeres. *EMBO J*. 2002;21(9):2207–19.
162. Coquelle A, Pipiras E, Toledo F, Buttin G, Debatisse M. Expression of Fragile Sites Triggers Intrachromosomal Mammalian Gene Amplification and Sets Boundaries to Early Amplicons. *Cell* [Internet]. 1997;89(2):215–25.
163. Arlt MF, Mülle JG, Schaibley VM, Ragland RL, Durkin SG, Warren ST, et al. Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants. *Am J Hum Genet* [Internet]. 2009;84(3):339–50.
164. McClintock B. Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol*. 1951;16:13–47.

165. Tanaka H YM. Palindromic gene amplification — an evolutionarily conserved role for DNA inverted repeats in the genome. *Nat Rev* [Internet]. 2009;9:215–24.
166. Viguera E, Canceill D, Ehrlich SD. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J*. 2001;20(10):2587–95.
167. Petruska J, Hartenstine MJ, Goodman MF. Analysis of Strand Slippage in DNA Polymerase Expansions of CAG / CTG Triplet Repeats Associated with Neurodegenerative Disease Analysis of Strand Slippage in DNA Polymerase Expansions of CAG /. *J Biol Chem*. 1998;273(9):1–8.
168. Jankovic M, Kostic T, Savic D. DNA sequence analysis of spontaneous histidine mutations in a polA1 strain of *Escherichia coli* K12 suggests a specific role of the GTGG sequence. *Mol Gen Genet*. 1990;223:481–6.
169. Cairns J, Fostert P. Adaptive Reversion of a Frameshift Mutation in *Escherichia coli*. *Genetics*. 1991;128:695–701.
170. Lee JA, Carvalho CMB, Lupski JR. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*. 2007;131(7):1235–47.
171. Zhou K, Aertsen A, Michiels C. The Role of Variable DNA Tandem Repeats in Bacterial Adaptation. *FEMS Microbiol Rev*. 2014;38(1):119–41.
172. Goodier JL, Kazazian HH. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. *Cell*. 2008;135(1):23–35.
173. Babushok D, Kazazian H. Progress in Understanding the Biology of the Human Mutagen LINE-1. *Hum Mutat* [Internet]. 2007;28(6):527–39.
174. Beck C, Collier P, Macfarlane C, Malig M, Kidd J, Eichler E, et al. LINE-1 Retrotransposition Activity in Human Genomes. *Cell* [Internet]. 2010;141(7):1159–70.
175. Ostertag EM, Jr. HHK. Biology of Mammalian L1. *Annu Rev Genet* [Internet]. 2001;35:501–38.
176. Cordaux R, Batzer M. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* [Internet]. 2009;10(10):691–703.
177. Konkel M, Batzer M. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* [Internet]. 2010;20(4):211–21.
178. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. Report SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *Am J Hum Genet*. 2003;73:1444–51.
179. Ayarpadikannan S, Kim H. The Impact of Transposable Elements in Genome Implications in Various Diseases. *Genomics Inform*. 2014;12(3):98–104.
180. Chance PF, Alderson MK, Leppig KA, Lensch MW, Matsunami N, Smith B, et al. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell*. 1993;72(1):143–51.
181. Haghghi K, Kolokathis F, Gramolini A, Waggoner J, Pater L, Lynch R, et al. A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. *PNAS*. 2006;103(5):1388–93.
182. Nathans J, Piantida TP, Eddy RL, Shows TB, Hogness DS. Molecular Genetics of Inherited Variation in Human Color Vision.pdf. *Science* (80-). 1986;236:203–11.
183. Lifton R, Dluhy R, Powers M, Rich G, Cook S, Ulick S, et al. A chimaeric 11beta-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. *Nature*. 1992;355:262–5.
184. Velagaleti G, Bien-willner G, Northup J, Lockhart L, Hawkins J, Jalal S, et al. Position Effects Due to Chromosome Breakpoints that Map ~ 900 Kb Upstream and ~ 1.3 Mb Downstream of SOX9 in Two Patients with Campomelic Dysplasia. *Am J Hum Genet*. 2005;76:652–62.
185. Lupiañez D, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of

- Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*. 2015;161:1–14.
186. Kurotaki N, Shen J, Touyama M, Kondoh T, Visser R, Ozaki T, et al. Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. *Genet Med*. 2005;7(7):479–83.
187. Duncan I. Transvection effects in *Drosophila*. *Annu Rev Genet*. 2002;36:521–56.
188. Yan J, Bi W, Lupski J. Penetrance of Craniofacial Anomalies in Mouse Models of Smith- Magenis Syndrome Is Modified by Genomic Sequence Surrounding Rai1: Not All Null Alleles Are Alike. *Am J Hum Genet*. 2007;80:518–25.
189. Alipanahi B, Delong A, Weirauch M, Frey B. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* [Internet]. 2015;33:831–838.
190. Zeng H, Edwards M, Liu G, Gifford D. Convolutional neural network architectures for predicting DNA – protein binding. *Bioinformatics*. 2016;32:i121–7.
191. Hagen JB. The origins of bioinformatics. *Nat Rev Genet*. 2000;1(3):231–6.
192. Ouzounis CA, Valencia A. Early bioinformatics: The birth of a discipline - A personal view. *Bioinformatics*. 2003;19(17):2176–90.
193. West M, Ponnampereuma C. Chemical evolution and the origin of life. *Sp Life Sci*. 1970;2:225–295.
194. Ohno S. *Evolution by Gene Duplication*. New York: Springer-Verlag; 1970.
195. Kabat E, Wu T. The influence of nearest-neighbor amino acid residues on aspects of secondary structure of proteins. Attempts to locate α -helices and β -sheets. *Biopolymers*. 1973;12:751–74.
196. Rossmann M, Liljas A. Recognition of Structural Domains in Globular Proteins. *J Mol Biol*. 1974;85:177–81.
197. Waterman M, Smith T. On the Similarity of Dendrograms. *J Theor Biol*. 1978;73:789–800.
198. Fox G, Stackebrandt E, Hespell R, Gibson J, Maniloff J, Dyer T, et al. The Phylogeny of Prokaryotes. *Science* (80-). 1980;209:457–63.
199. Crick F. Central dogma of molecular biology. *Nature*. 1970;227:561–3.
200. Gibbs A, Dale M, Kinns H, MacKenzie H. The Transition Matrix Method for Comparing Sequences; Its Use in Describing and Classifying Proteins by Their Amino Acid Sequences. *Syst Zool*. 1971;20(4):417–25.
201. Beyer W, Stein M, Smith T, Ulam S. A Molecular Sequence Metric and Evolutionary Trees. *Math Biosci*. 1974;19:9–25.
202. Waterman M, Smith T, Beyer W. Some Biological Sequence Metrics*. *Adv Math* (N Y). 1976;20:367–87.
203. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*. 1980;8(1):49–62.
204. Pain R, Robson B. Analysis of the Code Relating Sequence to Secondary Structure in Proteins. *Nature*. 1970;227:62–3.
205. Ptitsyn O. Statistical Analysis of the Distribution of Amino Acid Residues among Helical and Non-helical Regions in Globular Proteins. *J Mol Biol*. 1969;42:501–10.
206. Lee B, Richards F. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J Mol Biol*. 1971;55:379–400.
207. Dunnill P. The use of helical net-diagrams to represent protein structures. *Biophys J* [Internet]. 1968;8(7):865–75.
208. Smith T, Waterman M. Identification of Common Molecular Subsequences. *J Mol Biol*. 1981;195–7.
209. Smith T, Waterman M. Comparison of Biosequences. *Adv Appl Math*. 1981;2:482–9.
210. Lipman D, Pearson W. Rapid and Sensitive Protein Similarity Searches. *Science* (80-). 1985;227(4693):1435–41.

211. Walker J, Saraste M, Runswick M, Gay N. Distantly related sequences in the α - and β -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J*. 1982;1(8):945–51.
212. Klug A, Rhodes D. "Zinc fingers": a novel protein motif for nucleic acid recognition. *Trends Biochem Sci*. 1987;12:464–9.
213. Fickett J. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*. 1982;10(17):5303–18.
214. Shepherd J. Method to determine the reading frame of a protein from the purine / pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U S A*. 1981;78(3):1596–600.
215. Dumas J, Ninio J. Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Res*. 1981;10(1):197–206.
216. Turner D, Sugimoto N. Rna structure prediction. *Annu Rev Biophys Chem*. 1988;17:167–92.
217. Thornton J. Disulphide Bridges in Globular Proteins. *J Mol Biol*. 1981;151:261–87.
218. Cohen F, Sternberg M, Taylor W. Analysis of the Tertiary Structure of Protein beta-sheet Sandwiches. *J Mol Biol*. 1981;148:253–72.
219. Chothia C, Levitt M, Richardson D. Helix to Helix Packing. *J Mol Biol*. 1981;145:215–50.
220. Rothschild L, Ragan M, Coleman A, Heywood P, Gerbi S. Are rRNA Sequence Comparisons the Rosetta Stone of Phylogenetics? *Cell*. 1986;47:640.
221. Sogin M, Elwood H, Gunderson J. Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc Natl Acad Sci U S A*. 1986;83(5):1383–7.
222. Breslauer K, Frank R, Blocker H, Marky L. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A*. 1986;83:3746–50.
223. Sharp P. On the origin of RNA splicing and introns. *Cell*. 1985;42:397–400.
224. Bulmer M. A Statistical Analysis of Nucleotide Sequences of Introns and Exons in Human Genes. *Mol Biol Evol*. 1987;4(4):395–405.
225. Gilbert W, Marchionni M, Mcknight G. On the Antiquity of Introns. *Cell*. 1986;46:151–4.
226. Doolittle R, Feng D, Johnson M, McClure M. Origins and evolutionary relationships of retroviruses. *Q Rev Biol*. 1989;64(1):1–30.
227. Kelly J, Meyer E. Storage and retrieval of nucleic acid sequence data. *Comput Chem*. 1980;4:107–11.
228. Orcutt B, George D, G, Dayhoff M. Protein and nucleic acid sequence database systems. *Annu Rev Biophys Bioeng*. 1983;12:419–41.
229. Bilofsky H, Burks C. The GenBank® genetic sequence data bank *Nucleic Acids Research*. *Nucleic Acids Res*. 1988;16(5):1861–3.
230. Hamm G, Cameron G. The EMBL data library *Nucleic Acids Research*. *Nucleic Acids Res*. 1986;14(1):5–9.
231. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol [Internet]*. 1990;215(3):403–10.
232. Brunak S, Engelbrecht J, Knudsen S. Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res*. 1990;18(16):4797–802.
233. Guigó R, Knudsen S, Drake N, Smith T. Prediction of Gene Structure. *J Mol Biol*. 1992;226:141–57.
234. Callaway E. Platinum genome shapes up. *Nature*. 2014;515:20.
235. Omics Project List [Internet]. [cited 2017 May 16].
236. Romanoski C, Glass C. Epigenomics: Roadmap for regulation. *Nature*. 2015;518:314–6.
237. Relling M, Evans W. Pharmacogenomics in the clinic. *Nature*. 2015;526:343–50.
238. Neeha V, Kinth P. Nutrigenomics research : a review. *J Food Sci Technol*. 2013;50(3):415–

- 428.
239. OmicX Group. Adapter removal software tools [Internet]. [cited 2017 May 16].
240. Conesa A, Madrigal P, Tarazona S, Gomez-cabrero D, Cervera A, Mcpherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17(13):1–19.
241. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Biol.* 2013;9(11):1–9.
242. Parnell L, Lindenbaum P, Shameer K, Marco G, Swan D, Jensen L, et al. BioStar: An Online Question & Answer Resource for the Bioinformatics Community. *PLoS Comput Biol.* 2011;7(10):8–12.
243. Olivares E. Seqanswers.com [Internet]. [cited 2017 May 17].
244. Lejeune J, Marthe G, Raymond T. Étude des chromosomes somatiques des neuf enfants mongoliens. *C R Acad Sci.* 1959;248:1721–2.
245. Ford C, Jones K, Polani P, Almeida J, Briggs J. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet.* 1959;1(7075):711–3.
246. Jacobs P, Strong J. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature.* 1959;183:302–3.
247. Patau K, Smith D, Therman E, Inhorn S, Wagner HP. Multiple congenital anomaly caused by an extra autosome. *Lancet.* 1960;1:790–3.
248. Edwards J. A new trisomic syndrome. *Lancet.* 1960;1:787–90.
249. Lejeune J, Lafourcade J, Berger R. 3 cases of partial deletion of the short arm of a 5 chromosome. *C R Acad Sci.* 1963;257(5):3098–102.
250. Patil S, Merrick S, Lubs H. Identification of Each Human Chromosome with a Modified Giemsa Stain. *Science (80-).* 1971;173:821–2.
251. Drets M, Shaw M. Specific Banding Patterns of Human Chromosomes Specific Banding Patterns of Human Chromosomes. *Proc Natl Acad Sci.* 1971;68(9):2073–7.
252. Hook E. Exclusion of Chromosomal Mosaicism: Tables of 90 %, 95 %, and 99 % Confidence Limits and Comments on Use. *Am J Hum Genet.* 1977;29:94–7.
253. Vermeesch J, Brady P, Sanlaville D, Kok K, Hastings R. Genome-Wide Arrays: Quality Criteria and Platforms to be Used in Routine Diagnostics. *Hum Mutat.* 2012;33(6):906–15.
254. Pinkel D, Gray JW, Trask B. Cytogenetic Analysis by In Situ Hybridization with Fluorescently Labeled Nucleic Acid Probes Cytogenetic Analysis by In Situ Hybridization with Fluorescently Labeled Nucleic Acid Probes. *Cold Spring Harb Symp Quant Biol.* 1986;51:151–7.
255. Koboldt D. Exome-based Copy Number Analysis with VarScan 2 [Internet]. [cited 2017 Jun 2].
256. Shapiro M. The Beginner's Guide to Genetics Hacking [Internet]. [cited 2017 Jun 2].
257. Kallioniemi A, Kallioniemi O, Sudar D, Rutovitz D, Gray J, Waldman F, et al. Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors. *Science (80-).* 2016;258(5083):818–21.
258. Speicher R, Ballard S, Ward D. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat Genet.* 1996;12:368–75.
259. Liyanage M, Coleman A, Manoir S, Veldman T, McCormak S, Dickson R, et al. Multicolour spectral karyotyping of mouse chromosomes. *Nat Genet.* 1996;14:312–5.
260. Higuchi R, Fockler C, Dollinger G, Watson R. Kinetic PCR analysis: Real-time monitoring of DNA amplification reactions. *Bio/Technology.* 1993;11:1026–30.
261. Schouten J, Mcelgunn C, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* 2002;30(12):e57.
262. VanOpstal D, Boter M, DeJong D,

- VanDenBerg C, Brüggewirth H, Wildschut H, et al. Rapid aneuploidy detection with multiplex ligation-dependent probe amplification: a prospective study of 4000 amniotic fluid samples. *Eur J Hum Genet*. 2009;17:112–21.
263. Armengol L, Nevado J, Serra-juhe C, Plaja A, Mediano C, García-Santiago A, et al. Clinical utility of chromosomal microarray analysis in invasive prenatal diagnosis. *Hum Genet*. 2012;131:513–23.
264. Strassberg M, Fruhman G, Van den Veyer I. Copy-number changes in prenatal diagnosis. *Expert Rev*. 2011;11(6):579–92.
265. Solinas-toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, et al. Matrix-Based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. *Genes Chromosomes Cancer*. 1997;20:399–407.
266. Shaffer L, Dabell M, Fisher A, Coppinger J, Bandholz A, Ellison J, et al. Experience with microarray-based comparative genomic hybridization for prenatal diagnosis in over 5000 pregnancies. *Prenat Diagn*. 2012;32:1–10.
267. Lee C, Scherer S. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med*. 2010;12(March):1–29.
268. Boone P, Bacino C, Shaw C, Eng P, Hixson P, Pursley A, et al. Detection of Clinically Relevant Exonic Copy-Number Changes by Array CGH. *Hum Mutat*. 2015;31:1326–1342.
269. Miller D, Adam M, Aradhya S, Biasecker L, Brothman A, Carter N, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am J Hum Genet* [Internet]. 2010;86(5):749–64.
270. Breman A, Pursley A, Hixson P, Bi W, Ward P, Bacino C, et al. Prenatal chromosomal microarray analysis in a diagnostic laboratory; experience with >1000 cases and review of the literature. *Prenat Diagn*. 2012;32:351–61.
271. Srebniak M, Boter M, Oudesluijs G, Cohen-overbeek T, Govaerts L, Diderich K, et al. Genomic SNP array as a gold standard for prenatal diagnosis of foetal ultrasound abnormalities. *Mol Cytogenet*. 2012;5(14):5–8.
272. Retterer K, Scuffins J, Schmidt D, Lewis R, Pineda-alvarez D, Stafford A, et al. Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet Med*. 2014;17(8):623–9.
273. Aradhya S, Lewis R, Bonaga T, Nwokekeh N, Stafford A, Boggs B, et al. Exon-level array CGH in a large clinical cohort demonstrates increased sensitivity of diagnostic testing for Mendelian disorders. *Genet Med*. 2012;14(6):594–603.
274. Krumm N, Sudmant P, Ko A, Roak B, Malig M, Coe B, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res*. 2012;22:1525–32.
275. Sathirapongsasuti J, Lee H, Horst B, Brunner G, Coch- A, Binder S, et al. Exome Sequencing-Based Copy-Number Variation and Loss of Heterozygosity Detection: ExomeCNV. *Bioinformatics*. 2011;27(19):2648–2654.
276. Xie C, Tammi M. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*. 2009;10:1–9.
277. Zeitouni B, Boeva V, Janoueix-lerosey I, Loeillet S, Legoix-né P, Nicolas A, et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*. 2010;26(15):1895–6.
278. Karakoc E, Alkan C, Roak B, Dennis M, Vives L, Mark K, et al. Detection of structural variants and indels within exome data. *Nat Methods*. 2012;9(2):176–8.
279. Ye K, Schulz M, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(1):2865–71.
280. Li J, Lupat R, Amarasinghe K, Thompson E, Doyle M, Ryland G, et al. CONTRA: copy

- number analysis for targeted resequencing. *Bioinformatics*. 2012;28(10):1307–13.
281. Plagnol V, Curtis J, Epstein M, Mok K, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;28(21):2747–54.
282. Wu J, Grzeda K, Stewart C, Grubert F, Urban A, Snyder M, et al. Copy Number Variation detection from 1000 Genomes project exon capture sequencing data. *BMC Bioinformatics*. 2012;13(1):1–19.
283. Tattini L, Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*. 2015;3(June):1–8.
284. Zhang Z, Du J, Lam H, Abyzov A, Urban A, Snyder M, et al. Identification of genomic indels and structural variations using split reads. *BMC Bioinformatics*. 2011;12(375):1–12.
285. Oliva A, Brugada R, D'Aloja E, Boschi I, Partemi S, Brugada J, et al. State of the Art in Forensic Investigation of Sudden Cardiac Death. *Am J Forensic Med Pathol*. 2011;32(1):1–16.
286. Goldberger JJ, Cain ME, Hohnloser SH, Kadish AH, Knight BP, Lauer MS, et al. AHA / ACCF / HRS Scientific Statement American Heart Association / American College of Cardiology Foundation / Heart Rhythm Society Scientific Statement on Noninvasive Risk Stratification Techniques for Identifying Patients at Risk for Sudden Cardiac Death. *Heart Rhythm*. 2008;118:1497–518.
287. Brugada R, Brugada J, Brugada P. Clinical approach to sudden cardiac death syndromes. *JAMA*. 2010;304(15):1724–5.
288. Ackerman MJ, Priori SG, Willems S, Berul C, Brugada R, Calkins H, et al. HRS / EHRA EXPERT CONSENSUS STATEMENT HRS / EHRA Expert Consensus Statement on the State of Genetic Testing for the Channelopathies and Cardiomyopathies. *Europace*. 2011;13:1077–109.
289. Fishman GI, Chugh SS, DiMarco JP, Albert CM, Anderson ME, Bonow RO, et al. Sudden cardiac death prediction and prevention: report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop. *Circulation*. 2010;122:2335–48.
290. Mendis S, Puska P, Norrving B. Global Atlas on cardiovascular disease prevention and control. NBWHOWHFWWSOG. 2011;
291. Zipes DP, Camm AJ, Borggrefe M, Moss AJ, Buxton AE, Myerburg RJ, et al. ACC / AHA / ESC PRACTICE GUIDELINES ACC / AHA / ESC 2006 Guidelines for Management of Patients With Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death A Report of the American College of Cardiology / American Heart Association Task Force a. *Circulation*. 2006;114:e385–484.
292. Chugh SS, Reinier K, Teodorescu C, Evanado A, Kehr E, Samara M Al, et al. Epidemiology of Sudden Cardiac Death: Clinical and Research Implications. *Prog Cardiovasc Dis [Internet]*. 2008;51(3).
293. Deo R, Albert CM. Epidemiology and Genetics of Sudden Cardiac Death. *Circulation*. 2012;620–37.
294. Wong LCH, Behr ER. Sudden unexplained death in infants and children: the role of undiagnosed inherited cardiac conditions. *Europace*. 2014;1–8.
295. Sarquella-Brugada G, Campuzano O, Iglesias A, Sánchez-Malagón J, Guerra-Balic M, Brugada J, et al. Genetics of sudden cardiac death in children and young athletes. *Cardiol Young*. 2013;23:159–73.
296. Moon RY, Horne RSC, Hauck FR. Sudden infant death syndrome. *Lancet*. 2007;370:1578–87.
297. Tfelt-hansen J, Winkel G. Cardiac Channelopathies and Sudden Infant Death Syndrome. *Cardiology*. 2011;119:21–33.
298. Campuzano O, Allegue C, Sarquella-Brugada G, Coll M, Mates J, Alcalde M, et al. The role of clinical, genetic and segregation evaluation in sudden infant death. *Forensic Sci Int*. 2014;242:9–15.
299. Warland J, Mitchell E. A triple risk model for unexplained late stillbirth. *BMC Pregnancy Childbirth [Internet]*. 2014;14:142–7. Available from: BMC Pregnancy and Childbirth

300. Lodder EM, Bezzina CR. Arrhythmogenic Right Ventricular Cardiomyopathy : Growing Evidence for Complex Inheritance. *Circ Cardiovasc Genet*. 2013;6:525–8.
301. Gourraud J-B, Barc J, Thollet A, Le Scouarnec S, Le Marec H, Schott J-J, et al. The Brugada Syndrome : A Rare Arrhythmia Disorder with Complex inheritance. *Front Cardiovasc Med*. 2016;3(April):1–11.
302. Kapplinger JD, Tester DJ, Salisbury BA, Carr JL, Harris-kerr C, Pollevick GD, et al. Spectrum and prevalence of mutations from the first 2 , 500 consecutive unrelated patients referred for the FAMILION ® long QT syndrome genetic test. *Hear Rhythm [Internet]*. 2009;6(9):1297–303.
303. Chen Q, Kirsch GE, Zhang D, Brugada R, Brugada J, Brugada P, et al. Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Lett to Nat*. 1998;392:293–6.
304. Mann SA, Castro ML, Ohanian M, Guo G, Zodgekar P, Sheu A, et al. R222Q SCN5A Mutation Is Associated With Reversible Ventricular Ectopy and Dilated Cardiomyopathy. *J Am Coll Cardiol [Internet]*. 2012;60(16):1566–73.
305. Yu J, Hu J, Dai X, Cao Q, Xiong Q, Liu X, et al. SCN5A mutation in Chinese patients with arrhythmogenic right ventricular dysplasia. *Herz*. 2014;39:271–5.
306. Havndrup O, Christiansen M, Stoevring B, Jensen M, Hoffman-bang J, Andersen PS, et al. Fabry disease mimicking hypertrophic cardiomyopathy : genetic screening needed for establishing the diagnosis in women. *Eur J Heart Fail*. 2010;12:535–40.
307. Cobo-Marcos M, Cuenca S, Gámez Martínez JM, Bornstein B, Ripoll Vera T, Garcia-Pavia P. Usefulness of Genetic Testing for Hypertrophic Cardiomyopathy in Real-world Practice. *Rev Esp Cardiol*. 2013;66(9):746–7.
308. Alfares AA, Kelly MA, McDermott G, Funke BH, Lebo MS, Baxter SB, et al. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genet Med*. 2014;
309. Ingles J, Sarina T, Yeates L, Hunt L, Macciocca I, McCormack L, et al. Clinical predictors of genetic testing outcomes in hypertrophic cardiomyopathy. *Genet Med*. 2013;15(12):972–7.
310. Simpson S, Edwards J, Ferguson-mignan TFN, Cobb M, Mongan NP, Rutland CS. Genetics of Human and Canine Dilated Cardiomyopathy. *Int J Genomics*. 2015;1–13.
311. Lazzarini E, Jongbloed JDH, Pilichou K, Thiene G, Bikker H, Charbon B, et al. The ARVD / C Genetic Variants Database : 2014 Update. *Hum Mutat*. 2014;36(4):403–10.
312. Carrilho-ferreira P, Almeida AG. Non-compaction Cardiomyopathy : Prevalence , Prognosis , Pathoetiology , Genetics , and Risk of Cardioembolism. *Curr Hear Fail Rep*. 2014;11:393–403.
313. Campuzano O, Brugada R, Iglesias A. Genetics of Brugada syndrome. *Curr Opin Cardiol*. 2010;25:210–5.
314. Mizusawa Y, Horie M, Characteristics C. Genetic and Clinical Advances in Congenital Long QT Syndrome. *Circulation*. 2014;78:2827–33.
315. Mazzanti A, Kanthan A, Monteforte N, Memmi M, Bloise R, Novelli V, et al. Novel Insight Into the Natural History of Short QT Syndrome. *J Am Coll Cardiol*. 2014;63(13):1300–8.
316. Priori SG, Napolitano C, Memmi M, Colombi B, Drago F, Gasparini M, et al. Clinical and Molecular Characterization of Patients With Catecholaminergic Polymorphic Ventricular Tachycardia. *Circulation*. 2002;106:69–74.
317. Duijvenboden K Van, Ruijter JM, Christoffels VM. Gene regulatory elements of the cardiac conduction system. *Brief Funct Genomics*. 2013;13(1):28–38.
318. Bokil NJ, Baisden JM, Radford DJ, Summers KM. Molecular genetics of long QT syndrome. *Mol Genet Metab [Internet]*. 2010;101(1):1–8.
319. Zarrei M, MacDonald J, Merico D, Scherer S. A copy number variation map of the human genome. *Nat Publ Gr [Internet]*. 2015;16(3):172–83.

320. Jacoby D, McKenna WJ. Genetics of inherited cardiomyopathy. *Eur Heart J*. 2012;33:296–304.
321. Thiene G, Basso C. Arrhythmogenic right ventricular cardiomyopathy: An update. *Cardiovasc Pathol*. 2001;10:109–17.
322. Corrado D, Basso C, Rizzoli G, Schiavon M, Thiene G. Does Sports Activity Enhance the Risk of Sudden Death in Adolescents and Young Adults? *J Am Coll Cardiol*. 2003;42(11).
323. Maron B, Gardin J, Flack J, Gidding S, Kurosaki T, Bild D. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation*. 1995;92(4):785–9.
324. Van Driest SL, Ommen SR, Tajik AJ, Gersh BJ, Ackerman MJ. Sarcomeric Genotyping in Hypertrophic Cardiomyopathy. *Mayo Clin Proc*. 2005;80(April):463–9.
325. Lopes LR, Syrris P, Guttman OP, Mahony CO, Tang HC, Dalageorgou C, et al. Novel genotype – phenotype associations demonstrated by high-throughput sequencing in patients with hypertrophic cardiomyopathy. *Heart*. 2014;101(4):294–301.
326. Richard P, Charron P, Carrier L. Hypertrophic Cardiomyopathy: Distribution of Disease Genes , Spectrum of Mutations , and Implications for a Molecular Diagnosis Strategy A Novel Missense Mutation in the Myosin Binding Protein-C Gene Is Responsible for Hypertrophic Cardiomyopathy With Lef. *Circulation*. 2003;107(17):2227–32.
327. Veselka J, Anavekar NS, Charron P. Hypertrophic obstructive cardiomyopathy. *Lancet [Internet]*. 2016;6736(16):1–15.
328. Marian A, Yu Q-T, Mares A, Hill R, Roberts R, Perryman B. Detection of a New Mutation in the beta-Myosin Heavy Chain Gene in an Individual with Hypertrophic Cardiomyopathy. *J Clin Invest*. 1992;90:2156–65.
329. Jouven X, Hagege A, Charron P, Carrier L, Dubourg O, Langlard J, et al. Relation between QT duration and maximal wall thickness in familial hypertrophic cardiomyopathy. *Heart*. 2002;88:153–7.
330. Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, et al. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Publ Gr [Internet]*. 2009;6(7):507–10.
331. Coto E, Reguero JR, Palacín M, Gómez J, Alonso B, Iglesias S, et al. Resequencing the Whole MYH7 Gene (Including the Intronic , Promoter , and 3' = UTR Sequences) in Hypertrophic Cardiomyopathy. *JMDI [Internet]*. 2012;14(5):518–24.
332. Bagnall RD, Yeates L, Semsarian C. The role of large gene deletions and duplications in MYBPC3 and TNNT2 in patients with hypertrophic cardiomyopathy. *Int J Cardiol [Internet]*. 2009;145(1):150–3.
333. Chanavat V, Seronde MF, Bouvagnet P, Chevalier P, Rousson R, Millat G. Molecular characterization of a large MYBPC3 rearrangement in a cohort of 100 unrelated patients with hypertrophic cardiomyopathy. *Eur J Med Genet [Internet]*. 2012;55(3):163–6.
334. Pezzoli L, Elena M, Ferrazzi P, Iacone M. A new mutational mechanism for hypertrophic cardiomyopathy. *Gene [Internet]*. 2012;507(2):165–9.
335. Herman D, Lam L, Taylor M, Wang L, Teekakirikul P, Christodoulou D, et al. Truncations of Titin Causing Dilated Cardiomyopathy. *N Engl J Med*. 2012;366:619–28.
336. Truszkowska T, Bili ZT, Kosi J, Justyna Ś, Franaszczuk M, Bili M, et al. A study in Polish patients with cardiomyopathy emphasizes pathogenicity of phospholamban (PLN) mutations at amino acid position 9 and low penetrance of heterozygous null PLN mutations. *BMC Med Genet*. 2015;16(21):1–9.
337. Lopes LR, Murphy C, Syrris P, Dalageorgou C, McKenna WJ, Elliott PM, et al. Use of High-throughput Targeted Exome-sequencing to screen for Copy Number Variation in Hypertrophic Cardiomyopathy.

- Eur J Med Genet [Internet]. 2015;58(11):611–6.
338. Ceyhan-birsoy O, Pugh TJ, Bowser MJ, Hynes E, Frisella AL, Mahanta LM, et al. Next generation sequencing-based copy number analysis reveals low prevalence of deletions and duplications in 46 genes associated with genetic cardiomyopathies. *Mol Genet Genomic Med*. 2016;4(2):143–151.
339. Raju H, Alberg C, Sagoo GS, Burton H, Behr ER. Inherited cardiomyopathies. *BMJ*. 2011;1–7.
340. Guzzo-Merello G, Cobo-Marcos M, Gallego-Delgado M, Garcia-Pavia P. Alcoholic cardiomyopathy. *World J Cardiol*. 2014;6(8):771–81.
341. Lakdawala NK, Winterfield JR, Funke BH, Medic H, Universit L. Dilated Cardiomyopathy. *Circ Arrhythm Electrophysiol*. 2012;
342. Judge DP, Johnson NM. Genetic Evaluation of Familial Cardiomyopathy. *J Cardiovasc Trans Res*. 2008;1:144–54.
343. Jefferies JL, Towbin JA. Dilated cardiomyopathy. *Lancet*. 2010;375:752–62.
344. Hershberger RE, Morales A, Siegfried JD. Clinical and genetic issues in dilated cardiomyopathy: A review for genetics professionals. *Genet Med*. 2010;12(11).
345. Gupta P, Bilinska ZT, Sylvius N, Boudreau E, Veinot JP, Labib S, et al. Genetic and ultrastructural studies in dilated cardiomyopathy patients: a large deletion in the lamin A / C gene is associated with cardiomyocyte nuclear envelope disruption. *Basic Res Cardiol*. 2010;105:365–77.
346. Norton N, Siegfried J, Li D, Hershberger R. Assessment of LMNA Copy Number Variation in 58 Proband with Dilated Cardiomyopathy. *Clin Transl Sci*. 2011;4(5):351–2.
347. Sen-chowdhry S, Syrris P, Ward D, Asimaki A, Sevdalis E, McKenna WJ. Clinical and Genetic Characterization of Families With Arrhythmogenic Right Ventricular Dysplasia / Cardiomyopathy Provides Novel Insights Into Patterns of Disease Expression. *Circulation*. 2007;115:1710–21.
348. Marcus FI, Zareba W, Calkins H, Towbin JA, Basso C, Bluemke DA, et al. Arrhythmogenic right ventricular cardiomyopathy / dysplasia clinical presentation and diagnostic evaluation: Results from the North American Multidisciplinary Study. *Hear Rhythm [Internet]*. 2009;6(7):984–92.
349. Rossi PA. Arrhythmogenic right ventricular dysplasia — clinical features. *Eur Heart J*. 1989;10:7–9.
350. Demellawy D El, Nasr A, Alowami S. An Updated Review on the Clinicopathologic Aspects of. *Am J Forensic Med Pathol*. 2009;30(1):78–83.
351. Marcus FI, Co-chair WJM, Sherrill D, Basso C, Bauce B, Bluemke DA, et al. Diagnosis of arrhythmogenic right ventricular cardiomyopathy / dysplasia Proposed Modification of the Task Force Criteria. *Eur Heart J*. 2010;31:806–14.
352. Asimaki A, Tandri H, Huang H, Halushka MK, Gautam S, Basso C, et al. A New Diagnostic Test for Arrhythmogenic Right Ventricular Cardiomyopathy. *N Engl J Med*. 2009;360:1075–84.
353. Corrado D, Fontaine G, Marcus FI, McKenna WJ, Nava A, Thiene G, et al. Arrhythmogenic Right Ventricular Dysplasia/Cardiomyopathy Need for an International Registry. *Circulation*. 2000;101:e101–6.
354. Sen-Chowdhry S, Syrris P, McKenna W. Genetics of Right Ventricular Cardiomyopathy. *J Cardiovasc Electrophysiol*. 2005;16:927–35.
355. Dalal D, James C, Devanagondi R, Tichnell C, Tucker A, Prakasa K, et al. Penetrance of Mutations in Plakophilin-2 Among Families With Arrhythmogenic Right Ventricular Dysplasia / Cardiomyopathy. *Jour*. 2006;48(7).
356. Awad MM, Calkins H, Judge DP. Mechanisms of Disease: molecular genetics of arrhythmogenic right ventricular dysplasia / cardiomyopathy. *Nat Clin Pract*. 2008;5(5):258–67.
357. McKoy G, Protonotarios N, Crosby A, Tsatsopoulou A, Anastasakis A, Coonar A, et al. Identification of a deletion in plakoglobin

- in arrhythmogenic right ventricular cardiomyopathy with palmoplantar keratoderma and woolly hair (Naxos disease). *Lancet*. 2000;355:2119–24.
358. Norgett EE, Hatsell SJ, Carvajal-huerta L, Cabezas JR, Common J, Purkis PE, et al. Recessive mutation in desmoplakin disrupts desmoplakin – intermediate filament interactions and causes dilated cardiomyopathy , woolly hair and keratoderma. *Hum Mol Genet*. 2000;9(18):2761–6.
359. Xu T, Yang Z, Vatta M, Rampazzo A, Beffagna G, Pillichou K, et al. Compound and Digenic Heterozygosity Contributes to Arrhythmogenic Right Ventricular Cardiomyopathy. *J Am Coll Cardiol* [Internet]. 2010;55(6):587–97.
360. Cox M, van der Zwaag P, van der Werf C, van der Smagt J, Noorman M, Bhuiyan Z, et al. Arrhythmogenic Right Ventricular Dysplasia / Cardiomyopathy Pathogenic Desmosome Mutations in Index-Patients Predict Outcome of Family Screening : Dutch Arrhythmogenic Right Ventricular. *Circulation*. 2011;123:2690–2700.
361. Roberts J, Herkert J, Rutberg J, Nikkel S, Wiesfeld A, Dooijes D. Detection of genomic deletions of PKP2 in arrhythmogenic right ventricular cardiomyopathy. *Clin Genet*. 2012;83(5):452–6.
362. Sonoda K, Ohno S, Otuki S, Kato K, Yagihara N, Watanabe H, et al. Quantitative analysis of PKP2 and neighbouring genes in a patient with arrhythmogenic right ventricular cardiomyopathy caused by heterozygous PKP2 deletion. *Europace*. 2016;euw038.
363. Engberding R, Yelbuz TM, Breithardt G. Isolated noncompaction of the left ventricular myocardium. *Clin Res Cardiol*. 2007;96:481–8.
364. Maron BJ, Towbin JA, Thiene G, Antzelevitch C, Corrado D, Arnett D, et al. Contemporary Definitions and Classification of the Cardiomyopathies An American Heart Association Scientific Statement From the Council on Clinical Cardiology , Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functio. *Circulation*. 2006;113:1807–16.
365. Pignatelli RH, McMahon CJ, Dreyer WJ, Denfield SW, Price J, Belmont JW, et al. Clinical Characterization of Left Ventricular Noncompaction in Children A Relatively Common Form of Cardiomyopathy. *Circulation*. 2003;108:2672–9.
366. Scaglia F, Towbin JA, Craigen WJ, Belmont JW, Smith EOB, Neish SR, et al. Clinical Spectrum, Morbidity, and Mortality in 113 Pediatric Patients With Mitochondrial Disease. *Pediatrics*. 2004;114(4):925–31.
367. Klaassen S, Probst S, Oechslin E, Gerull B, Krings G, Schuler P, et al. Mutations in Sarcomere Protein Genes in Left Ventricular Noncompaction. *Circulation*. 2008;117:2893–901.
368. Hoedemaekers YM, Caliskan K, Michels M, Frohn-mulder I, Jasper J, Smagt V Der, et al. The Importance of Genetic Counseling , DNA Diagnostics , and Cardiologic Family Screening in Left Ventricular. *Circ Cardiovasc Genet*. 2010;3:232–9.
369. Bleyl SB, Mumford BR, Thompson V, Carey JC, Pysker TJ, Chin TK, et al. Neonatal , Lethal Noncompaction of the Left Ventricular Myocardium Is Allelic with Barth Syndrome. *Am J Hum Genet*. 1997;61:868–72.
370. Tang S, Batra A, Zhang Y, Ebenroth ES, Huang T. Left ventricular noncompaction is associated with mutations in the mitochondrial genome. *Mitochondrion* [Internet]. 2010;10(4):350–7.
371. Campbell MJ, Czosek RJ, Hinton RB, Miller EM. Exon 3 Deletion of Ryanodine Receptor Causes Left Ventricular Noncompaction , Worsening Catecholaminergic Polymorphic Ventricular Tachycardia , and Sudden Cardiac Arrest. *Am J Med Genet*. 2015;167A:2197–200.
372. Ohno S, Omura M, Kawamura M, Kimura H, Itoh H, Makiyama T, et al. Exon 3 deletion of RYR2 encoding cardiac ryanodine receptor is associated with left ventricular non-compaction. *Europace*. 2014;16:1646–54.
373. Kushwaha S, Fallon J, Fuster V. Restrictive

- Cardiomyopathy. *N Engl J Med*. 1997;336(4):267–76.
374. Stöllberger C, Finsterer J. Extracardiac Medical and Neuromuscular Implications in Restrictive Cardiomyopathy. *Clin Cardiol*. 2007;30:375–80.
375. American Heart Association [Internet]. 2017 [cited 2017 Apr 21].
376. Abriel H, Zaklyazminskaya E V. Cardiac channelopathies: Genetic and molecular mechanisms. *Gene* [Internet]. 2013;517(1):1–11.
377. Antzelevitch C, Brugada P, Brugada J, Brugada R, Shimizu W, Gussak I, et al. Brugada Syndrome: A Decade of Progress. *Circ Res*. 2002;91:1114–9.
378. Brugada J, Brugada R, Antzelevitch C, Towbin J, Nademanee K, Brugada P. Long-Term Follow-Up of Individuals With the Electrocardiographic Pattern of Right Bundle-Branch Block and ST-Segment Elevation in Precordial Leads V₁ to V₃. *Circulation*. 2002;105:73–8.
379. Berne P, Brugada J. Brugada Syndrome 2012. *Circulation*. 2012;76:1563–71.
380. Brugada R, Brugada J, Antzelevitch C, Kirsch GE, Potenza D, Towbin JA, et al. Sodium Channel Blockers Identify Risk for Sudden Death in Patients With ST-Segment Elevation and Right Bundle Branch Block but Structurally Normal Hearts. *Circulation*. 2000;101:510–5.
381. Kapplinger JD, Tester DJ, Alders M, Benito B, Berthet M, Brugada J, et al. An international compendium of mutations in the SCN5A - encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. *Heart Rhythm* [Internet]. 2010;7(1):33–46.
382. Probst V, Hoorntje TM, Hulsbeek M, Wilde AAM, Alshinawi C, Kyndt F. Cardiac conduction defects associate with mutations in SCN5A. *Nat Genet*. 1999;23(september):20–1.
383. Remme CA, Verkerk AO, Nuyens D, Ginneken ACG Van, Brunschot S Van, Belterman CNW, et al. Overlap Syndrome of Cardiac Sodium Channel Disease in Mice Carrying the Equivalent Mutation of Human. *Circulation*. 2006;114:2584–95.
384. Nielsen MW, Holst AG, Olesen S, Olesen MS. The genetic component of Brugada syndrome. *Front Physiol*. 2013;4(July):1–12.
385. Eastaugh L, James P, Phelan D, Hons B, Davis A. Brugada Syndrome Caused by a Large Deletion in SCN5A Only Detected by Multiplex Ligation-Dependent Probe Amplification. *J Cardiovasc Electrophysiology*. 2011;22(9):1073–6.
386. Mademont-soler I, Mates J, Pinsach-abuin M, Riuro H, Coll M, Porres M, et al. Large Genomic Imbalances in Brugada Syndrome. *PLoS One*. 2016;11(9):e0163514.
387. Schwartz PJ, Stramba-badiale M, Crotti L, Pedrazzini M, Besana A, Bosi G, et al. Prevalence of the Congenital Long-QT Syndrome. *Circulation*. 2009;120:1761–7.
388. Roden DM. Long-QT Syndrome. *N Engl J Med*. 2008;358:169–76.
389. Zipes DP. The Long QT Interval Syndrome A Rosetta Stone for Sympathetic related ventricular tachyarrhythmias. *Circulation*. 1991;84:1414–9.
390. Koopmann TT, Alders M, Jongbloed RJ, Guerrero S, Mannens MMAM, Wilde AAM, et al. Long QT syndrome caused by a large duplication in the KCNH2 (HERG) gene undetectable by current polymerase chain reaction-based exon-scanning methodologies. *Heart Rhythm*. 2006;3:52–5.
391. Eddy C, McCormick JM, Chung S, Crawford JR, Love DR, Rees MI, et al. Identification of large gene deletions and duplications in KCNQ1 and KCNH2 in patients with long QT syndrome. *Heart Rhythm*. 2008;5:1275–81.
392. Tester DJ, Benton AJ, Train L, Deal B, Baudhuin LM, Ackerman MJ. Prevalence and Spectrum of Large Deletions or Duplications in the Major Long QT Syndrome-Susceptibility Genes and Implications for Long QT Syndrome Genetic Testing. *Am J Cardiol* [Internet]. 2010;106:1124–8.
393. Barc J, Briec F, Schmitt S, Kyndt F, LeCunff M, Baron E, et al. Screening for Copy

- Number Variation in Genes Associated With the Long QT Syndrome Clinical Relevance. *J Am Coll Cardiol* [Internet]. 2011;57(1):40–7.
394. Campuzano O, Sarquella-Brugada G, Mademont-Soler I, Allegue C, Cesar S, Ferrer-Costa C, et al. Identification of genetic alterations, as causative genetic defects in long QT syndrome, using next generation sequencing technology. *PLoS One*. 2014;9(12).
395. Stattin E, Boström IM, Winbo A, Cederquist K, Jonasson J, Jonsson B, et al. Founder mutations characterise the mutation panorama in 200 Swedish index cases referred for Long QT syndrome genetic testing. *BMC Cardiovasc Disord*. 2012;12:1–12.
396. Williams VS, Cresswell CJ, Ruspi G, Yang T, Atak TC, Mcloughlin M, et al. Multiplex ligation-dependent probe amplification copy number variant analysis in patients with acquired long QT syndrome. *Europace*. 2015;17(4):635–41.
397. Morita H, Wu J, Zipes DP. The QT syndromes: long and short. *Lancet*. 2008;372:750–63.
398. Guerrier K, Kwiatkowski D, Czosek RJ, Spar DS, Anderson JB, Knilans TK. Short QT Interval Prevalence and Clinical Outcomes in a Pediatric Population. *Circ Arrhythm Electrophysiol*. 2015;
399. Moriya M, Seto S, Yano K, Akahoshi M. Two Cases of Short QT Interval. *PACE*. 2007;30(December):1522–6.
400. Crotti L, Taravelli E, Girardengo G, Schwartz PJ. Congenital Short QT Syndrome. *Indian Pacing J*. 2010;10(2):86–95.
401. Patel C, Yan G, Antzelevitch C. Basic Science for the Clinical Electrophysiologist Short QT Syndrome: From Bench to Bedside. *Circ Arrhythm Electrophysiol*. 2010;3:401–8.
402. Rudic B, Schimpf R, Borggreffe M. Short QT Syndrome – Review of Diagnosis and Treatment. *Arrhythmia Mech*. 2014;3(2):76–9.
403. Giustetto C, Schimpf R, Mazzanti A, Scrocco C, Maury P, Anttonen O, et al. Long-Term Follow-Up of Patients With Short QT Syndrome. *J Am Coll Cardiol* [Internet]. 2011;58(6):587–95.
404. Katz G, Arad M, Eldar M. Catecholaminergic Polymorphic Ventricular Tachycardia from Bedside to Bench and Beyond. *Curr Probl Cardiol* [Internet]. 2009;34(1):9–43.
405. Leenhardt A, Denjoy I, Guicheney P. Catecholaminergic Polymorphic Ventricular Tachycardia. *Circ Arrhythm Electrophysiol*. 2011;5:1044–52.
406. Priori SG, Napolitano C, Tiso N, Memmi M, Vignati G, Sorrentino V, et al. Mutations in the Cardiac Ryanodine Receptor Gene (hRyR2) Underlie Catecholaminergic Polymorphic Ventricular Tachycardia. *Circulation*. 2001;103:196–200.
407. Priori SG, Chairperson HRS, Wilde AA, Chairperson E, Horie M, Chairperson A, et al. HRS / EHRA / APHRS Expert Consensus Statement on the Diagnosis and Management of Patients with Inherited Primary Arrhythmia. *J Arrhythmia* [Internet]. 2014;30(1):1–28.
408. Bhuiyan ZA, Berg MP Van Den, Tintelen JP Van, Alders M, Postma A V, Langen I Van, et al. Expanding Spectrum of Human RYR2 - Related Disease New Electrocardiographic , Structural , and Genetic Features. *Circulation*. 2007;116:1569–76.
409. Marjamaa A, Laitinen-forsblom P, Lahtinen AM, Viitasalo M, Toivonen L, Kontula K, et al. Search for cardiac calcium cycling gene mutations in familial ventricular arrhythmias resembling catecholaminergic polymorphic ventricular tachycardia. *BMC Med Genet*. 2009;10:1–9.
410. Medeiros-Domingo A, Bhuiyan ZA, Tester DJ, Hofman N, Bikker H, Tintelen JP Van, et al. The RYR2-Encoded Ryanodine Receptor / Calcium Release Channel in Patients Diagnosed Previously With Either Catecholaminergic Polymorphic Ventricular Tachycardia or Genotype Negative , Exercise-Induced Long QT Syndrome A Comprehensive Open Reading Frame Mu. *J Am Coll Cardiol* [Internet]. 2009;54(22):2065–74.

411. Tang Y, Tian X, Wang R, Fill M, Chen S. Abnormal Termination of Ca²⁺ Release Is a Common Defect of RyR2 Mutations Associated With Cardiomyopathies. *Circ Res*. 2012;110:968–77.
412. Pérez-Serra A, Campuzano O, Brugada R. Update about atrial fibrillation genetics. *Curr Opin Cardiol*. 2017;32(3):246–52.
413. Fuster V, Rydén LE, Cannom DS, Crijns HJ, Curtis AB, Ellenbogen KA, et al. ACC / AHA / ESC Practice Guidelines ACC / AHA / ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation A Report of the American College of Cardiology / American Heart Association Task Force on Practice Guidelines and the European Soci. *Circulation*. 2006;114:e257–354.
414. Brugada R, Tapscott T, Czernuszewicz GZ, Marian A, Iglesias A, Mont L, et al. Identification of a genetic locus for familial atrial fibrillation. *N Engl J Med*. 1997;336:905–11.
415. Tsai C, Lai L, Lin J, Chiang F. Molecular Genetics of Atrial Fibrillation. *J Am Coll Cardiol*. 2008;52(4):241–50.
416. Gollob MH, Jones DL, Krahn AD, Danis L, Gong X-Q, Shao Q, et al. Somatic Mutations in the Connexin 40 Gene (GJA5) in Atrial Fibrillation. *N Engl J Med*. 2006;354:2677–88.
417. Tsai C, Hsieh C, Chang S, Chuang EY, Ueng K, Tsai C, et al. Genome-wide screening identifies a KCNIP1 copy number variant as a genetic predictor for atrial fibrillation. *Nat Commun [Internet]*. 2016;7(110):1–9.
418. Yetman AT, Temple J, Erickson CC. Radiofrequency ablation of a left-sided atrioventricular pathway in a patient with Marfan syndrome. *Cardiol Young*. 2002;12:494–5.
419. Universal Mutation Database [Internet]. [cited 2017 Apr 3].
420. Child A, Aragon-Martin J, Sage K. Genetic testing in Marfan syndrome. *Br J Hosp Med*. 2016;77(1):38–41.
421. Macdonald J, Ziman R, Yuen R, Feuk L, Scherer S. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42:986–92.
422. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J. ACMG Standards and Guidelines Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–24.
423. Sinard JH. Accounting for the Professional Work of Pathologists Performing Autopsies. *Arch Pathol Lab Med*. 2012;
424. Basso C, Burke M, Fornes P, Gallagher PJ, Gouveia RH De, Sheppard M, et al. Guidelines for autopsy investigation of sudden cardiac death. *Virchows Arch*. 2008;11–8.
425. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol [Internet]*. 2013;31(11):1023–31.
426. Samorodnitsky E, Datta J, Jewell BM, Hagopian R, Miya J, Wing MR, et al. Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. *J Mol diagnostics*. 2015;17(1):64–75.
427. Rehm HL, Bale SJ, Bayrak-toydemir P, Jonathan S, Brown KK, Deignan JL, et al. sequencing. *Genet Med*. 2013;15(9):733–47.
428. ENSEMBL [Internet]. [cited 2017 Jul 5]. Available from: <http://www.ensembl.org/>
429. NCBI-RefSeq Website [Internet]. [cited 2017 Jul 5]. Available from: <http://www.ncbi.nlm.nih.gov/refseq/>
430. NCBI-CCDS Website [Internet]. [cited 2017 Jul 5]. Available from: <https://www.ncbi.nlm.nih.gov/CCDS/>
431. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using Phred . I . Accuracy Assessment. *Genome Res*. 1998;175–85.

432. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* 1998;(206):186–94.
433. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods.* 2010;6(11).
434. Ferragina P, Manzini G. Opportunistic Data Structures with Applications. *Proc 41st Symp Found Comput Sci (FOCS 2000).* 2000;390–8.
435. Burrows M, Wheeler DJ. A Block-sorting Lossless Data Compression Algorithm. *SRC Res Rep.* 1994;
436. Marco-sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Commun.* 2012;9(12):1185–92.
437. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment / Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.
438. Broad Institute –Picard–.
439. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008;18:1851–8.
440. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A general approach to single-nucleotide polymorphism discovery. *Nat Genet.* 1999;23(december):452–6.
441. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
442. Depristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491–8.
443. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Bork P, Kondrashov AS, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7(4):248–9.
444. Choi Y, Sims G, Murphy S, Miller J, Chan A. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One.* 2012;7(10).
445. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods [Internet].* 2014;11(4):361–2.
446. NCBI-dbSNP Website.
447. Exome Sequencing Project.
448. 1000 Genomes Project.
449. Exome Aggregation Consortium.
450. Human Gene Mutation Database [Internet]. [cited 2017 Jul 5]. Available from: <http://www.hgmd.cf.ac.uk/ac/index.php>
451. ClinVar Website.
452. Beck T, Mullikin J, Blesecker L. Systematic Evaluation of Sanger Validation of NextGen Sequencing Variants. *Clin Chem.* 2016;0:1–8.
453. Kalari K, Casavant M, Bair T, Keen H. First Exons and Introns – A Survey of GC Content and Gene Structure in the Human Genome. *In Silico Biol.* 2006;6:237–42.
454. Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, et al. Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Rep [Internet].* 2012;1(5):543–56.
455. Aird D, Ross M, Chen W, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol [Internet].* 2011;12(2):R18.
456. Benjamini Y, Speed T. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 2012;40(10):e72.
457. Inc. G. GenoHub [Internet]. 2017 [cited 2017 Jun 28].
458. Olshen A, Venkatraman E. Circular binary segmentation for the analysis of array-based

- DNA copy number data. *Biostatistics*. 2004;5(4):557–72.
459. The Comprehensive R Archive Network [Internet]. [cited 2017 Jul 13]. Available from: <https://cran.r-project.org/>
460. Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
461. Talevich E, Shain A, Botton T, Bastian B. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873.
462. Lek M, Karczewski K, Eric V, Hill A, Cummings B, Tukiainen T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nat Publ Gr* [Internet]. 2016;536(7616):285–91.
463. Cooper D, Ball E, Krawczak M. The human gene mutation database. *Nucleic Acids Res*. 1998;26(1):285–7.
464. ClinGen [Internet]. [cited 2017 Jul 15]. Available from: <https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/>
465. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. [Internet]. [cited 2017 Jun 25]. Available from: <http://www.repeatmasker.org>
466. Mademont-soler I, Mates J, Yotti R, Espinosa M, Fernandez-avila A, Coll M, et al. Additional value of screening for minor genes and copy number variants in hypertrophic cardiomyopathy. *PLoS One*. 2017;12(8):e0181465.
467. Kalman L, Leonard J, Gerry N, Tarleton J, Bridges C, Gastier-foster J, et al. Quality Assurance for Duchenne and Becker Muscular Dystrophy Genetic Testing Development of a Genomic DNA Reference Material Panel. *J Mol diagnostics*. 2011;13(2):167–74.
468. Florian A, Rösch S, Bietenbeck M, Engelen M, Stypmann J, Waltenberger J, et al. Cardiac involvement in female Duchenne and Becker muscular dystrophy carriers in comparison to their first-degree male relatives: a comparative cardiovascular magnetic resonance study. *Eur Heart J*. 2016;17(3):326–33.
469. Medin M, Hermida-prieto M, Monserrat L, Laredo R, Rodriguez-rey J, Fernandez X, et al. Mutational screening of phospholamban gene in hypertrophic and idiopathic dilated cardiomyopathy and functional study of the PLN À 42 C > G mutation. *Eur J Heart Fail*. 2007;9:37–43.
470. Gene Cards [Internet]. [cited 2017 Jul 31]. Available from: <http://www.genecards.org/>
471. Hayashi T, Arimura T, Itoh-Satoh M. Tcap Gene Mutations in Hypertrophic Cardiomyopathy and Dilated Cardiomyopathy. *J Am Coll Cardiol*. 2004;44:2192–201.
472. Hastings R, Villiers C, Hooper C, Ormondroyd L, Pagnamenta A, Lise S, et al. Combination of Whole Genome Sequencing , Linkage and Functional Studies Implicates a Missense Mutation in Titin as a Cause of Autosomal Dominant Cardiomyopathy with Features of Left Ventricular Non-Compaction. *Circ Cardiovasc Genet*. 2016;9(5):426–35.
473. Karst M, Herron K, Olson T. X-Linked Nonsyndromic Sinus Node Dysfunction and Atrial Fibrillation Caused by Emerin Mutation. *J Cardiovasc Electrophysiology*. 2008;19(5):510–5.
474. Wang C, Wu M, Qian J, Li B, Tu X, Xu C, et al. Identification of rare variants in TNNI3 with atrial fibrillation in a Chinese GeneID population. *Mol Genet Genomics*. 2015;291(1):79–92.
475. Toruner G, Kurvathi R, Sugalski R, Shulman L, Twersky S, Pearson P, et al. Copy number variations in three children with sudden infant death. *Clin Genet*. 2009;76(3):63–8.
476. Quintela I, Eiris J, Gómez-lado C, Pérez- L, Dacruz D, Cruz R, et al. Copy number variation analysis of patients with intellectual disability from North-West Spain. *Gene* [Internet]. 2017;30(626):189–99.
477. Costain G, Silversides C, Bassett A. The importance of copy number variation in congenital heart disease. *Genomic Med*.

- 2016;(April).
478. Broad Institute Genomic Services [Internet]. [cited 2017 Aug 10]. Available from: <http://genomics.broadinstitute.org/>
479. Pfundt R, Rosario M, Vissers L, Kwint M, Janssen I, Leeuw Nd, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet Med*. 2016;(March):1–9.
480. Pugh T, Amr S, Bowser M, Gowrisankar S, Hynes E, Mahanta L, et al. VisCap : inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet Med*. 2015;18:712–719.
481. Kearney H, Thorland E, Brown K, Quintero-Rivera F, South S. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med*. 2011;13(7):680–5.
482. Newman S, Hermetz K, Weckselblatt B, Rudd M. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* [Internet]. 2015;96(2):208–20.
483. Moncayo-arlandi J, Brugada R. Unmasking the molecular link between arrhythmogenic cardiomyopathy and Brugada syndrome. *Nat Rev Cardiol* [Internet]. 2017;
484. La Marche P, Heisler A, Kronemer N. Disappearing mosaicism. Suggested mechanism is growth advantage of normal over abnormal cell population. *R I Med J*. 1967;50(3):184–9.
485. Ma H, Marti-gutierrez N, Park S, Wu J, Lee Y, Suzuki K, et al. Correction of a pathogenic gene mutation in human embryos. *Nature* [Internet]. 2017;0(0):1–7.
486. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna J, Charpentier E. A Programmable Dual-RNA – Guided. *Science* (80-). 2012;337(August):816–22.
487. Cong L, Ran F, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR / Cas Systems. *Scienceexpress*. 2013;339:819–23.
488. DellEra P, Benzoni P, Crescini E, Valle M, Xia E, Consiglio A, et al. Cardiac disease modeling using induced pluripotent stem cell-derived human cardiomyocytes. *World J Stem Cells*. 2015;7(2):329–42.
489. Goodwin S, Mcpherson J, McCombie W. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* [Internet]. 2016;17(6):333–51.
490. Schaaf C, Wiszniewska J, Beaudet A. Copy Number and SNP Arrays in Clinical Diagnostics. *Annu Rev Genomics Hum Genet*. 2011;12:25–51.
491. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent Q, Antipenko A. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci*. 2015;112(17):5473–5478.
492. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet*. 2016;2:5–8.

ANNEX 1

HOJA DE INFORMACION AL PACIENTE

Introducción

Se le ha ofrecido la posibilidad de proceder al análisis genético a partir de una muestra de saliva o sangre a fin de determinar sus variantes genéticas asociadas a enfermedades del corazón como las miocardiopatías y canalopatías hereditarias.

Se considera que el resultado del test, junto a otros factores clínicos y patológicos, es un factor pronóstico de la aparición de la enfermedad.

Para este análisis se emplea Sudd inCode®, un servicio de análisis genético que analiza genes implicados en la función del músculo cardíaco, en muestras de sangre o saliva. Sudd inCode® comprobará sus características genéticas y podrá dar información de gran valor que ayudará a su médico a encontrar el plan de tratamiento más específico y adecuado para su situación particular.

Propósito

Este sistema de análisis ha sido desarrollado con la intención de conseguir un mayor conocimiento de su posible riesgo de sufrir una disfunción coronaria mediante la identificación de variantes genéticas.

Su médico valorará la información genética derivada de este análisis y junto con otros factores bioquímicos, clínicos y funcionales y le permitirá efectuar una recomendación terapéutica o de hábitos de vida.

Procedimiento de estudio

Si está de acuerdo en la utilización de este análisis, su médico le indicará los pasos a seguir para obtener una muestra de su saliva en la misma consulta o bien como proceder para extraer una muestra de 5 ml de sangre periférica. Su médico posteriormente enviará la muestra a un laboratorio centralizado en Girona.

El empleo de este análisis no implica ningún otro examen, ni intervención, ni otro procedimiento médico, ni otra molestia ni ningún otro riesgo adicional para Vd.

Al término de la fase analítica, la muestra se mantendrá congelada a disposición de usted y de su médico durante 1 año a efectos de poder efectuar un contra análisis. Transcurrido dicho período y a no ser que usted así lo autorice expresamente, solo podrá conservarse si los datos de carácter personal de la misma han sido sometidos a su disociación y por tanto mantienen el anonimato.

Información sobre los resultados y consejo genético

Se le informará del resultado y de los datos genéticos de carácter personal que se obtengan del análisis. Si lo desea puede revocar el consentimiento o prescindir de conocer los resultados del análisis. En este último supuesto se suministrarán al facultativo principal sólo aquellos datos estrictamente necesarios aceptados por usted. Se le garantiza el correspondiente asesoramiento genético sobre el resultado del análisis.

Es importante que tenga en cuenta que las enfermedades genéticas pueden heredarse en la familia y por tanto, los resultados de su test pueden tener implicaciones para su propia familia.

En el caso del estudio genético de una mutación, la identificación de la mutación tiene carácter diagnóstico, mientras que la no identificación no es excluyente de la patología. Un test negativo no excluye la posibilidad de tener la enfermedad (algunas enfermedades tienen múltiples causas y en la actualidad, no es posible probarlas todas).

Riesgos asociados

No existen riesgos significativos asociados con la extracción de una muestra de su saliva o de su sangre. La extracción de sangre, en el caso de que sea necesario, puede ser incómoda, ocasionalmente puede producir cierto dolor y, raramente, desmayo. Sólo personal experto será el responsable de extraer una muestra de su sangre.

En algunas ocasiones, el laboratorio podría tener dificultades en analizar la muestra y podría ser requerida una segunda extracción. Aunque los métodos empleados para hacer este diagnóstico genético son altamente sensibles y específicos, siempre existe una pequeña posibilidad de fracaso de la técnica o error de interpretación. En ocasiones, pueden existir ciertas alteraciones en la estructura del ADN de determinados individuos que pueda llevar a resultados de difícil interpretación, dificultando el diagnóstico e incluso haciendo imposible la obtención de un resultado concluyente.

Confidencialidad y manejo de datos personales por el laboratorio de genética

La confidencialidad y privacidad serán respetadas. Ningún tipo de información que pueda revelar su identidad será publicada sin su consentimiento específico. Su identidad no será empleada en ningún informe del análisis. En los registros que partan de este centro será identificado solamente por un código.

El facultativo principal, y su equipo, serán los únicos agentes que tendrán acceso a sus datos de carácter personal, quienes están sometidos al deber de reserva y confidencialidad. El tratamiento de los datos de carácter personal y genéticos estará reversiblemente dissociado o codificado, de tal manera que únicamente puedan identificarle el facultativo principal y su equipo.

Sus datos de carácter personal y genéticos se conservarán durante un periodo de 5 años y, transcurrido dicho periodo, mientras sean necesarios para preservar su salud, si no ha ejercitado su derecho de cancelación. Lo anterior es sin perjuicio de que dichos datos puedan conservarse con fines de investigación, de forma que los datos se mantengan en el anonimato, es decir, sin que sea posible su identificación.

Finalmente, le informamos que sus de datos de carácter personal quedarán recogidos en un fichero cuyo responsable es [Centro/Hospital] que los utilizará únicamente y exclusivamente para la consecución del propósito descrito en la presente Hoja Informativa, salvo que medie autorización expresa suya.

No obstante, usted puede ejercitar los derechos de acceso, rectificación, cancelación y oposición comunicándolo por escrito a [Centro/Hospital] a la siguiente dirección: [] a la atención del departamento [].

FORMULARIO DE CONSENTIMIENTO INFORMADO

- He recibido una copia de la Hoja de Información al Paciente y he comprendido la información.
- He tenido suficiente tiempo para tomar mi decisión.
- Autorizo a mi médico a enviar una muestra biológica de mi saliva o sangre y la información clínica relevante para su procesamiento por el servicio Sudd inCode®. La información clínica podría ser relevante para la correcta interpretación del análisis genético.
- Comprendo que en algunas ocasiones, el laboratorio podría tener dificultades en analizar mi muestra y que una segunda muestra podría ser requerida.
- Mi consentimiento es completamente voluntario y no afectará mi relación con el médico que me trata. Los datos que se obtengan a partir de mi serán estrictamente confidenciales y tratados de acuerdo con la Ley Orgánica, 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal y legislación que la desarrolla, así como de conformidad a la Ley 14/2007, de 3 de julio, de Investigación Biomédica. Mi consentimiento no libera de sus responsabilidades a las personas y/o entidades implicadas en todo el proceso del análisis dejando a salvo todos mis derechos garantizados por ley.
- Comprendo que, en el interés de la ciencia, resúmenes de los resultados de este análisis podrían ser publicados en ámbitos científicos. No obstante, en ningún caso será posible identificar en ellos información confidencial relativa a mi persona, a no ser que previamente la haya expresamente autorizado por escrito.

Autorizo a que mi muestra pueda conservarse con fines de investigación, de forma que mis datos se mantengan en el anonimato, es decir, sin que sea posible su identificación.

En cualquier momento puedo cambiar de parecer y denegar la autorización para el estudio genético que doy en este documento, y revocar así mi decisión de continuar con el análisis.

Nombre del/la paciente: _____

Firma del/la paciente: _____ Fecha: _____

Nombre del representante legal: _____

Firma del representante legal: _____ Fecha: _____

Persona designada por el clínico para participar en el proceso de consentimiento informado:

Nombre: _____

Firma: _____ Fecha: _____

Nombre del médico: _____ Título/ Posición: _____

Firma del médico: _____ Fecha: _____

ANNEX 2

Taula A-2 | Resum dels gens (i les isoformes) inclosos en els diferents panells de gens utilitzats.

Panell	Gens i isoformes
Tots	<i>ACTC1</i> (NM_005159), <i>ACTN2</i> (NM_001103), <i>ANK2</i> (NM_001148), <i>CACNA1C</i> (NM_001129827;NM_000719), <i>CACNB2</i> (NM_201596;NM_201590), <i>CASQ2</i> (NM_001232), <i>CAV3</i> (NM_033337), <i>CRYAB</i> (NM_001885), <i>CSRP3</i> (NM_003476), <i>DES</i> (NM_001927), <i>DMD</i> (NM_004006), <i>DSC2</i> (NM_024422), <i>DSG2</i> (NM_001943), <i>DSP</i> (NM_004415), <i>EMD</i> (NM_000117), <i>FBN1</i> (NM_000138), <i>GLA</i> (NM_000169), <i>GPD1L</i> (NM_015141), <i>HCN4</i> (NM_005477), <i>JPH2</i> (NM_020433), <i>JUP</i> (NM_002230), <i>KCNE1</i> (NM_000219), <i>KCNE2</i> (NM_172201), <i>KCNH2</i> (NM_000238), <i>KCNJ2</i> (NM_000891), <i>KCNQ1</i> (NM_000218), <i>LAMP2</i> (NM_002294), <i>LDB3</i> (NM_001080116), <i>LMNA</i> (NM_170707), <i>MYBPC3</i> (NM_000256), <i>MYH6</i> (NM_002471), <i>MYH7</i> (NM_000257), <i>MYL2</i> (NM_000432), <i>MYL3</i> (NM_000258), <i>MYOZ2</i> (NM_016599), <i>PDLIM3</i> (NM_014476), <i>PKP2</i> (NM_004572), <i>PLN</i> (NM_002667), <i>PRKAG2</i> (NM_016203), <i>RYR2</i> (NM_001035), <i>SCN4B</i> (NM_174934), <i>SCN5A</i> (NM_198056), <i>SGCD</i> (NM_000337), <i>TAZ</i> (NM_000116), <i>TGFB3</i> (NM_003239), <i>TGFBR2</i> (NM_003242), <i>TNNC1</i> (NM_003280), <i>TNNI3</i> (NM_000363), <i>TNNT2</i> (NM_001001430), <i>TPM1</i> (NM_001018005), <i>TTN</i> (NM_133378), <i>VCL</i> (NM_014000)
78, 85, 118, 147	<i>ABCC9</i> (NM_005691), <i>AKAP9</i> (NM_005751), <i>BAG3</i> (NM_004281), <i>CACNA2D1</i> (NM_000722), <i>FKTN</i> (NM_001079802), <i>KCND3</i> (NM_004980), <i>KCNE3</i> (NM_005472), <i>KCNE5</i> (NM_012282), <i>KCNJ5</i> (NM_000890), <i>KCNJ8</i> (NM_004982), <i>MYPN</i> (NM_032578), <i>NEBL</i> (NM_006393), <i>NEXN</i> (NM_144573), <i>NOSTAP</i> (NM_014697;NM_001164757), <i>RANGRF</i> (NM_016492), <i>RBM20</i> (NM_001134363), <i>SCN1B</i> (NM_001037;NM_199037), <i>SCN2B</i> (NM_004588), <i>SLMAP</i> (NM_007159), <i>SNTA1</i> (NM_003098), <i>TCAP</i> (NM_003673), <i>TMEM43</i> (NM_024334), <i>TMPO</i> (NM_003276;NM_001032283), <i>TP63</i> (NM_003722), <i>TRDN</i> (NM_006073), <i>TRPM4</i> (NM_017636)
85, 118, 147	<i>SCN3B</i> (NM_018400), <i>TTR</i> (NM_000371)
85, 147	<i>DMPK</i> (NM_001081563), <i>FLNA</i> (NM_001110556), <i>FLNC</i> (NM_001458), <i>SCN10A</i> (NM_006514), <i>TRIM63</i> (NM_032588)
118, 147	<i>ACTA2</i> (NM_001613;NM_001141945), <i>CACNA1G</i> (NM_018896), <i>CACNA1H</i> (NM_021098), <i>CACNA1I</i> (NM_021096), <i>CALM1</i> (NM_006888), <i>COL3A1</i> (NM_000090), <i>CTF1</i> (NM_001330), <i>CTNNA3</i> (NM_013266), <i>DPP6</i> (NM_001936;NM_001039350), <i>ECE1</i> (NM_001397), <i>EN1</i> (NM_001426), <i>EYA4</i> (NM_172105;NM_004100), <i>FBN2</i> (NM_001999), <i>FHL2</i> (NM_201555;NM_001450), <i>GJA1</i> (NM_000165), <i>GJA5</i> (NM_005266), <i>HCN1</i> (NM_021072), <i>HCN2</i> (NM_001194), <i>KCNA5</i> (NM_002234), <i>KCNE4</i> (NM_080671), <i>LAMA4</i> (NM_001105206;NM_002290), <i>MYH11</i> (NM_002474;NM_001040114), <i>MYLK2</i> (NM_033118), <i>NOTCH1</i> (NM_017617), <i>NPPA</i> (NM_006172), <i>NUP155</i> (NM_153485), <i>PHOX2A</i> (NM_005169), <i>PHOX2B</i> (NM_003924), <i>PITX2</i> (NM_153426), <i>RET</i> (NM_020975), <i>SLC6A4</i> (NM_001045), <i>SLC8A1</i> (NM_021097), <i>SLN</i> (NM_003063), <i>SMAD3</i> (NM_005902), <i>TGFB2</i> (NM_001135599;NM_003238), <i>TGFBR1</i> (NM_004612), <i>TGFBR3</i> (NM_003243), <i>TLX3</i> (NM_021025)
55	<i>SGCA</i> (NM_000023), <i>SGCB</i> (NM_000232)
147	<i>ANKRD1</i> (NM_014391), <i>BRAF</i> (NM_004333), <i>CALM2</i> (NM_001743), <i>CALM3</i> (NM_005184), <i>CALR3</i> (NM_145046), <i>CBL</i> (NM_005188), <i>DTNA</i> (NM_032975), <i>EYA1</i> (NM_000503), <i>GAA</i> (NM_000152), <i>HRAS</i> (NM_176795), <i>KRAS</i> (NM_033360), <i>MAP2K1</i> (NM_002755), <i>MAP2K2</i> (NM_030662), <i>NF1</i> (NM_001042492), <i>NRAS</i> (NM_002524), <i>PTPN11</i> (NM_002834), <i>RAF1</i> (NM_002880), <i>RIT1</i> (NM_001256821), <i>SDHA</i> (NM_004168), <i>SHOC2</i> (NM_007373), <i>SLC22A5</i> (NM_003060), <i>SOS1</i> (NM_005633), <i>SOS2</i> (NM_006939), <i>SPRED1</i> (NM_152594)

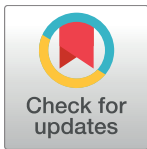
ANNEX 3

RESEARCH ARTICLE

Additional value of screening for minor genes and copy number variants in hypertrophic cardiomyopathy

Irene Mademont-Soler^{1,2}, Jesus Mates¹, Raquel Yotti^{2,3*}, Maria Angeles Espinosa^{2,3}, Alexandra Pérez-Serra^{1,2}, Ana Isabel Fernandez-Avila^{2,3}, Monica Coll^{1,2}, Irene Méndez^{2,3}, Anna Iglesias^{1,2}, Bernat del Olmo¹, Helena Riuró¹, Sofía Cuenca^{2,3}, Catarina Allegue¹, Oscar Campuzano^{1,2,4}, Ferran Pico¹, Carles Ferrer-Costa⁵, Patricia Álvarez⁵, Sergio Castillo⁵, Pablo Garcia-Pavia^{2,6}, Esther Gonzalez-Lopez^{2,6}, Laura Padron-Barthe^{2,6}, Aranzazu Díaz de Bustamante⁷, María Teresa Darnaude⁷, José Ignacio González-Hevia⁸, Josep Brugada^{2,9}, Francisco Fernandez-Aviles^{2,3}, Ramon Brugada^{1,2,4,10}

1 Cardiovascular Genetics Center, University of Girona-IDIBGI, Girona, Spain, **2** Centro de Investigación Biomédica en Red de Enfermedades Cardiovasculares (CIBERCV), Madrid, Spain, **3** Department of Cardiology, Hospital General Universitario Gregorio Marañón, Instituto de Investigación Sanitaria Gregorio Marañón. Universidad Complutense, Madrid, Spain, **4** Department of Medical Sciences, School of Medicine, University of Girona, Girona, Spain, **5** Gendiag.exe SL, Barcelona, Spain, **6** Inherited Cardiac Diseases Unit, Department of Cardiology, Hospital Universitario Puerta de Hierro, Francisco de Vitoria University, Madrid, Spain, **7** Genetics Unit, Hospital Universitario de Móstoles, Madrid, Spain, **8** Hospital Universitario Miguel Servet, Zaragoza, Spain, **9** Arrhythmia Unit, Hospital Clinic de Barcelona, University of Barcelona, Barcelona, Spain, **10** Cardiovascular Genetics Unit, Hospital Universitari Dr. Josep Trueta, Girona, Spain



OPEN ACCESS

Citation: Mademont-Soler I, Mates J, Yotti R, Espinosa MA, Pérez-Serra A, Fernandez-Avila AI, et al. (2017) Additional value of screening for minor genes and copy number variants in hypertrophic cardiomyopathy. PLoS ONE 12(8): e0181465. <https://doi.org/10.1371/journal.pone.0181465>

Editor: Chunhua Song, Pennsylvania State University, UNITED STATES

Received: March 22, 2017

Accepted: June 30, 2017

Published: August 3, 2017

Copyright: © 2017 Mademont-Soler et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Individual-level variant data is queryable from Cardiovascular Genetics Center under terms determined by Hospital General Universitario Gregorio Marañón and Institut d'Investigació Biomèdica de Girona for scientific collaboration with not-for profit entities. The data sharing would take place after acceptance of the conditions stipulated through a material transfer agreement. This was imposed by an Institutional Review Board. For more details, please contact Dr. Ramon Brugada (ramon@brugada).

These authors contributed equally to this work.

* raquel.yotti@salud.madrid.org

Abstract

Introduction

Hypertrophic cardiomyopathy (HCM) is the most prevalent inherited heart disease. Next-generation sequencing (NGS) is the preferred genetic test, but the diagnostic value of screening for minor and candidate genes, and the role of copy number variants (CNVs) deserves further evaluation.

Methods

Three hundred and eighty-seven consecutive unrelated patients with HCM were screened for genetic variants in the 5 most frequent genes (*MYBPC3*, *MYH7*, *TNNT2*, *TNNI3* and *TPM1*) using Sanger sequencing (N = 84) or NGS (N = 303). In the NGS cohort we analyzed 20 additional minor or candidate genes, and applied a proprietary bioinformatics algorithm for detecting CNVs. Additionally, the rate and classification of *TTN* variants in HCM were compared with 427 patients without structural heart disease.

Results

The percentage of patients with pathogenic/likely pathogenic (P/LP) variants in the main genes was 33.3%, without significant differences between the Sanger sequencing and NGS cohorts. The screening for 20 additional genes revealed LP variants in *ACTC1*, *MYL2*,

com) or Dr. Raquel Yotti (raquel.yotti@salud.madrid.org).

Funding: This work was supported by: ió “Obra social La Caixa”: Ramon Brugada; Instituto de Salud Carlos III (Fondo Investigación Sanitaria -FIS- (PI14/01773): Ramon Brugada; Sociedad Española de Cardiología (Proyecto Investigación Básica Cardiología 2015 de los Socios Estratégicos SEC); Oscar Campuzano; Instituto de Salud Carlos III (Fondo Investigación Sanitaria -FIS- (PI15/02222; BA16/00032): Raquel Yotti. The funder provided support in the form of salaries for authors CF-C, PA and SC, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the ‘author contributions’ section.

Competing interests: I have read the journal’s policy and the authors of this manuscript have the following competing interests: Dr. Ramon Brugada is consultant of Ferrer-inCode; and Dr. Ferrer-Costa, Dr. Patricia Álvarez and Dr. Sergio Castillo are employed by Gendiag.exe SL. This commercial affiliation does not alter our adherence to PLOS ONE policies on sharing data and materials. The other authors declare no conflicts of interest to disclose.

MYL3, *TNNC1*, *GLA* and *PRKAG2* in 12 patients. This approach resulted in more inconclusive tests (36.0% vs. 9.6%, $p < 0.001$), mostly due to variants of unknown significance (VUS) in *TTN*. The detection rate of rare variants in *TTN* was not significantly different to that found in the group of patients without structural heart disease. In the NGS cohort, 4 patients (1.3%) had pathogenic CNVs: 2 deletions in *MYBPC3* and 2 deletions involving the complete coding region of *PLN*.

Conclusions

A small percentage of HCM cases without point mutations in the 5 main genes are explained by P/LP variants in minor or candidate genes and CNVs. Screening for variants in *TTN* in HCM patients drastically increases the number of inconclusive tests, and shows a rate of VUS that is similar to patients without structural heart disease, suggesting that this gene should not be analyzed for clinical purposes in HCM.

Introduction

Hypertrophic cardiomyopathy (HCM) is characterized by left ventricular hypertrophy with histologic features of cellular hypertrophy, myofibrillar disarray, and interstitial fibrosis. With a prevalence of 0.2% in the adult population, HCM is the most common inherited cardiac disease and a major cause of sudden cardiac death (SCD) in young people [1]. The disease has marked phenotypic variability, and clinical manifestations range from asymptomatic clinical course to severe heart failure and SCD. The identification of a disease-causing variant in a patient is crucial for diagnosis confirmation in borderline cases, early management of at-risk family members, genetic counseling and avoidance of unnecessary follow-up of non-carriers. The latter, besides doubtless clinical benefit, enables significant health-care costs saving [2–4]. For all these reasons, current guidelines recommend genetic testing in patients fulfilling diagnostic criteria for HCM, but the advantages of screening for genes without a definitive evidence of disease association versus more conservative approaches remain to be determined [5].

Overall, in patients fulfilling HCM diagnostic criteria, genetic testing leads to the identification of disease-causing genetic variants in 32–78.9% of cases, depending on the clinical characteristics of the patients, the number of genes studied, and the criteria used for variant classification [4, 6–19]. Most HCM cases are caused by mutations in genes that encode sarcomere proteins [19–21]. Among them, about 85% of pathogenic variants are found in *MYBPC3* and *MYH7*, 10% in cardiac troponin T (*TNNT2*) and troponin I (*TNNI3*), up to 2% in *TPM1*, and less than 3% in other sarcomere genes (*MYL2*, *MYL3*, *ACTC1* and *TNNC1*). For this reason, initial studies using Sanger sequencing in HCM recommended to focus on the 5 principal sarcomere genes [20]. Recent improvements in DNA-sequencing technologies offer the opportunity to screen for a larger number of genes in a time and cost-effective manner. However, this approach also results in an increase of the number of rare genetic variants of unknown significance (VUS), which may entail a clinical challenge.

The analysis of a predefined panel of HCM-related genes using Next-Generation Sequencing (NGS) technologies has emerged as the preferred genetic testing methodology for clinical purposes in HCM. This approach allows the additional screening for genes that have been previously proposed to cause a relatively small number of HCM cases (minor genes) and genes

with a controversial role in the disease (candidate genes) [12, 19, 22]. Moreover, panels can easily include genes related to metabolic disorders that account for rare cases of unexplained left ventricular hypertrophy in adults (<5%) but whose identification is of great clinical relevance [5]. Finally, NGS enables the detection of alterations in the number of copies of large genomic regions, known as Copy Number Variants (CNVs). Recently, two large NGS series involving the screening for these variants in HCM-associated genes have shown that 0.56–0.8% of HCM cases may be explained by these large imbalances [23, 24].

The aim of the present study was to determine the prevalence and spectrum of clinically relevant genetic variants in a Spanish cohort of HCM patients and analyze the additional clinical value provided by the screening for minor and candidate HCM genes and CNVs using NGS. The value and clinical challenges derived from the screening for variants in *TTN* were specifically addressed and compared with an independent cohort of patients without structural heart disease.

Materials and methods

Study population

The study cohort includes 387 consecutive unrelated Spanish patients with clinical diagnosis of HCM according to current clinical criteria [5], referred for genetic testing between 2012 and 2016. The study was approved by the ethical committee of Hospital Universitari Dr. Josep Trueta de Girona (Spain) and conformed to the ethical guidelines of the Declaration of Helsinki 2008. Informed written consent was obtained from all subjects. Patients were recruited at the Inherited Heart Diseases Units from Hospital General Universitario Gregorio Marañón, Hospital Universitari Dr. Josep Trueta, Hospital Clínic de Barcelona and Hospital Universitario Puerta del Hierro. The mean age at the time of the genetic study was 48 ± 20 years, 255 patients (65.9%) were men and 132 (34.1%) women. Family members of carriers of rare non-synonymous variants, indels and/or CNVs were invited to undergo genetic analysis. During the period of the study 180 relatives were referred for genetic testing and were studied by Sanger sequencing or MLPA.

The detection rate and classification of rare variants in *TTN* in the HCM cohort was compared with the results obtained in an independent group of 427 unrelated patients without echocardiographic evidence of structural heart disease (30 healthy subjects, 191 patients with Brugada syndrome, 138 with long QT syndrome, 20 with catecholaminergic polymorphic ventricular tachycardia, 8 with short QT syndrome, 9 with atrioventricular block, 7 with idiopathic ventricular tachycardia/ventricular fibrillation, 5 with atrial fibrillation and 19 with other arrhythmias).

Genetic analysis

Total genomic DNA was isolated from blood or saliva samples using Chemagen MSM I (PerkinElmer, Germany). All patients were screened for the 5 more frequent sarcomere genes (*MYBPC3*, *MYH7*, *TNNI3*, *TNNT2* and *TPM1*) (isoforms analyzed are listed in [S1 Table](#)). The first 84 patients underwent Sanger sequencing of these 5 genes and the remaining 303 patients were studied using expanded NGS panels.

Sanger sequencing. The coding regions and exon-intron boundaries (± 10 bp) of the 5 main sarcomere genes were amplified by PCR and, after purification, the PCR products were directly sequenced in both directions using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, TX, USA). Sequencing products were run on 3130XL Genetic Analyzer (Applied Biosystems) and analyzed by means of SeqScape Software v2.5 (Life Technologies, CA, USA).

NGS. Three hundred and three patients were analyzed with custom NGS panels (55 or 78 genes) including the coding regions and exon-intron boundaries (± 10 bp) of the most prevalent genes associated with inherited cardiac diseases (the 55-gene panel includes the UTR sequences of some genes). Coordinates of sequence data were based on UCSC human genome version hg19 (GRCh37). Both NGS panels were developed by Gendiag.exe S.L. and commercialized by Ferrer InCode as SudD inCode®.

Using either panel, we focused on the analysis of 25 genes previously associated or candidate for HCM. Reference sequence transcripts listed in [S1 Table](#) were analyzed. Only eight sarcomere genes definitively associated with HCM (*MYBPC3*, *MYH7*, *TNNI3*, *TNNT2*, *TPM1*, *ACTC1*, *MYL2* and *MYL3*) and genes robustly associated with metabolic diseases that can mimic HCM (*GLA*, *LAMP2* and *PRKAG2*) were considered validated genes [22]. The additional 14 genes (*CSRP3*, *PLN*, *ACTN2*, *MYOZ2*, *MYH6*, *TNNC1*, *CAV3*, *JPH2*, *LDB3*, *RYR2*, *TCAP*, *VCL*, *PDLIM3* and *TTN*) were classified as candidate genes. Following the criteria described by Walsh et al. [22], these candidate genes have different levels of evidence for their association with HCM, with the exception of *PDLIM3* and *VCL* (no supporting evidence), and *RYR2* and *TTN* (evidence not analyzed). We analyzed separately the additional value and clinical challenges derived from the screening for variants in *TTN*, due to the limited evidence for its association with HCM and the high background variation of this gene.

Sample libraries were prepared following the SureSelect XT Target Enrichment System for Illumina Paired-End Sequencing Library protocol (Agilent Technologies, CA, USA). Indexed libraries were sequenced in ten-sample pools on a MiSeq platform (Illumina, CA, USA), with 2x75 bp reads length.

An algorithm developed by Gendiag.exe SL was used to process the FASTQ files to obtain clean BAM files for the subsequent analysis of both SNVs and indels. In brief, the processed raw reads obtained after sequencing were trimmed and mapped with BWA-MEM [25]. The output from mapping steps was joined and sorted, and only the uniquely and properly mapped read pairs were selected. Then the variant call was performed with SAMtools v.1.2 [26], together with an *ad hoc* developed script. Both the custom NGS gene panels and the bioinformatics algorithm used for the detection of SNVs and indels had been previously validated in our center, obtaining a sensitivity of 100% and a specificity of 99.5% (unpublished data). The identified SNVs and indels were annotated with dbSNP [27], Exome Sequencing Project (ESP) [28], 1000 Genomes Project [29], Exome Aggregation Consortium (ExAC) [30], Human Gene Mutation Database (HGMD) [31] and ClinVar [32]. Sanger sequencing was performed to sequence regions with coverage lower than 30X, as well as to validate the uncommon non-synonymous variants identified. Genetic variants were reported following the recommendations of the Human Genome Variation Society.

We used a bioinformatics algorithm developed in our laboratory to detect CNVs using NGS data. The approach focuses on capturing significant differences between the expected and observed normalized coverage for a given sample in every exon of the genes included in the NGS panels. Raw coverage is first normalized by the amount of DNA yielded for each sample in the run. Then the insert size and the low probe affinity bias for targeted regions with a too high or too low GC content ($>75\%$ and $<45\%$, respectively) are corrected. Finally, the ratio between each sample and a built-in baseline is evaluated. If the ratio falls outside a signal-to-noise window and is greater or lower than the duplication or deletion cut-offs (0.45 and -0.8, respectively), the gain or loss is inferred. Each potential CNV was visually reviewed to discard possible false positives due to artefacts caused by samples with enrichment inconsistencies generated during the library preparation protocol. Sensitivity and specificity of the method were assessed in an independent cohort including 108 patients with different cardiovascular

diseases (16 of them with known CNVs and the remaining without this type of rearrangements), and they were 100% and 90.7%, respectively (unpublished data).

Each CNV identified by NGS was validated by an alternative method (Multiplex Ligation-dependent Probe Amplification -MLPA- or quantitative PCR -qPCR-). MLPA analyses were performed using commercially available SALSA MLPA probemixes and following manufacturer's instructions (MRC-Holland, Amsterdam, The Netherlands). After the multiplex PCR reaction, electrophoresis was performed using ABI3130XL Genetic Analyzer with LIZ500 size standard (both from Applied Biosystems), and results were analyzed using [Coffalyser.Net](#) (MRC-Holland). qPCR analyses were performed with the QuantStudio 7 Flex System using Power Up Sybr Green master mix (both from Life Technologies), following manufacturer's recommendations. Results were analyzed with QuantStudio Real-Time PCR Software v1.2 (Life Technologies).

For precise characterization of the CNVs, the breakpoints were assessed using NGS split-read data when the breakpoint regions were covered by the NGS panels, and then they were confirmed by Sanger sequencing. When no split-read data were available, Sanger sequencing was performed in an attempt to characterize the breakpoints using primers located in the non-altered regions of the gene of interest.

Summary of NGS data results. In the present study, including all MiSeq runs, the average call rate achieved at 30x with the custom enrichment gene designs of 55 and 78 genes was 99.7% and 99.8%, respectively. The median percentage of reads overlapping our target regions was 48% (range 39% to 51%) for the first panel and 66% (range 52% to 69%) for the second one. The median coverage per sample was 870 (721 to 1069) and 679 (479 to 867), respectively. The 25 and 75 percentiles were 571 and 1099 for the 55 gene design, and 509 and 843 for the 78 gene design.

Criteria for interpretation of SNVs, indels and CNVs

Rare variants (SNVs and indels) were defined as variants with a minor allele frequency (MAF) <0.002 [19, 33] in the databases dbSNP [27], ESP [28], 1000 Genomes Project [29] and Exome Aggregation Consortium (ExAC) [30]. We chose this conservative and inclusive cut-off to ensure the selection of all potentially relevant variants for the subsequent process of individual variant classification (see below). Additionally, to analyze the impact of the MAF filter applied on the final number and classification of genetic variants, we compared this approach with a more restrictive hard filtering recently proposed by Walsh et al., based on the frequency of the most common pathogenic HCM variant (MAF <0.0001 in ExAC) [22]. Both MAF criteria were also used to compare the detection rate of rare variants in *TTN* in HCM patients and individuals without a structural heart disease.

We used the updated American College of Medical Genetics and Genomics (ACMG) 2015 guidelines for variant interpretation to classify variants in 5 categories: pathogenic (P), likely pathogenic (LP), VUS, likely benign (LB) or benign (B) [34]. For the assessment of the clinical significance of previously reported variants we first searched for information in public variant databases, population cohorts and scientific literature. Then, available clinical, experimental and computational data were integrated with potential additional information obtained from the study of the particular family to reach a final clinical conclusion. The strength of the association with the disease at the gene-level was classified as strong, moderate, weak, only supported in functional data or no evidence [22].

Novel variants were defined as variants not previously reported in patients (published literature, HGMD [31] or ClinVar [32]) and absent from controls in ESP, 1000 Genomes Project [29], ExAC and Genome Aggregate Database [30]. Novel variants that did not meet strict

ACMG criteria of pathogenicity (VUS) but exhibited at least one supportive criteria were denominated novel candidate variants. For this purpose we considered the following criteria: 1) location in a mutational hot spot and/or critical and well-established functional domain, 2) protein length changes as a result of in-frame deletions/insertions in a nonrepeat region, 3) missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before, 4) cosegregation with disease in ≥ 2 affected family members, or 5) multiple lines of computational evidence support a deleterious effect (probably or possibly damaging/deleterious/disease causing by three *in silico* prediction tools: PolyPhen-2 [35], Provean [36] and Mutation Taster [37]).

Finally, if the segregation study of a large family enabled the classification of a variant as LB (no segregation), such labeling was extrapolated to other index cases with the same genetic variant. In the description of *TTN* variants, we included their location in the main protein domains and the percentage spliced in (PSI) of the affected exon (estimation of the percentage of *TTN* transcripts that incorporate the mutation) [38], but we did not use this information to modify the variant classification.

Confirmed CNVs were first compared with published literature and databases HGMD [31], ClinVar [32], DECIPHER [39], Database of Genomic Variants [40] and ClinGen [41]. If the CNV (identical after precise characterization) had been previously robustly classified as P/LP or B/LB, the classification was extrapolated to our case. Novel CNVs were classified as pathogenic variants if: 1) it was a deletion in/of a gene where loss of function is a known mechanism of patient's disease, 2) it was an intragenic in tandem duplication (not involving the last exon of the gene) in a gene where loss of function is a known mechanism of disease, or 3) it was a whole gene duplication in a gene for which triplosensitivity is known to cause patient's disease.

Statistical analysis

Categorical variables were compared using the chi-square test and two-sided p values < 0.05 were considered significant. Specifically, we compared the percentage of patients with rare variants observed after the screening for 25 genes with the percentage obtained when only the 5 main sarcomere genes were analyzed in the same cohort. We also compared the rate of P/LP variants found using these two different approaches. The same comparison was performed excluding *TTN* (set of 24 genes). Contingence tables were built to identify the number of patients with rare variants in *TTN* that also carried rare or P/LP variants in sarcomere genes. The role of the MAF hard filter on the final number and classification of genetic variants was analyzed comparing the proportions obtained using two different MAFs (< 0.002 vs. < 0.0001). We also used both MAF cut-offs to compare the detection rate of rare variants in *TTN* in patients with HCM and patients without structural heart disease. The statistical analysis was performed using R version 3.3.2.

Results

Genetic variants in main sarcomere genes

Overall, including both the Sanger sequencing and the NGS cohorts ($n = 387$ patients), we found 187 rare variants in the 5 principal sarcomere genes (*MYBPC3*, *MYH7*, *TNNI3*, *TNNT2* and *TPM1*) in 269 patients (69.5%) (S1 and S2 Tables). After applying the ACMG criteria of causality, 135 variants were classified as P/LP (72.2%), 41 (21.9%) were considered VUS, and only 11 (5.9%) were LB. No significant differences in the percentage of rare, P/LP or novel variants in the five main sarcomere genes were observed between the Sanger and the NGS cohorts (split data are shown in Table 1). The percentage of patients with at least one P/LP variant

Table 1. Rare variants (MAF <0.002) in the 5 most frequent sarcomere genes, 25 genes associated with or candidate for HCM and 24 genes (same panel excluding *TTM*).

	Main Sarcomere Genes			25 Gene Panel	24 Genes (excluding <i>TTM</i>)
	Pooled Data	Sanger cohort	NGS cohort	NGS cohort	NGS cohort
Patients n	387	84	303	303	303
Positive test n (%)	129 (33.3)	30 (35.7)	99 (32.7)	111 (36.6)	111 (36.6)
Positive test or novel candidates n (%)	130 (33.6)	30 (35.7)	100 (33.0)	121 (39.9)	115 (38.0)
Test with non-benign variants ⁽¹⁾ n (%)	163 (42.1)	35 (41.7)	128 (42.2)	220 (72.6)*	171 (56.4)*#
Inconclusive test n (%)	34 (8.8)	5 (6.0)	29 (9.6)	109 (36.0)*	60 (19.8)*#
Number of rare variants	187	39	148	398	235*#
Pathogenic n (%)	68 (36.4)	14 (35.9)	54 (36.5)	57 (14.3)*	57 (24.3)*#
Likely Pathogenic n (%)	67 (35.8)	16 (41.0)	51 (34.5)	60 (15.1)*	60 (25.5)#
VUS n (%)	41 (21.9)	6 (15.4)	35 (23.6)	243 (61.1)*	97 (41.3)*#
Benign/Likely Benign n (%)	11 (5.9)	3 (7.7)	8 (5.4)	38 (9.5)	21 (8.9)
Novel variants	48	10	38	110*	61#
Pathogenic/Likely Pathogenic n (%)	35 (72.9)	9 (90)	26 (68.4)	29 (26.4)*	29 (47.5)#
Novel Candidate variants	1 (2.1)	0	1 (2.6)	27 (24.5)*	7 (11.5)#
VUS (excluding candidate variants) n (%)	11 (22.9)	1 (10)	10 (26.3)	51 (46.4)*	24 (39.3)
Benign/Likely Benign n (%)	1 (2.1)	0	1 (2.6)	3 (2.7)	1 (1.6)

(1) All rare variants excluding benign and likely benign variants. n: number; NGS: next generation sequencing; VUS: variant of unknown significant.

*p<0.05 vs. analysis of 5 genes (*MYBPC3*, *MYH7*, *TNNI3*, *TNNT2* and *TPM1*) in the NGS cohort.

#p<0.05 vs. panel including 25 genes.

<https://doi.org/10.1371/journal.pone.0181465.t001>

(positive test) was 33.3%. We found 48 novel variants in these sarcomere genes. Among them, 35 were classified as P/LP (Table 2) and 1 as candidate novel variant.

The distribution of the 187 rare variants found in the 5 main sarcomere genes (pooled data from Sanger sequencing and NGS cohort, 387 patients) and their clinical classification is shown in Fig 1. We found 114 rare variants in *MYBPC3*, 48 in *MYH7*, 11 in *TNNT2*, 4 in *TNNI3* and 10 in *TPM1*. As expected, most P/LP variants were found in *MYBPC3* (64.2%) and *MYH7* (27%). In accordance with a well-known loss of function mechanism for *MYBPC3*, most P/LP variants in this gene (60.2%) were radical variants, whereas missense variants were the most frequent variants among the other genes. Cascade genetic screening allowed us to establish the penetrance of 40 P/LP variants: 17 with complete penetrance and 23 with incomplete penetrance. One variant was *de novo*.

We identified several sarcomere variants previously associated with other cardiomyopathies (S2 Table). Among P/LP variants, *MYH7*_p.R249G has been previously described in association with left ventricular non-compaction cardiomyopathy (LVNC) [42], and *MYH7*_p.T1019N in association with both dilated cardiomyopathy (DCM) and HCM [43, 44]. Additionally, cascade screening for the pathogenic variant *TNNT2*_p.E163del found in one HCM index case showed that 1 of the 3 genotype-positive relatives had HCM but the other 2 had a LVNC phenotype.

Rare genetic variants identified in the NGS cohort

The screening for 25 genes in the NGS cohort showed 401 rare variants, but 3 variants were not confirmed by Sanger sequencing (false discovery rate = 0.75%). Therefore, 398 confirmed variants were reported in 231 patients. In comparison with the screening for only the 5 main sarcomere genes, the percentage of patients with non-benign variants was significantly higher

Table 2. Novel pathogenic/likely pathogenic variants found in validated sarcomere genes.

Gene	cDNA	Aminoacid	Exon	Type	Probands
MYBPC3					
	c.323delC	p.108Lfs*51	3	Frameshift	1
	c.313dupG	p.A105Gfs*8	3	Frameshift	1
	c.572G>T	p.W191L	5	Missense	1
	c.1421_1424delAGTG	p.E474Vfs*13	16	Frameshift	1
	c.1471delG	p.V491Wfs*3	17	Frameshift	1
	c.2190delC	p.K731Rfs*23	23	Frameshift	2
	c.2329dupG	p.A777Gfs*56	24	Frameshift	2
	c.2591delT	p.F864Sfs*15	25	Frameshift	1
	c.2512G>T	p.E838*	25	Nonsense	1
	c.2724_2725delCTinsGCTGTA	p.Y908*	26	Nonsense	1
	c.2603-2A>G		26	Splice site	2
	c.2905+2T>C		27	Splice site	1
	c.3066dupC	p.N1023Qfs*28	29	Frameshift	2
	c.3190+5G>C		29	Intronic	1
	c.3182_3190+4delAGGTTGTTGGTGC		29	Long indel	1
	c.3020G>A	p.W1007*	29	Nonsense	2
	c.3190+2T>C		29	Splice site	1
	c.3328delA	p.M1110Wfs*79	30	Frameshift	3
	c.3620_3623dupGCCC	p.K1209Pfs*34	32	Frameshift	1
	c.3719T>A	p.I1240N	33	Missense	1
MYH7					
	c.530C>G	p.T177S	6	Missense	1
	c.920C>G	p.P307R	11	Missense	2
	c.1207C>G	p.R403G	13	Missense	1
	c.1580C>T	p.P527L	16	Missense	1
	c.2596T>C	p.S866P	22	Missense	1
TNNT2					
	c.311C>A	p.A104E	9	Missense	1
TNNI3					
	c.602T>C	p.M201T	8	Missense	1

<https://doi.org/10.1371/journal.pone.0181465.t002>

(72.6% vs. 42.2%, $p < 0.001$). Among them, 117 variants in 111 patients (36.6%) were classified as P/LP (Table 1). The proportion of patients with a positive test was not significantly different to the proportion found when screening for only the 5 main sarcomere genes (Table 1). The distribution of the 398 rare variants in the 25 genes in the NGS cohort and their clinical classification is shown in Fig 2. All genes except *CSRP3*, *LDB3*, *MYOZ2* and *PLN* showed at least one rare variant. The classification of the novel variants identified in this cohort is shown in Fig 3.

In 99 patients who did not have rare variants in the 5 main sarcomere genes the screening for minor and candidate genes identified 250 rare variants. Among them, 12 patients carried P/LP variants. Most of these P/LP variants were found in validated minor sarcomere genes (*ACTC1*, *MYL2* and *MYL3*) and genes related to metabolic diseases (*GLA* and *PRKAG2*) (Figs 2 and 4). The lack of definitive association with HCM implied that most rare variants found in candidate genes did not meet enough standardized ACMG criteria to be considered P/LP. However, at the individual variant level, we identified 2 unrelated HCM patients with the variant *TNNC1*_p.A8V, which has been previously reported in at least 7 additional unrelated HCM patients, is absent in control population (ExAC), and has functional studies supportive

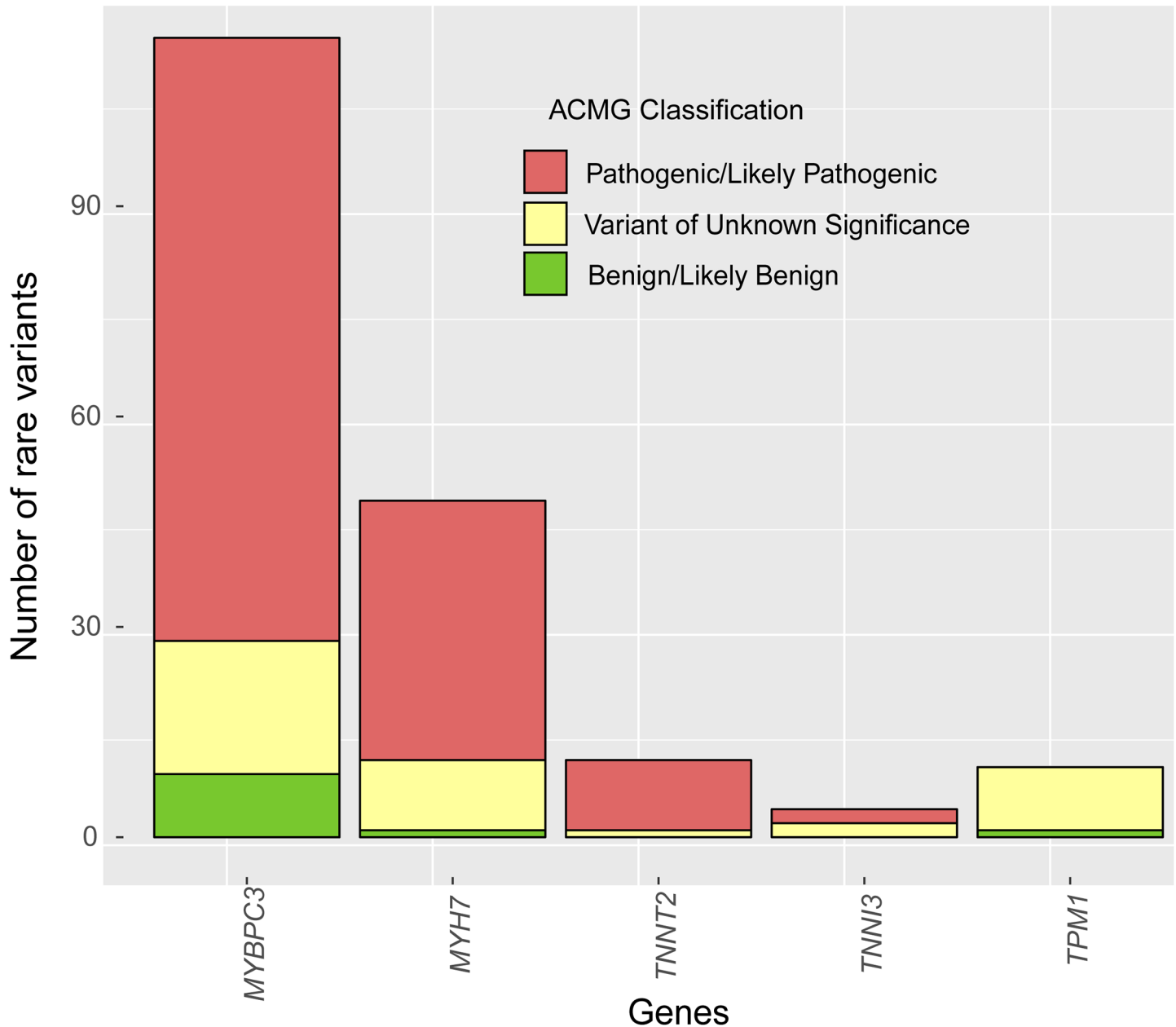


Fig 1. Classification of rare variants in MYBPC3, MYH7, TNNI3, TNNT2 and TPM1 (pooled data from Sanger sequencing and NGS cohorts).

<https://doi.org/10.1371/journal.pone.0181465.g001>

of a damaging effect. The findings of our cohort may help to increase the supportive evidence for enrichment of this infrequent variant in HCM cases.

The expanded genetic study identified 72 novel variants out of the 5 more frequent genes. On the other hand, the number of VUS drastically increased with the screening of 25 genes (23.6% vs. 61.1%, $p < 0.001$). Most of the additional VUS were *TTN* variants (97.9% of them missense). Accordingly, the number of inconclusive tests also increased (9.6% vs. 36.0%, $p < 0.001$) (Table 1).

Effect of MAF cut-off on the number and classification of rare variants. Using a ExAC MAF < 0.0001 we identified 308 variants in 205 patients (67.7%) from the NGS cohort. This

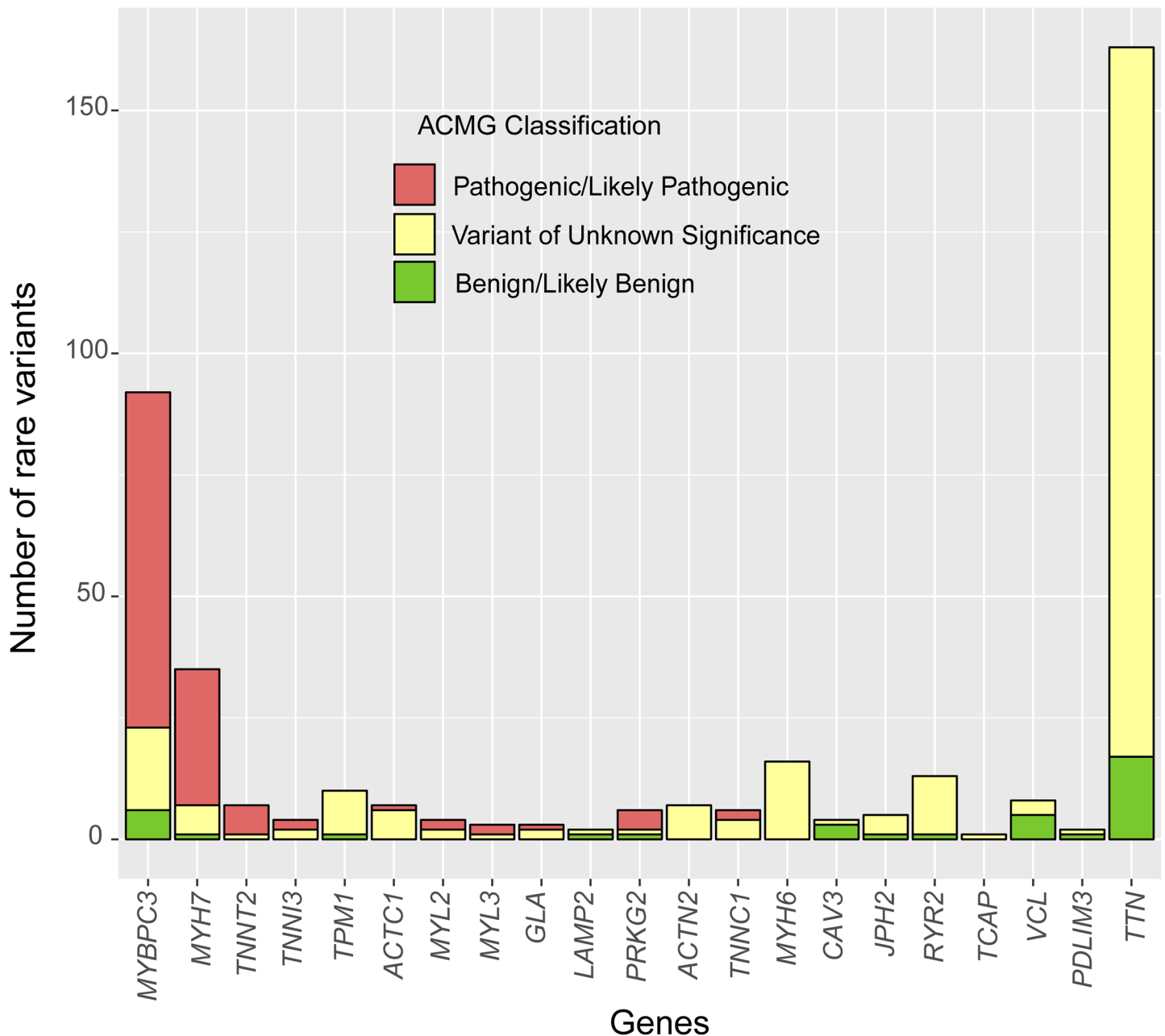


Fig 2. Classification of the rare variants found in the 25 genes screened in the NGS cohort.

<https://doi.org/10.1371/journal.pone.0181465.g002>

cut-off filtered out 28 of the 39 LB variants and 59 of the 241 VUS with $MAF < 0.002$. However, the rate of inconclusive studies did not significantly change (31.7% vs. 36.0%, $p = 0.26$). Three LP variants in 3 different patients were missed (*MYBPC3*_p.V771M in two patients and *TNNT2*_p.R278C in another one). Although there is enough supporting evidence to consider both variants as LP, published data suggest incomplete penetrance and a mild effect in isolation. Complete analysis with $MAF < 0.0001$ is provided in [S3 Table](#). The effect of the different MAF cut-offs on the number and distribution of variants according to the gene-level association with the disease is presented in [Fig 4](#).

Identification of multiple variants: Compound and double heterozygotes. Overall, considering only the 8 validated sarcomere genes and excluding LB variants, 14 patients carried two variants. In 6 cases we found 2 P/LP variants, and in 8 cases one P/LP variant in combination with one VUS. Six patients had two variants in *MYBPC3* (1 was a compound heterozygote, while in the remaining cases we cannot discard the possibility of both variants being located in the same allele), 1 patient had two variants in *MYH7*, and 7 patients were double-heterozygous for validated sarcomere genes. In the NGS cohort, 105 patients had more than one non-benign variant. After excluding *TTN* from analysis, 38 patients had multiple non-benign variants.

Detection rate of *TTN* variants in HCM and comparison with patients without structural heart disease

***TTN* variant classification and filtering.** We found 163 rare variants in *TTN* in 117 patients, 49 novel and 114 previously described in clinical and/or population databases. After a

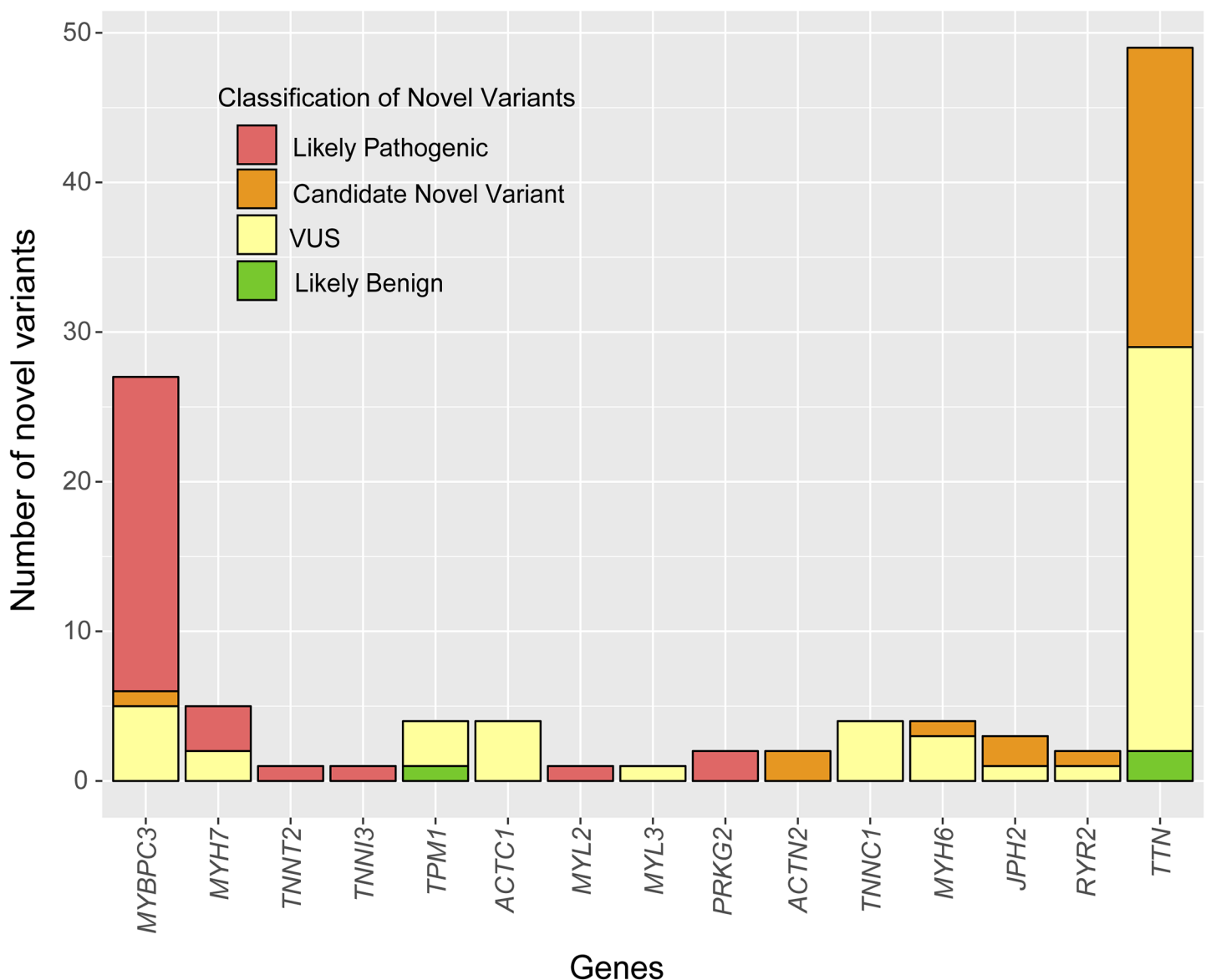
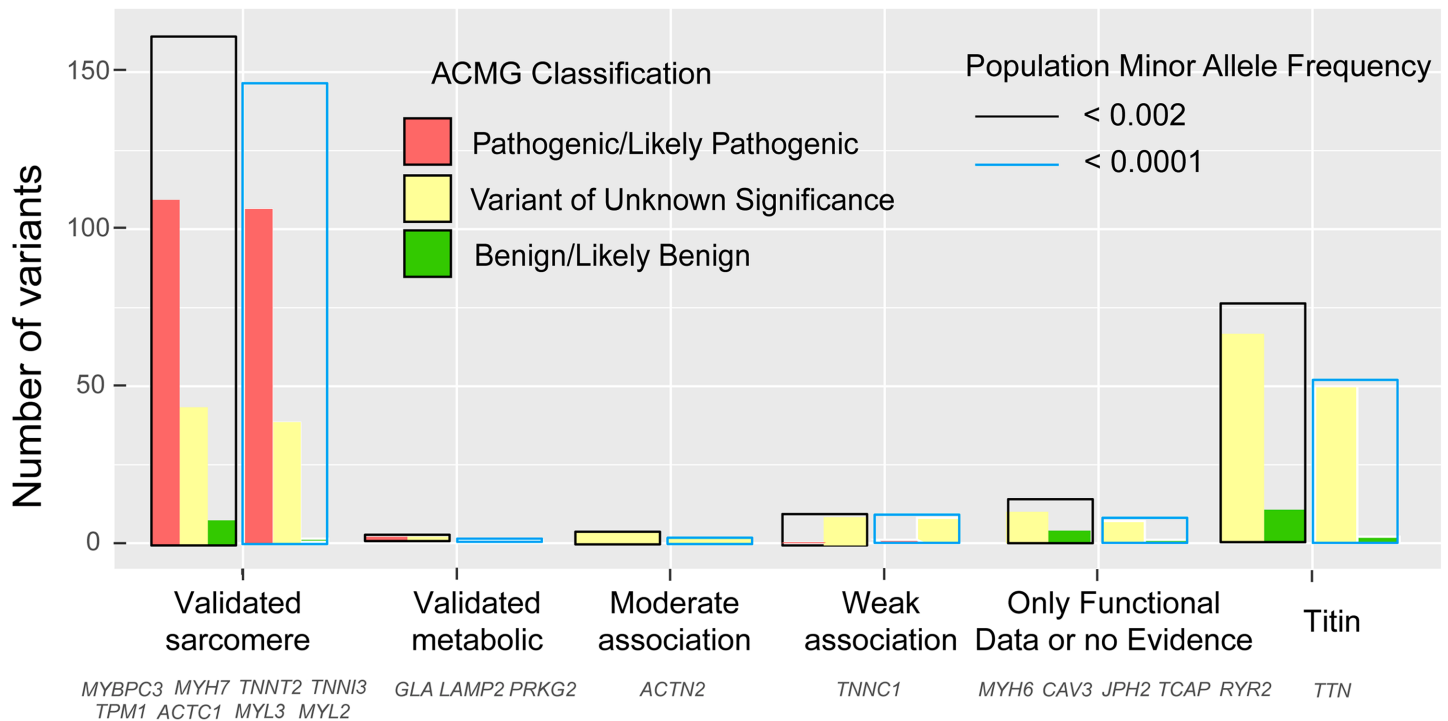


Fig 3. Classification of the novel variants identified in the NGS cohort.

<https://doi.org/10.1371/journal.pone.0181465.g003>



Classification of Genes based on evidence for association with HCM

Fig 4. Distribution of rare variants according to gene-level supporting evidence, ACMG clinical classification and minor allele frequency filtering.

<https://doi.org/10.1371/journal.pone.0181465.g004>

careful revision of published data and family studies, 17 variants were classified as LB (15 previously reported and 2 novel variants). In particular, segregation studies allowed us to classify the variant as LB in 8 families. The remaining 146 variants in 108 patients were classified as VUS (28 patients had 2 additional VUS, and 5 had 3 VUS). Fifty-eight patients also carried variants in the main sarcomere genes, classified as P/LP in 40 cases. In an attempt to select the VUS in *TTN* with a higher potential clinical significance we focused on 20 novel candidate variants found in 18 patients in which three *in silico* tools consistently predicted a deleterious effect (Table 3). Fourteen of these variants were found in 13 patients who also carried variants in sarcomere genes (9 patients had a P/LP variant and 4 a VUS). Only 6 of these filtered *TTN* variants in 5 patients were found in cases without variants in the main sarcomere genes.

Comparison with patients without structural heart disease. The number of HCM patients with non-synonymous rare variants in *TTN* (38.6% with MAF <0.002 and 29% with MAF <0.0001) was not significantly different to the group of patients without evidence of structural heart disease (39.3% with MAF <0.002, and 26% with MAF <0.0001) ($p > 0.3$ for both case-control comparisons) (Table 4). The detection rate of VUS and novel variants was also similar (Table 4). A detailed description of the *TTN* variants analyzed in this cohort is provided in S4 Table.

With respect to the total number of *TTN* variants, filtering with MAF <0.002, the percentage of LB variants was higher in the group of patients without structural heart disease, and the relative percentage of novel variants was lower. However, considering only variants with MAF <0.0001, the percentage of novel variants was not significantly different. The location of the variants in the protein and the percentage of variants located in constitutive exons (PSI = 100) were similar in both groups when using MAF <0.0001 (Table 4). Applying the same step-wise

Table 3. Novel variants of unknown significance in *TTN* gene that are deleterious according to multiple in silico predictors.

Patient ID	cDNA	Aminoacid	Exon	PSI	Domain	Variants in sarcomere genes in the same patient ⁽¹⁾
44	c.78293C>T	p.T26098I	275	100	A-band	Pathogenic: <i>MYBPC3</i> c.1505G>A p.R502Q
52	c.62924A>T	p.D20975V	275	100	A-band	Pathogenic: <i>MYBPC3</i> c.1624G>C p.E542Q
53	c.81646G>T	p.V27216F	283	100	A-band	Likely pathogenic: <i>MYBPC3</i> c.2724_2725delCTinsGCTGTA p.Y908*
60	c.89236A>G	p.K29746E	297	100	A-band	VUS Novel: <i>MYBPC3</i> c.631G>A p.D211N
66	c.85081G>A	p.A28361T	288	100	A-band	Pathogenic: <i>MYBPC3</i> c.162delG K54Nfs*13
66	c.93829T>C	p.Y31277H	307	100	M-band	Pathogenic: <i>MYBPC3</i> c.162delG K54Nfs*13
71	c.78293C>T	p.T26098I	275	100	A-band	Pathogenic: <i>ACTC1</i> c.889G>T p.A297S
95	c.8920A>G	p.M2974V	38	100	I-band	Likely Pathogenic: <i>MYH7</i> c.2608C>T p.R870C
98	c.758C>T	p.T253I	6	100	Z-disc	None
98	c.70579C>G	p.P23527A	275	100	A-band	None
101	c.51661G>A	p.D17221N	250	100	A-band	VUS Novel: <i>TNNC1</i> c.121C>A p.L41M
104	c.40364C>T	p.S13455F	205	100	A-band	None
108	c.46801C>A	p.P15601T	231	100	A-band	VUS Novel: <i>TPM1</i> c.632C>T p.A211V
112	c.90118A>G	p.R30040G	300	100	A-band	None
146	c.89159C>G	p.P29720R	296	100	A-band	None
170	c.16069C>T	p.P5357S	65	6	I-band	Likely pathogenic: <i>TNNT2</i> c.857G>A p.R286H
189	c.72563G>C	p.R24188T	275	100	A-band	Likely pathogenic: <i>MYL3</i> c.427G>A p.E143K
217	c.72098G>C	p.G24033A	275	100	A-band	Pathogenic: <i>MYBPC3</i> c.2308G>A p.D770N
245	c.37807G>C	p.G12603R	195	100	I-band	VUS: <i>TNNI3</i> c.304G>A p.A102T
260	c.47179C>T	p.P15727S	232	100	A-band	None

PSI: percent of splice in.

In bold: patients with 2 different novel variants in *TTN*.

⁽¹⁾ additional information available in [S2 Table](#).

<https://doi.org/10.1371/journal.pone.0181465.t003>

algorithm in the non-structural cohort we found 17 novel VUS consistently predicted as deleterious, 3 of them non-sense variants located in the A-band, I-band and M-band, respectively ([S4 Table](#)). The detection rate of patients with this type of selected novel variants was not significantly different between HCM patients and patients without structural heart disease.

Identification, characterization and classification of CNVs

Screening for CNVs in our NGS cohort revealed that 4 out of the 303 patients had a validated CNV in one of the 25 genes analyzed (1.3%). Twelve additional signals were detected but they were not validated by MLPA or qPCR (false discovery rate = 75%). Among confirmed CNVs, 2 patients had deletions involving *MYBPC3* gene, and 2 patients had a deletion of the entire coding region of the *PLN* gene. According to our criteria for interpretation of CNVs, all of them were considered pathogenic variants.

CNVs in *MYBPC3*. One case (P168) had a deletion of the entire exon 27 ([Fig 5. Panel A](#)), and the other one (P259) had a deletion spanning from exon 4 to exon 12 ([Fig 5. Panel B](#)). Both deletions were confirmed by MLPA. No split-read data were available for the establishment of the rearrangement breakpoints of these cases, but both CNVs could be precisely characterized by Sanger sequencing: c.2737+148_2905+40del727insG for P168 and c.406+69_1091-1154del5654 for P259. None of the deletions has been previously described.

The case P168 also carried two previously reported variants in *MYBPC3* (p.V771M –classified as LP– and p.A522T –classified as LB–), and a VUS in *TTN*. The case P259 also carried two missense variants in *TTN*: p.E2055K (classified as LB) and p.R25906C (classified as VUS).

Table 4. Detection rate and classification of variants in *TTN* in patients with hypertrophic cardiomyopathy and patients without structural heart disease.

	HCM (N = 303)		Non-structural (N = 427)	
	MAF <0.002	MAF <0.0001	MAF <0.002	MAF <0.0001
Patients				
with Rare Variants in <i>TTN</i>	117 (38.6)	88 (29.0)	168 (39.3)	111 (26.0)
with VUS in <i>TTN</i>	108 (35.6)	86 (28.4)	146 (34.2)	111 (26.0)
with Novel Variants in <i>TTN</i>	42 (13.9)	42(13.9)	50 (11.7)	50 (11.7)
<i>TTN</i> variants				
Likely Benign	163	109	274	151
VUS	17 (10.4)	3 (2.8)	63 (23.0)*	0 (0.0)
Novel	146 (89.6)	106 (97.2)	211 (77.0)*	151 (100)*
Novel VUS deleterious <i>in silico</i>	49 (30.1)	49 (45.0)	58 (21.2)*	58 (38.4)
Novel VUS deleterious <i>in silico</i>	20 (12.3)	20 (18.3)	17 (6.2)*	17 (11.2)
Truncating variants ⁽¹⁾	0 (0.0)	0 (0.0)	3 (1.1)	3 (2.0)
In consitutive Exons (PSI = 100)	123 (75.5)	81 (74.3)	230 (83.9)*	124 (82.1)
Location in protein				
A band	97 (59.5)	66 (60.6)	190 (69.3)*	98 (64.9)
I band	45 (27.6)	32 (29.4)	64 (23.4)	42 (27.8)
M band	13 (8)	8 (7.3)	17 (6.2)	10 (6.6)
Z disk	8 (4.9)	3 (2.8)	3 (1.1)*	1 (0.7)

HCM: hypertrophic cardiomyopathy; MAF: minor allele frequency in ExAC; PSI: percent of splice in; VUS: Variant of unknown significance.

⁽¹⁾ Truncating: Nonsense, frameshift or canonical splicing.

* p<0.05 non-structural vs. HCM using the same MAF filter.

<https://doi.org/10.1371/journal.pone.0181465.t004>

Cascade genetic testing in this family showed that the deletion from exon 4 to exon 12 cosegregated with the disease in one affected relative; this patient also carried the variant *TTN*_p.R25906C but not the variant *TTN*_p.E2055K (S2 Table).

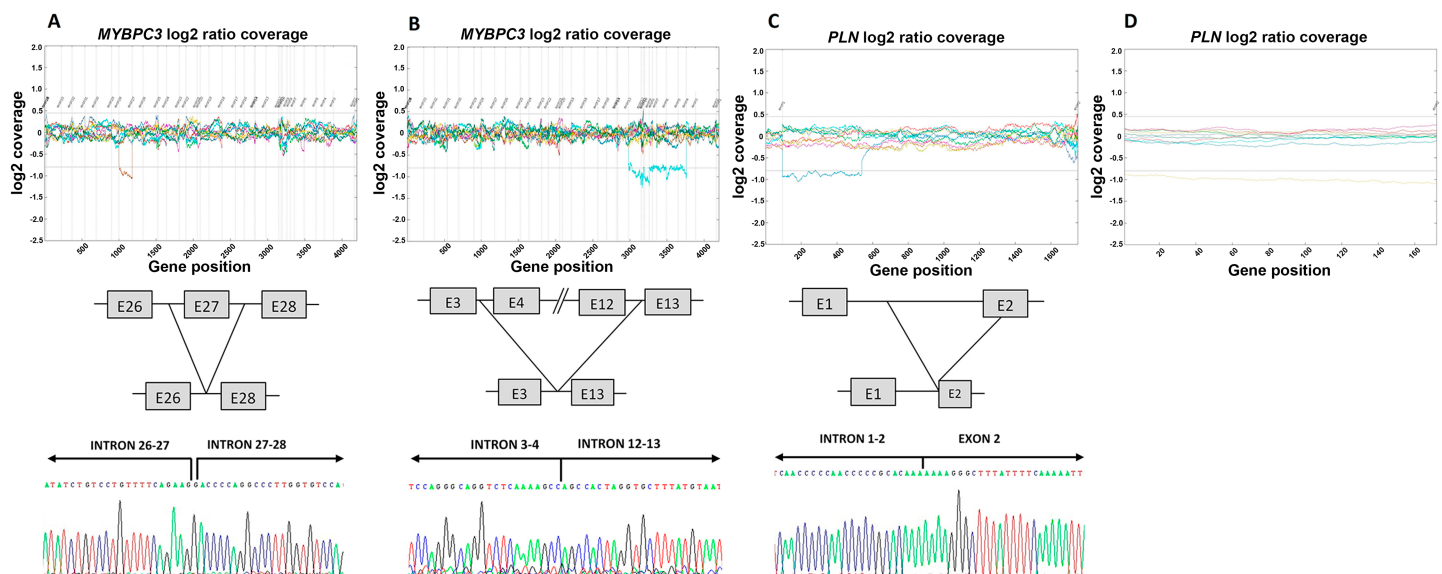


Fig 5. Cases with confirmed CNVs. NGS results, schematic representation of the breakpoints and precise characterization by Sanger sequencing of (A) the deletion of exon 27 of *MYBPC3* (P168, brown sample in the graph), (B) the deletion spanning from exon 4 to exon 12 of *MYBPC3* (P259, turquoise sample in the graph), and (C) the well-characterized *PLN* deletion (blue sample in the graph). (D) NGS results are shown for the non-characterized *PLN* deletion (orange sample in the graph).

<https://doi.org/10.1371/journal.pone.0181465.g005>

CNVs in *PLN*. In both cases, the deletions involved the entire coding region of the *PLN* gene and were confirmed by qPCR using the pair of primers 5' CTCAACAAGCAGTCAAAA GC3' and 5' GCATCACGATGATACAGATCAGC3'. None of the patients had any uncommon SNP or indel. The first patient had a deletion of 7936bp, involving a portion of the first intron and the second exon of the gene (Fig 5. Panel C). The deletion was identified with the 55-gene panel, which included the UTR regions for this gene. The breakpoints could be identified using NGS split-read data, and were confirmed by Sanger sequencing. The precise description of the rearrangement was c.1-7587_159+190del7936. Such rearrangement has not been previously described. The second deletion of *PLN* was detected with the 78-gene panel, which does not include the UTR regions (Fig 5. Panel D). The rearrangement could not be further characterized using neither split-read data nor Sanger sequencing, but most probably is different from the one diagnosed in the other patient, as widening the coding region coordinates 500bp upstream and downstream the deletion signal was still present, and Sanger sequencing with the same pair of primers did not amplify any fragment. Overlapping rearrangements have not been previously described in HGMD, but there are several overlapping deletions described in DGV [40], DECIPHER [39], ClinGen [41] and ClinVar [32]. The variant found in DGV is a deletion of 58Kb involving the genes *CEP85L* and *PLN*, and the variants found in the other databases are larger and have been found in patients with congenital abnormalities.

Discussion

In this study, we report the results of the genetic screening of 387 unrelated Spanish patients clinically diagnosed with HCM. Using NGS, we focus on the additional diagnostic value of screening for minor and candidate genes for HCM, and expose a comprehensive study of CNVs. Our data show that screening for these genes and CNVs in HCM patients identifies the genetic cause of the disease in a small number of cases, but this approach does not increase the global detection rate. The screening for variants in *TTN* in HCM patients shows a high number of VUS and increases the rate of inconclusive test. In an independent cohort without structural heart disease, we have found a number and classification of rare variants in *TTN* similar to that found in HCM patients; adding evidence against the role of this gene in HCM.

Detection rate and clinical classification of rare variants

In the present study, the analysis of the most frequent sarcomere genes (*MYBPC3*, *MYH7*, *TNNI3*, *TNNT2* and *TPM1*) in patients with HCM identified a potentially relevant variant in 42.2% of the patients. Applying recommended criteria for clinical classification, P/LP variants were identified in only one third of the patients. The screening for additional sarcomere genes and other known and putative HCM genes using a 25-gene NGS panel showed potentially relevant variants in about two thirds of the patients, but identified an additional P/LP mutation in a small number of patients, without significantly increasing the percentage of patients with a positive test. Although some initial studies using Sanger sequencing reported a detection rate of pathogenic variants of 63% [7, 20], most studies have reported rates below 50% [6, 8, 9, 11, 13–18]. The detection rate in recent NGS studies including different additional genes range from 32% to 78.9% [4, 12, 17, 19, 44–52]. This huge variation is most likely due to selection bias in these studies, differences in the clinical characteristics of the patients included and, importantly, the different criteria applied for the classification of genetic variants. As extensively recognized, NGS offers a high reliability [44, 47], so the sequencing process itself does not justify the differences in the reported yields.

Interestingly, the percentage of patients with P/LP variants found in our study is in agreement with the largest published NGS series in patients with HCM, which includes more than

2900 unrelated HCM patients (with genetic screening involving from 10 to 51 known or putative HCM genes) [4]. The criteria for the assessment of genetic variants used in their study was similar to the ACMG classification followed in the present study. It has been recognized that this strict classification may result in a larger proportion of variants being considered VUS [34], but the main purpose of this tool is to guide clinical decisions, which must be always based on strong supporting evidence. Nevertheless, we recognize that these restrictive clinical classifications do not fulfill the requirements of a research study to identify new disease-causing variants, especially when genes with a non-definitive association with the disease are included. For this reason, an additional effort to weight the evidence of pathogenicity during the assessment of novel candidate variants identified by NGS should be attempted.

Genetic spectrum of the disease. In agreement with most series [4, 7–9, 14, 15, 17], our study shows that *MYBPC3* is the gene with a higher proportion of P/LP variants, followed by *MYH7*. While *MYH7* variants are almost exclusively missense (in our cohort all of them), *MYBPC3* is characterized by a significant incidence of radical variants [3, 4]. Overall, considering only the 8 validated sarcomere genes, the 3.6% of our patients carried two non-benign variants, which is consistent with published data [17, 20]. The screening for additional candidate genes using NGS increased the proportion of carriers of multiple non-benign variants to 34.7%, mostly due to the existence of rare missense *TTN* variants. It has been demonstrated that the presence of multiple pathogenic variants in the 8 validated sarcomere genes may confer a more severe form of disease with a higher incidence of adverse outcomes including heart failure and sudden death [53]. However, the clinical significance of the presence of multiple variants in the remaining genes is unknown and this information should not currently be used for prognostic purposes.

Variants of unknown significance and role of *TTN*

While the screening for 25 genes provides a definitive diagnostic in particular cases without P/LP variants in the main sarcomere genes, the proportion of cases with VUS increases exponentially. These VUS represent nowadays a major clinical challenge, as proper genetic diagnosis and genetic counseling cannot be provided. In the assessment of VUS, the study of large pedigrees for segregation analyses and *in vitro* assays may provide useful information, but these studies are not possible or feasible in most cases.

Current available data are not enough to support pathogenicity of novel variants in candidate genes, but the absence in controls and the existence of consistent computational data supporting a deleterious effect should be taken into consideration to undertake segregation and/or functional studies. It is important to underline that variants in candidate genes should not be used for clinical purposes, such as genetic counseling or cascade screening testing. However, reporting and carefully addressing them are necessary steps for the improvement of genetic diagnosis in HCM. Additionally, scientific literature and databases should be periodically searched for new information to reclassify these variants.

In the present study, the drastic increase of VUS in the NGS cohort is mainly due to the analysis of *TTN*, which is the gene with the largest coding sequence in the human genome. Whereas the pathogenic role of truncating variants in *TTN* has been demonstrated for DCM [54], the frequency of these radical variants in patients with HCM is similar to that found in control populations [54], and the pathogenic role of *TTN* missense variants is unknown [55]. In the present study we show that the number and classification of missense variants in *TTN* is not significantly different in patients with HCM and patients without structural heart disease. Even the rate of occurrence of novel variants that are consistently predicted as deleterious by *in silico* tools is not significantly different between both cohorts. Moreover, whereas a potential modifier role of selected missense *TTN* variants cannot be definitively ruled out, the finding of

a high proportion of rare *TTN* variants in patients with variants in sarcomere genes increases the concerns about their actual pathogenic role. Altogether, these data may advise against the inclusion of this gene in clinical HCM panels.

Role of CNVs in HCM

To date, the role of CNVs in patients with HCM is a relatively unexplored field. The first CNV in a patient with HCM was reported in 1992 and consisted of a 2.4Kb deletion in *MYH7*, which was identified by Southern blotting, analyzing restriction fragment length polymorphisms [56]. A second patient with HCM and two CNVs in *MYBPC3* was reported in 2009 [57]. Since then, few series studying CNVs in HCM patients have been published, and most studies have evaluated only 1 or 2 genes [58–63]. The search for single-exon deletions by long-range PCR in *MYH7* in a cohort of 150 patients did not identify CNVs [58]. Three studies that performed MLPA of *MYBPC3* (and in some cases *TNNT2*) reported a detection rate for CNVs of 0% (0/108) [59], 1% (1/100) [60] and 1.4% (1/72) [61]. Interestingly, the CNV identified in the last two studies was an identical *MYBPC3* large deletion involving several exons (starting in the intron 27 and ending 485 bp after the *MYBPC3* stop codon). Three other cases with CNVs in *MYBPC3* have been reported, but detailed information of each case is not available [62, 63].

The first comprehensive study that searched for CNVs in multiple genes in a large group of HCM patients was published in 2015 by Lopes et al. [23]. They analyzed 19 HCM-related or candidate genes by NGS in 505 patients and detected 4 CNVs (0.8%): 1 deletion in *MYBPC3*, 1 deletion in *PDLIM3*, 1 duplication of the entire *TNNT2* gene, and 1 duplication in *LMNA*. Deletions were considered LP variants, while duplications were considered VUS. Recently, Ceyhan-Birsoy et al. screened 708 HCM patients for CNVs using a NGS panel including 18 HCM-related (or putative) genes or 46 genes covering the full spectrum of cardiomyopathies, and detected CNVs in 4 of them (0.56%): a duplication in *MYOZ2*; a deletion in *MYBPC3*; a whole gene duplication of *NEXN*; and a whole gene duplication of *GLA*, *LAMP2*, *EMD* and *TAZ* (patient with trisomy X) [24]. Only the deletion of *MYBPC3* was classified as pathogenic. In the present study, to further elucidate the role of CNVs in HCM we screened 303 HCM patients for CNVs in 25 genes associated with or candidate for HCM. Among them, we detected 4 CNVs (1.3% of our patients). Two CNVs were novel deletions in *MYBPC3*, one of them involving exon 27 and the other one ranging from exon 4 to exon 12. Both CNVs were classified as pathogenic variants, as radical variants in *MYBPC3* are a well-known cause of HCM [4]. Interestingly, the first patient also harbored one LP variant and one LB variant in *MYBPC3* (p.V771M and p.A522T, respectively), and a VUS in *TTN*. As no family members were available, we were not able to determinate if the *MYBPC3* variants were located in the same allele. To the best of our knowledge, this kind of complex genotype in *MYBPC3* has not been reported before. A previous study in 113 patients designed to search for CNVs in *MYBPC3* in HCM patients carrying one pathogenic point variant did not identify any large rearrangement [61]. Our study demonstrates that even in patients with a LP variant in a main sarcomere gene, the screening for CNVs may add valuable information.

The other two rearrangements identified in our study were deletions of the entire coding region of *PLN* gene. Such *PLN* deletions were classified as pathogenic variants, because the patients only have a single functional copy of the gene, and reduction of the expression of *PLN* (due to nonsense and promoter pathogenic variants) has been previously associated with the development of HCM [64, 65]. These patients did not harbor any other P/LP variant that could explain the HCM phenotype.

The CNV prevalence in our cohort (1.3%) is not significantly different to that reported by Lopes et al. [23] (0.8%, $p = 0.4630$) and Ceyhan-Birsoy et al. (0.56%, $p = 0.2144$) [24]. These

data suggest that large rearrangements explain a small number of cases that do not carry SNVs and indels, and in selected cases can be part of complex genotypes in combination with variants in sarcomere genes.

Limitations

Only the protein-coding and flanking intronic regions of known or putative HCM genes were analyzed, and some HCM cases may be explained by pathogenic variants in non-coding regions or other genes. In fact, during the enrollment period of this study new genes that were not included in the cardiomyopathy panels used, such as *FLNC* [66] or *FHL1* [67], have been associated with HCM. Additionally, the inclusion in the HCM panels of genes related to other inherited diseases that involve left ventricular hypertrophy (i.e. Pompe disease, amyloidosis, mitochondrial cardiomyopathies or rasopathies) could also increase the diagnosis of some unexplained cases. The two steps process used for discovery and validation of CNVs showed a high false discovery rate, but this finding might be at least partially related to the inclusion of signals with low quality score in the validation step, in an attempt to minimize the rate of false negatives in a clinical scenario. Even using this low threshold, the number of CNV signals identified is small and their validation does not significantly impact the total cost of genetic testing in the whole sample.

Conclusion

Only a small percentage of HCM cases without point mutations in the 5 principal sarcomere genes are explained by pathogenic variants in minor or candidate genes for HCM or CNVs, but their identification is of major clinical relevance and can be easily performed by widely available NGS techniques. Screening for *TTN* in HCM patients drastically increases the number of inconclusive tests, and provides a rate of rare variants similar to that found in patients without structural heart disease, suggesting that this gene should not be analyzed for clinical purposes in HCM patients.

Supporting information

S1 Table. Isoforms analysed of the 25 known or candidate HCM genes included in the custom NGS panels.

(XLSX)

S2 Table. Genetic variants with ExAC MAF <0.002 identified in 387 consecutive unrelated Spanish patients with hypertrophic cardiomyopathy.

(XLSX)

S3 Table. Rare variants (MAF <0.0001) in the 5 most frequent sarcomere genes, 25 genes associated with or candidates for HCM and 24 genes (same panel excluding *TTN*).

(XLSX)

S4 Table. Nonsynonymous variants in *TTN* with ExAC MAF <0.002 identified in 427 consecutive unrelated Spanish patients without structural heart disease.

(XLSX)

Acknowledgments

The authors acknowledge Instituto de Salud Carlos III (ISCIII). The CIBERCV is an initiative of the ISCIII, Spanish Ministry of Economy and Competitiveness.

Author Contributions

Conceptualization: Irene Mademont-Soler, Jesus Mates, Raquel Yotti, Catarina Allegue, Oscar Campuzano, Ramon Brugada.

Data curation: Jesus Mates, Anna Iglesias, Bernat del Olmo.

Formal analysis: Irene Mademont-Soler, Jesus Mates, Raquel Yotti, Bernat del Olmo.

Funding acquisition: Raquel Yotti, Oscar Campuzano, Francisco Fernandez-Aviles, Ramon Brugada.

Investigation: Irene Mademont-Soler, Jesus Mates, Alexandra Pérez-Serra, Monica Coll, Helena Riuró, Ferran Picó.

Methodology: Irene Mademont-Soler, Jesus Mates, Bernat del Olmo.

Project administration: Irene Mademont-Soler, Jesus Mates, Oscar Campuzano, Ramon Brugada.

Resources: Jesus Mates, Raquel Yotti, Maria Angeles Espinosa, Ana Isabel Fernandez-Avila, Irene Méndez, Sofía Cuenca, Carles Ferrer-Costa, Patricia Álvarez, Sergio Castillo, Pablo Garcia-Pavia, Esther Gonzalez-Lopez, Laura Padron-Barthe, Aranzazu Díaz de Bustamante, María Teresa Darnaude, José Ignacio González-Hevia, Josep Brugada, Francisco Fernandez-Aviles, Ramon Brugada.

Software: Jesus Mates, Bernat del Olmo.

Supervision: Raquel Yotti, Francisco Fernandez-Aviles, Ramon Brugada.

Validation: Irene Mademont-Soler, Jesus Mates, Alexandra Pérez-Serra, Monica Coll, Helena Riuró.

Visualization: Irene Mademont-Soler, Jesus Mates, Raquel Yotti.

Writing – original draft: Irene Mademont-Soler, Jesus Mates, Raquel Yotti.

Writing – review & editing: Irene Mademont-Soler, Jesus Mates, Raquel Yotti, Oscar Campuzano, Ramon Brugada.

References

1. Maron BJ, Gardin JM, Flack JM, Gidding SS, Kurosaki TT, Bild DE. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation*. 1995; 92(4):785–9. PMID: [7641357](https://pubmed.ncbi.nlm.nih.gov/7641357/)
2. Cobo-Marcos M, Cuenca S, Gamez Martinez JM, Bornstein B, Ripoll Vera T, Garcia-Pavia P. Usefulness of genetic testing for hypertrophic cardiomyopathy in real-world practice. *Rev Esp Cardiol (Engl Ed)*. 2013; 66(9):746–7.
3. Das KJ, Ingles J, Bagnall RD, Semsarian C. Determining pathogenicity of genetic variants in hypertrophic cardiomyopathy: importance of periodic reassessment. *Genet Med*. 2014; 16(4):286–93. <https://doi.org/10.1038/gim.2013.138> PMID: [24113344](https://pubmed.ncbi.nlm.nih.gov/24113344/)
4. Alfares AA, Kelly MA, McDermott G, Funke BH, Lebo MS, Baxter SB, et al. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. *Genet Med*. 2015; 17(11):880–8. <https://doi.org/10.1038/gim.2014.205> PMID: [25611685](https://pubmed.ncbi.nlm.nih.gov/25611685/)
5. Authors/Task Force m, Elliott PM, Anastakis A, Borger MA, Borggrefe M, Cecchi F, et al. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *Eur Heart J*. 2014; 35(39):2733–79. <https://doi.org/10.1093/eurheartj/ehu284> PMID: [25173338](https://pubmed.ncbi.nlm.nih.gov/25173338/)

6. Van Driest SL, Ommen SR, Tajik AJ, Gersh BJ, Ackerman MJ. Sarcomeric genotyping in hypertrophic cardiomyopathy. *Mayo Clin Proc.* 2005; 80(4):463–9. [https://doi.org/10.1016/S0025-6196\(11\)63196-0](https://doi.org/10.1016/S0025-6196(11)63196-0) PMID: 15819282
7. Olivetto I, Girolami F, Ackerman MJ, Nistri S, Bos JM, Zachara E, et al. Myofilament protein gene mutation screening and outcome of patients with hypertrophic cardiomyopathy. *Mayo Clin Proc.* 2008; 83(6):630–8. <https://doi.org/10.4065/83.6.630> PMID: 18533079
8. Millat G, Bouvagnet P, Chevalier P, Dauphin C, Jouk PS, Da Costa A, et al. Prevalence and spectrum of mutations in a cohort of 192 unrelated patients with hypertrophic cardiomyopathy. *Eur J Med Genet.* 2010; 53(5):261–7. <https://doi.org/10.1016/j.ejmg.2010.07.007> PMID: 20624503
9. Waldmuller S, Erdmann J, Binner P, Gelbrich G, Pankuweit S, Geier C, et al. Novel correlations between the genotype and the phenotype of hypertrophic and dilated cardiomyopathy: results from the German Competence Network Heart Failure. *Eur J Heart Fail.* 2011; 13(11):1185–92. <https://doi.org/10.1093/eurjhf/hfr074> PMID: 21750094
10. Garcia-Pavia P, Vazquez ME, Segovia J, Salas C, Avellana P, Gomez-Bueno M, et al. Genetic basis of end-stage hypertrophic cardiomyopathy. *Eur J Heart Fail.* 2011; 13(11):1193–201. <https://doi.org/10.1093/eurjhf/hfr110> PMID: 21896538
11. Gruner C, Ivanov J, Care M, Williams L, Moravsky G, Yang H, et al. Toronto hypertrophic cardiomyopathy genotype score for prediction of a positive genotype in hypertrophic cardiomyopathy. *Circ Cardiovasc Genet.* 2013; 6(1):19–26. <https://doi.org/10.1161/CIRCGENETICS.112.963363> PMID: 23239831
12. Lopes LR, Zekavati A, Syrris P, Hubank M, Giambartolomei C, Dalageorgou C, et al. Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *J Med Genet.* 2013; 50(4):228–39. <https://doi.org/10.1136/jmedgenet-2012-101270> PMID: 23396983
13. Marsiglia JD, Credidio FL, de Oliveira TG, Reis RF, Antunes Mde O, de Araujo AQ, et al. Screening of MYH7, MYBPC3, and TNNT2 genes in Brazilian patients with hypertrophic cardiomyopathy. *Am Heart J.* 2013; 166(4):775–82. <https://doi.org/10.1016/j.ahj.2013.07.029> PMID: 24093860
14. Kassem H, Azer RS, Saber-Ayad M, Moharem-Elgamal S, Magdy G, Elguindy A, et al. Early results of sarcomeric gene screening from the Egyptian National BA-HCM Program. *J Cardiovasc Transl Res.* 2013; 6(1):65–80. <https://doi.org/10.1007/s12265-012-9425-0> PMID: 23233322
15. Reguero JR, Gomez J, Martin M, Florez JP, Moris C, Iglesias S, et al. The G263X MYBPC3 mutation is a common and low-penetrant mutation for hypertrophic cardiomyopathy in the region of Asturias (Northern Spain). *Int J Cardiol.* 2013; 168(4):4555–6. <https://doi.org/10.1016/j.ijcard.2013.06.085> PMID: 23870641
16. Kapplinger JD, Landstrom AP, Bos JM, Salisbury BA, Callis TE, Ackerman MJ. Distinguishing hypertrophic cardiomyopathy-associated mutations from background genetic noise. *J Cardiovasc Transl Res.* 2014; 7(3):347–61. <https://doi.org/10.1007/s12265-014-9542-z> PMID: 24510615
17. Li Z, Huang J, Zhao J, Chen C, Wang H, Ding H, et al. Rapid molecular genetic diagnosis of hypertrophic cardiomyopathy by semiconductor sequencing. *J Transl Med.* 2014; 12:173. <https://doi.org/10.1186/1479-5876-12-173> PMID: 24938736
18. Berge KE, Leren TP. Genetics of hypertrophic cardiomyopathy in Norway. *Clin Genet.* 2014; 86(4):355–60. <https://doi.org/10.1111/cge.12286> PMID: 24111713
19. Lopes LR, Syrris P, Guttman OP, O'Mahony C, Tang HC, Dalageorgou C, et al. Novel genotype-phenotype associations demonstrated by high-throughput sequencing in patients with hypertrophic cardiomyopathy. *Heart.* 2015; 101(4):294–301. <https://doi.org/10.1136/heartjnl-2014-306387> PMID: 25351510
20. Richard P, Charron P, Carrier L, Ledeuil C, Cheav T, Pichereau C, et al. Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy. *Circulation.* 2003; 107(17):2227–32. <https://doi.org/10.1161/01.CIR.0000066323.15244.54> PMID: 12707239
21. Veselka J, Anavekar NS, Charron P. Hypertrophic obstructive cardiomyopathy. *Lancet.* 2016. Epub 2016/12/04.
22. Walsh R, Buchan R, Wilk A, John S, Felkin LE, Thomson KL, et al. Defining the genetic architecture of hypertrophic cardiomyopathy: re-evaluating the role of non-sarcomeric genes. *Eur Heart J.* 2017. Epub 2017/01/14.
23. Lopes LR, Murphy C, Syrris P, Dalageorgou C, McKenna WJ, Elliott PM, et al. Use of high-throughput targeted exome-sequencing to screen for copy number variation in hypertrophic cardiomyopathy. *Eur J Med Genet.* 2015; 58(11):611–6. <https://doi.org/10.1016/j.ejmg.2015.10.001> PMID: 26455666
24. Ceyhan-Birsoy O, Pugh TJ, Bowser MJ, Hynes E, Frisella AL, Mahanta LM, et al. Next generation sequencing-based copy number analysis reveals low prevalence of deletions and duplications in 46 genes associated with genetic cardiomyopathies. *Mol Genet Genomic Med.* 2016; 4(2):143–51. <https://doi.org/10.1002/mgg3.187> PMID: 27066507

25. Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*. 2012; 9(12):1185–8. <https://doi.org/10.1038/nmeth.2221> PMID: [23103880](https://pubmed.ncbi.nlm.nih.gov/23103880/)
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352> PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/)
27. 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(1):308–11. PMID: [11125122](https://pubmed.ncbi.nlm.nih.gov/11125122/)
28. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP) [database on the Internet]. 2016. Available from: <http://evs.gs.washington.edu/EVS/>.
29. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature*. 2015; 526(7571):68–74. <https://doi.org/10.1038/nature15393> PMID: [26432245](https://pubmed.ncbi.nlm.nih.gov/26432245/)
30. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016; 536(7616):285–91. <https://doi.org/10.1038/nature19057> PMID: [27535533](https://pubmed.ncbi.nlm.nih.gov/27535533/)
31. The Human Gene Mutation Database [database on the Internet]. Institute of Medical Genetics in Cardiff. 2016. Available from: <http://www.hgmd.cf.ac.uk/ac/index.php>.
32. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014; 42(Database issue):D980–5. <https://doi.org/10.1093/nar/gkt1113> PMID: [24234437](https://pubmed.ncbi.nlm.nih.gov/24234437/)
33. Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, et al. A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet*. 2013; 84(5):453–63. <https://doi.org/10.1111/cge.12257> PMID: [24033266](https://pubmed.ncbi.nlm.nih.gov/24033266/)
34. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17(5):405–24. <https://doi.org/10.1038/gim.2015.30> PMID: [25741868](https://pubmed.ncbi.nlm.nih.gov/25741868/)
35. 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(4):248–9. <https://doi.org/10.1038/nmeth0410-248> PMID: [20354512](https://pubmed.ncbi.nlm.nih.gov/20354512/)
36. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012; 7(10):e46688. <https://doi.org/10.1371/journal.pone.0046688> PMID: [23056405](https://pubmed.ncbi.nlm.nih.gov/23056405/)
37. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014; 11(4):361–2. <https://doi.org/10.1038/nmeth.2890> PMID: [24681721](https://pubmed.ncbi.nlm.nih.gov/24681721/)
38. Roberts AM, Ware JS, Herman DS, Schafer S, Baksi J, Bick AG, et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Sci Transl Med*. 2015; 7(270):270ra6. <https://doi.org/10.1126/scitranslmed.3010134> PMID: [25589632](https://pubmed.ncbi.nlm.nih.gov/25589632/)
39. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. 2009; 84(4):524–33. <https://doi.org/10.1016/j.ajhg.2009.03.010> PMID: [19344873](https://pubmed.ncbi.nlm.nih.gov/19344873/)
40. Database of Genomic Variants [database on the Internet]. 2016. Available from: <http://dgv.tcag.ca/dgv/app/home>.
41. ClinGen: Clinical Genome Resource [database on the Internet]. 2016. Available from: <http://clinicalgenome.org/>.
42. Tian T, Wang J, Wang H, Sun K, Wang Y, Jia L, et al. A low prevalence of sarcomeric gene variants in a Chinese cohort with left ventricular non-compaction. *Heart Vessels*. 2015; 30(2):258–64. <https://doi.org/10.1007/s00380-014-0503-x> PMID: [24691700](https://pubmed.ncbi.nlm.nih.gov/24691700/)
43. Villard E, Duboscq-Bidot L, Charron P, Benaiche A, Conraads V, Sylvius N, et al. Mutation screening in dilated cardiomyopathy: prominent role of the beta myosin heavy chain gene. *Eur Heart J*. 2005; 26(8):794–803. <https://doi.org/10.1093/eurheartj/ehi193> PMID: [15769782](https://pubmed.ncbi.nlm.nih.gov/15769782/)
44. Gomez J, Reguero JR, Moris C, Martin M, Alvarez V, Alonso B, et al. Mutation analysis of the main hypertrophic cardiomyopathy genes using multiplex amplification and semiconductor next-generation sequencing. *Circ J*. 2014; 78(12):2963–71. PMID: [25342278](https://pubmed.ncbi.nlm.nih.gov/25342278/)
45. Meder B, Haas J, Keller A, Heid C, Just S, Borries A, et al. Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. *Circ Cardiovasc Genet*. 2011; 4(2):110–22. <https://doi.org/10.1161/CIRCGENETICS.110.958322> PMID: [21252143](https://pubmed.ncbi.nlm.nih.gov/21252143/)



46. Mook OR, Haagmans MA, Soucy JF, van de Meerakker JB, Baas F, Jakobs ME, et al. Targeted sequence capture and GS-FLX Titanium sequencing of 23 hypertrophic and dilated cardiomyopathy genes: implementation into diagnostics. *J Med Genet.* 2013; 50(9):614–26. <https://doi.org/10.1136/jmedgenet-2012-101231> PMID: 23785128
47. D'Argenio V, Frisso G, Precone V, Boccia A, Fienga A, Pacileo G, et al. DNA sequence capture and next-generation sequencing for the molecular diagnosis of genetic cardiomyopathies. *J Mol Diagn.* 2014; 16(1):32–44. <https://doi.org/10.1016/j.jmoldx.2013.07.008> PMID: 24183960
48. Glotov AS, Kazakov SV, Zhukova EA, Alexandrov AV, Glotov OS, Pakin VS, et al. Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group. *Clin Chim Acta.* 2015; 446:132–40. <https://doi.org/10.1016/j.cca.2015.04.014> PMID: 25892673
49. Liu X, Jiang T, Piao C, Li X, Guo J, Zheng S, et al. Screening Mutations of MYBPC3 in 114 Unrelated Patients with Hypertrophic Cardiomyopathy by Targeted Capture and Next-generation Sequencing. *Sci Rep.* 2015; 5:11411. <https://doi.org/10.1038/srep11411> PMID: 26090888
50. Bottillo I, D'Angelantonio D, Caputo V, Paiardini A, Lipari M, De Bernardo C, et al. Molecular analysis of sarcomeric and non-sarcomeric genes in patients with hypertrophic cardiomyopathy. *Gene.* 2016; 577(2):227–35. <https://doi.org/10.1016/j.gene.2015.11.048> PMID: 26656175
51. Cecconi M, Parodi MI, Formisano F, Spirito P, Autore C, Musumeci MB, et al. Targeted next-generation sequencing helps to decipher the genetic and phenotypic heterogeneity of hypertrophic cardiomyopathy. *Int J Mol Med.* 2016; 38(4):1111–24. <https://doi.org/10.3892/ijmm.2016.2732> PMID: 27600940
52. Rubattu S, Bozzao C, Pennacchini E, Pagannone E, Musumeci BM, Piane M, et al. A Next-Generation Sequencing Approach to Identify Gene Mutations in Early- and Late-Onset Hypertrophic Cardiomyopathy Patients of an Italian Cohort. *Int J Mol Sci.* 2016; 17(8).
53. Maron BJ, Maron MS, Semsarian C. Double or compound sarcomere mutations in hypertrophic cardiomyopathy: a potential link to sudden death in the absence of conventional risk factors. *Heart rhythm: the official journal of the Heart Rhythm Society.* 2012; 9(1):57–63.
54. Herman DS, Lam L, Taylor MR, Wang L, Teekakirikul P, Christodoulou D, et al. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med.* 2012; 366(7):619–28. <https://doi.org/10.1056/NEJMoa1110186> PMID: 22335739
55. Gigli M, Begay RL, Morea G, Graw SL, Sinagra G, Taylor MR, et al. A Review of the Giant Protein Titin in Clinical Molecular Diagnostics of Cardiomyopathies. *Front Cardiovasc Med.* 2016; 3:21. <https://doi.org/10.3389/fcvm.2016.00021> PMID: 27493940
56. Marian AJ, Yu QT, Mares A Jr., Hill R, Roberts R, Perryman MB. Detection of a new mutation in the beta-myosin heavy chain gene in an individual with hypertrophic cardiomyopathy. *J Clin Invest.* 1992; 90(6):2156–65. <https://doi.org/10.1172/JCI116101> PMID: 1361491
57. Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, et al. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods.* 2009; 6(7):507–10. <https://doi.org/10.1038/nmeth.1343> PMID: 19543287
58. Coto E, Reguero JR, Palacin M, Gomez J, Alonso B, Iglesias S, et al. Resequencing the whole MYH7 gene (including the intronic, promoter, and 3' UTR sequences) in hypertrophic cardiomyopathy. *J Mol Diagn.* 2012; 14(5):518–24. <https://doi.org/10.1016/j.jmoldx.2012.04.001> PMID: 22765922
59. Bagnall RD, Yeates L, Semsarian C. The role of large gene deletions and duplications in MYBPC3 and TNNT2 in patients with hypertrophic cardiomyopathy. *Int J Cardiol.* 2010; 145(1):150–3. <https://doi.org/10.1016/j.ijcard.2009.07.009> PMID: 19666196
60. Chanavat V, Seronde MF, Bouvagnet P, Chevalier P, Rousson R, Millat G. Molecular characterization of a large MYBPC3 rearrangement in a cohort of 100 unrelated patients with hypertrophic cardiomyopathy. *Eur J Med Genet.* 2012; 55(3):163–6. <https://doi.org/10.1016/j.ejmg.2012.01.002> PMID: 22314326
61. Pezzoli L, Sana ME, Ferrazzi P, Iacone M. A new mutational mechanism for hypertrophic cardiomyopathy. *Gene.* 2012; 507(2):165–9. <https://doi.org/10.1016/j.gene.2012.06.097> PMID: 22820391
62. Jouven X, Hagege A, Charron P, Carrier L, Dubourg O, Langlard JM, et al. Relation between QT duration and maximal wall thickness in familial hypertrophic cardiomyopathy. *Heart.* 2002; 88(2):153–7. PMID: 12117842
63. Nannenberg EA, Michels M, Christiaans I, Majoor-Krakauer D, Hoedemaekers YM, van Tintelen JP, et al. Mortality risk of untreated myosin-binding protein C-related hypertrophic cardiomyopathy: insight into the natural history. *J Am Coll Cardiol.* 2011; 58(23):2406–14. <https://doi.org/10.1016/j.jacc.2011.07.044> PMID: 22115648
64. Haghghi K, Kolokathis F, Gramolini AO, Waggoner JR, Pater L, Lynch RA, et al. A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. *Proc Natl Acad Sci U S A.* 2006; 103(5):1388–93. <https://doi.org/10.1073/pnas.0510519103> PMID: 16432188

65. Medin M, Hermida-Prieto M, Monserrat L, Laredo R, Rodriguez-Rey JC, Fernandez X, et al. Mutational screening of phospholamban gene in hypertrophic and idiopathic dilated cardiomyopathy and functional study of the PLN -42 C>G mutation. *Eur J Heart Fail.* 2007; 9(1):37–43. <https://doi.org/10.1016/j.ejheart.2006.04.007> PMID: 16829191
66. Valdes-Mas R, Gutierrez-Fernandez A, Gomez J, Coto E, Astudillo A, Puente DA, et al. Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. *Nat Commun.* 2014; 5:5326. <https://doi.org/10.1038/ncomms6326> PMID: 25351925
67. Friedrich FW, Wilding BR, Reischmann S, Crocini C, Lang P, Charron P, et al. Evidence for FHL1 as a novel disease gene for isolated hypertrophic cardiomyopathy. *Human molecular genetics.* 2012; 21(14):3237–54. <https://doi.org/10.1093/hmg/dds157> PMID: 22523091


ANNEX 4

RESEARCH ARTICLE

Large Genomic Imbalances in Brugada Syndrome

Irene Mademont-Soler¹ , Mel·lina Pinsach-Abuin¹ , Helena Riuró¹, Jesus Mates¹, Alexandra Pérez-Serra¹, Mònica Coll¹, José Manuel Porres², Bernat del Olmo¹, Anna Iglesias¹, Elisabet Selga¹, Ferran Picó¹, Sara Pagans^{1,3}, Carles Ferrer-Costa⁴, Geòrgia Sarquella-Brugada⁵, Elena Arbelo⁶, Sergi Cesar⁶, Josep Brugada^{5,6}, Óscar Campuzano^{1,3}, Ramon Brugada^{1,3,7*}

1 Cardiovascular Genetics Center, University of Girona-IDIBGI, Girona, Spain, **2** Arrhythmia Unit, Hospital Universitario Donostia, San Sebastian, Spain, **3** Department of Medical Sciences, School of Medicine, University of Girona, Girona, Spain, **4** Gendiag SL, Barcelona, Spain, **5** Arrhythmia Unit, Hospital Sant Joan de Déu, University of Barcelona, Barcelona, Spain, **6** Arrhythmia Unit, Hospital Clinic de Barcelona, University of Barcelona, Barcelona, Spain, **7** Cardiovascular Genetics Unit, Hospital Josep Trueta, Girona, Spain

 These authors contributed equally to this work.

* ramon@brugada.org



 OPEN ACCESS

Citation: Mademont-Soler I, Pinsach-Abuin M, Riuró H, Mates J, Pérez-Serra A, Coll M, et al. (2016) Large Genomic Imbalances in Brugada Syndrome. PLoS ONE 11(9): e0163514. doi:10.1371/journal.pone.0163514

Editor: Chunhua Song, Pennsylvania State University, UNITED STATES

Received: June 2, 2016

Accepted: September 9, 2016

Published: September 29, 2016

Copyright: © 2016 Mademont-Soler et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: NGS raw data have been uploaded to Figshare. Accession numbers have been included in the Results section of the manuscript: (<https://dx.doi.org/10.6084/m9.figshare.3564141.v3> and <https://dx.doi.org/10.6084/m9.figshare.3565980.v1>).

Funding: This work has been partially supported by Obra social la Caixa; Ministerio de Economía y Competitividad / Instituto de Salud Carlos III (Spain) (PI14/01773); and Proyecto Investigación Básica Cardiología 2015 de los Socios Estratégicos de la Sociedad Española de Cardiología (Spain). MP-A acknowledges a predoctoral fellowship from

Abstract

Purpose

Brugada syndrome (BrS) is a form of cardiac arrhythmia which may lead to sudden cardiac death. The recommended genetic testing (direct sequencing of *SCN5A*) uncovers disease-causing SNVs and/or indels in ~20% of cases. Limited information exists about the frequency of copy number variants (CNVs) in *SCN5A* in BrS patients, and the role of CNVs in BrS-minor genes is a completely unexplored field.

Methods

220 BrS patients with negative genetic results were studied to detect CNVs in *SCN5A*. 63 cases were also screened for CNVs in BrS-minor genes. Studies were performed by Multiplex ligation-dependent probe amplification or Next-Generation Sequencing (NGS).

Results

The detection rate for CNVs in *SCN5A* was 0.45% (1/220). The detected imbalance consisted of a duplication from exon 15 to exon 28, and could potentially explain the BrS phenotype. No CNVs were found in BrS-minor genes.

Conclusion

CNVs in current BrS-related genes are uncommon among BrS patients. However, as these rearrangements may underlie a portion of cases and they undergo unnoticed by traditional sequencing, an appealing alternative to conventional studies in these patients could be

Generalitat de Catalunya (2014FI_B 00586); and ES acknowledges a Sara Borrell postdoctoral fellowship from Instituto de Salud Carlos III. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Gendiag SL provided support in the form of salary for author CF-C, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific role of this author is articulated in the 'author contributions' section.

Competing Interests: Author RB received funding from FerrerInCode, a commercial company, for this study. Author CF-C received funding from Gendiag SL, a commercial company, for this study. There are no patents, products in development or marketed products to declare. This does not alter our adherence to all the PLOS ONE policies on sharing data and materials.

targeted NGS, including in a single experiment the study of SNVs, indels and CNVs in all the known BrS-related genes.

Introduction

Brugada syndrome (BrS) is a form of cardiac arrhythmia, characterized by a typical electrocardiographic pattern of ST segment elevation in leads V1 to V3, and incomplete or complete right bundle branch block [1]. A common presentation of BrS is syncope, which is caused by fast polymorphic ventricular tachycardia. Such syncope typically occurs in the third and fourth decade of life, and usually at rest or during sleep. In some cases, tachycardia does not terminate spontaneously, and it may degenerate into ventricular fibrillation and lead to sudden death [2].

BrS exhibits an autosomal dominant pattern of inheritance, with incomplete penetrance and variable expressivity. Currently, its global prevalence is estimated at 3–5 in 10 000 people, although the incidence is higher in Southeast Asian countries than in the United States and Europe. The syndrome is genetically heterogeneous and can arise from pathogenic variants in at least 19 different genes [3]. The major gene associated with BrS is *SCN5A*, which encodes for the α -subunit of the voltage-gated cardiac sodium channel $Na_v1.5$. Screening for pathogenic variants in this gene uncovers mutations in approximately 20% of BrS patients [4,5]. An additional 15% of patients can be molecularly diagnosed if the minor genes described as causing BrS are included in the genetic analysis (*ABCC9*, *CACNA1C*, *CACNA2D1*, *CACNB2*, *FGF12*, *GPD1L*, *HCN4*, *KCND3*, *KCNE1L*, *KCNE3*, *KCNH2*, *KCNJ8*, *PKP2*, *RANGRF*, *SCN1B*, *SCN2B*, *SCN3B*, *SCN10A*, *SLMAP* and *TRPM4*) [3,6,7]. Hence, a causal genetic variant is not found in a high percentage of patients with BrS.

Genetic testing of BrS patients generally involves sequencing of protein-coding portions and flanking intronic regions of *SCN5A*, according to current ESC Guidelines [8]. The diagnosis is usually performed by direct sequencing, which does not enable the detection of large genomic imbalances (Copy Number Variants, CNVs), which could explain a portion of BrS cases. Although for other arrhythmogenic disorders (such as Long QT syndrome) a relevant contribution of CNVs to the disease has been described (2–11.5%) [9], limited information is available about their contribution to BrS. In 2011, Eastaugh *et al.* [10] reported the first BrS patient (with a concomitant conduction system disease) with a large rearrangement, consisting of a deletion of exons 9 and 10 of *SCN5A*. This rearrangement underwent unnoticed by traditional sequencing and was detected by a quantitative approach. In addition, it was predicted to cause no functional protein to appear in the membrane, resulting in haploinsufficiency. Such finding led the authors to suggest that assessment of CNVs in *SCN5A* should be considered as a standard part of genetic testing in BrS patients. However, after this first report of a CNV in BrS, only three series have been published regarding the frequency of CNVs in *SCN5A* in genotype-negative BrS patients, and no further large deletions or duplications were identified [11–13]. Although the cohorts studied were relatively small ($N = 38$; $N = 68$; and $N = 37$), the authors concluded that such imbalances do not seem to have a major contribution to BrS. On the other hand, the role of CNVs in minor genes related to BrS is a completely unexplored field.

In this report, we present the largest screening for CNVs in *SCN5A* in genotype-negative BrS patients. We also assess, for the first time, the contribution of large genomic imbalances in BrS-associated minor genes.

Materials and Methods

Study population

Two hundred and twenty non-related patients of European descent with a definite BrS phenotype and negative genetic results (for Single Nucleotide Variants -SNVs- and small insertions/deletions -indels-) were studied to detect CNVs in *SCN5A*. For a portion of cases (N = 63), minor genes related to BrS were also screened. The mean age of the patients at the time of clinical diagnosis was 43.53 ± 13.54 years, and at the time of genetic ascertainment 48.69 ± 14.52 years. The 79% of patients were males. The genetic analyses previously performed in the 220 BrS patients that led to their classification as genotype-negative were: a) in 120 cases, conventional Sanger sequencing of *SCN5A*; b) in 37 cases, Sanger sequencing of the following BrS-related genes: *CACNA1C*, *CACNB2*, *GPD1L*, *HCN4*, *KCNE1L*, *KCNE3*, *KCND3*, *KCNJ8*, *RANGRF*, *SCN1B*, *SCN2B*, *SCN3B* and *SCN5A* (results published by Selga et al., 2015 [13]); c) in 43 cases, Next-Generation Sequencing (NGS) analysis using a custom panel that included the genes *CACNA1C*, *CACNB2*, *GPD1L*, *HCN4*, *PKP2* and *SCN5A* (NGS panel 1); and d) in 20 cases, NGS analysis with a custom panel that included the genes *ABCC9*, *CACNA1C*, *CACNA2D1*, *CACNB2*, *GPD1L*, *HCN4*, *KCND3*, *KCNE1L*, *KCNE3*, *KCNJ8*, *PKP2*, *RANGRF*, *SCN1B*, *SCN2B*, *SCN5A*, *SLMAP* and *TRPM4* (NGS panel 2). All assays were performed on total genomic DNA isolated from blood or saliva samples using Chemagen MSM I (PerkinElmer, Germany). For the genes mentioned, only coding regions and flanking intronic sequences were analyzed. NGS panels were developed by Gendiag.exe SL and commercialized by Ferrer InCode as SudD inCode[®]. For patients in the categories a) and b) (N = 157), CNVs were only assessed in *SCN5A* by Multiplex ligation-dependent probe amplification (MLPA). For patients in the categories c) and d) (N = 63), CNVs in *SCN5A* as well as the minor genes included in each custom panel were studied after analysis of NGS data with an algorithm developed in our laboratory to detect large genomic imbalances. In cases where a CNV was detected, clinical data and blood samples from relatives were analyzed for segregation studies and interpretation of results. Clinical investigation of relatives included medical history, clinical examination and 12-lead electrocardiogram (ECG). The study was approved by the ethical committee of Hospital Universitari Dr. Josep Trueta de Girona (Spain) and conformed to the ethical guidelines of the Declaration of Helsinki 2008. Informed written consent was obtained from all patients.

Detection of CNVs by MLPA

MLPA analysis was carried out in 157 BrS patients using the commercially available SALSA MLPA P108 *SCN5A* probemix (MRC-Holland, Amsterdam, The Netherlands). This kit contains probes for each exon of *SCN5A* and one probe upstream of this gene (isoform NM_198056.2). Remarkably, for exon 1 the probe is within the intron (beginning 209 nt after exon 1), and for exon 28 two probes are included. For each experiment, 3 reference DNAs from healthy individuals were used. The MLPA protocol was carried out according to manufacturer's instructions (MRC-Holland). After the multiplex PCR reaction, electrophoresis was performed using the ABI3130XL genetic analyzer with LIZ500 size standard (both from Applied Biosystems, Waltham, MA, USA), and results were analyzed using Coffalyser.Net (MRC-Holland). A reduction or increase in the relative signal strength of $>30\%$ was considered as a deletion or duplication of the locus, respectively. For confirmation, each CNV identified was studied by an alternative method.

Detection of CNVs by NGS

Sequencing data from 63 BrS patients that were prepared with NGS custom panels 1 and 2 (including *SCN5A* and several BrS-associated minor genes) were analyzed for detection of

CNVs. Sample libraries had been prepared following the SureSelect XT Target Enrichment System for Illumina Paired-End Sequencing Library protocol (Agilent Technologies, Santa Clara, CA, USA). Indexed libraries had been sequenced in a ten-sample pool on a MiSeq platform (Illumina, San Diego, CA, USA), with 2x75 bp reads length. For the detection of CNVs, a bioinformatic algorithm developed in our laboratory was used. In brief, the approach focuses on capturing significant differences between the expected and the obtained normalized coverages for a given sample in every exon of the genes of interest. Raw coverage is first normalized by the amount of DNA yielded for each sample in the run. Then the insert size and the low probe affinity bias for targeted regions with a too high GC content (>75%) or too low GC content (<45%) are corrected. Finally, the ratio between each sample and a built-in baseline is evaluated. If the ratio falls outside a signal-to-noise window and is greater or lower than the duplication or deletion cut-offs (0.45 and -0.8, respectively), the gain or loss is inferred. For confirmation, each CNV identified was studied by an alternative method.

Results

CNVs in *SCN5A*

Among the 220 genotype-negative BrS patients investigated for CNVs in *SCN5A* by NGS or MLPA, one large genomic imbalance was detected. Thus, the detection rate for CNVs in *SCN5A* in this cohort was 0.45%. The imbalance consisted of a large duplication spanning from exon 15 to exon 28 of *SCN5A*. The rearrangement was first detected by MLPA (repeated four times) and then confirmed by NGS (using the custom NGS panel 1) (Fig 1A and 1B). For both techniques, DNA obtained from the same fresh whole blood sample was used. The NGS analysis did not reveal any SNV, indel or CNV different from that of *SCN5A* that could potentially explain the BrS phenotype. Interestingly, the imbalance was found in a mosaic state, as the signals of both techniques for all duplicated exons were lower than expected for a heterozygous duplication (ratio of 1.5 for MLPA and log₂ ratio of 0.6 for NGS). Moreover, the signals for the first and last exons of the duplicated region were lower than those for the other exons, suggesting that the rearrangement could be more complex than a typical duplication. To further characterize this rearrangement, new fresh whole blood and saliva samples were requested 11 years later to obtain DNA and RNA. When MLPA (performed in blood and saliva) and NGS (performed in blood, with custom panel 2) were performed in these new samples, the imbalance was not detected. Again, no SNVs, indels and CNVs that could potentially explain the BrS phenotype were detected. To discard sample swapping, the NGS results for SNVs and indels of the first and second blood samples were compared. The analysis of the genes included in both panels (as each sample was prepared with a different NGS panel) revealed the identical 107 SNVs and 3 indels in both samples, so the possibility of sample swapping was discarded. For further studying the case, a dermal biopsy was collected from the patient and kept frozen until DNA extraction (using Chemagen MSM I, PerkinElmer). The rearrangement was not detected in this tissue.

The patient with the *SCN5A* rearrangement was a 48-year-old woman diagnosed with BrS after a syncopal episode. On ECG examination, a type I Brugada pattern was detected in V1 and V2. Electrophysiological study showed that she was inducible into non-sustained ventricular arrhythmias. Flecainide test was positive (Fig 2A and 2B, basal ECG and flecainide test, respectively). She was implanted with a defibrillator. Clinical evaluation and genetic study by MLPA of the index case relatives (parents, two siblings, one son and one daughter) revealed that they presented neither a BrS phenotype nor the *SCN5A* rearrangement (Fig 2C).

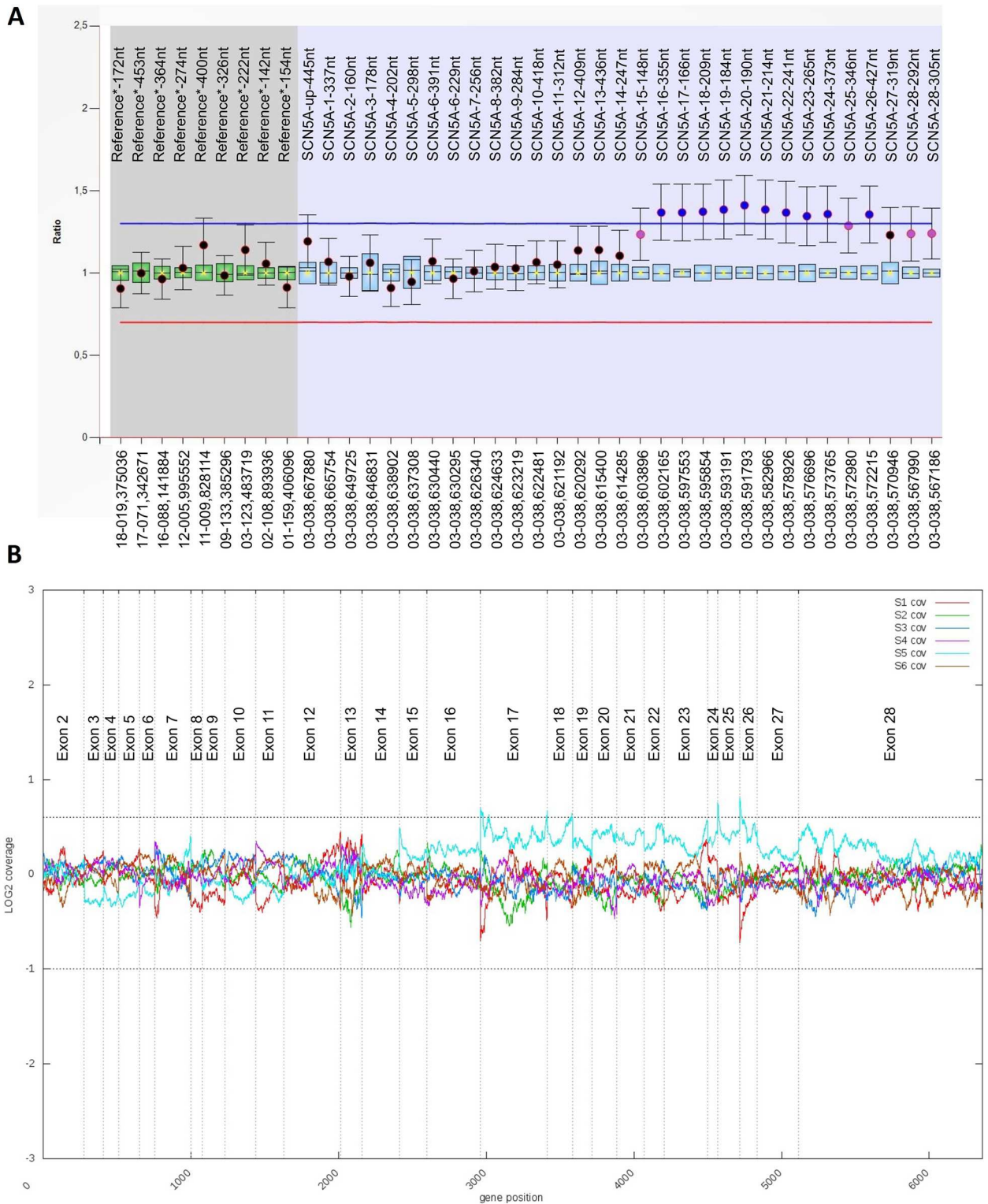


Fig 1. *SCN5A* duplication in a Brugada syndrome patient. Results of Multiplex ligation-dependent probe amplification (A) and Next-Generation Sequencing (B, patient in light blue) showing the duplication from exon 15 to 28 of *SCN5A*. Exon numbering according to isoform NM_198056.

doi:10.1371/journal.pone.0163514.g001

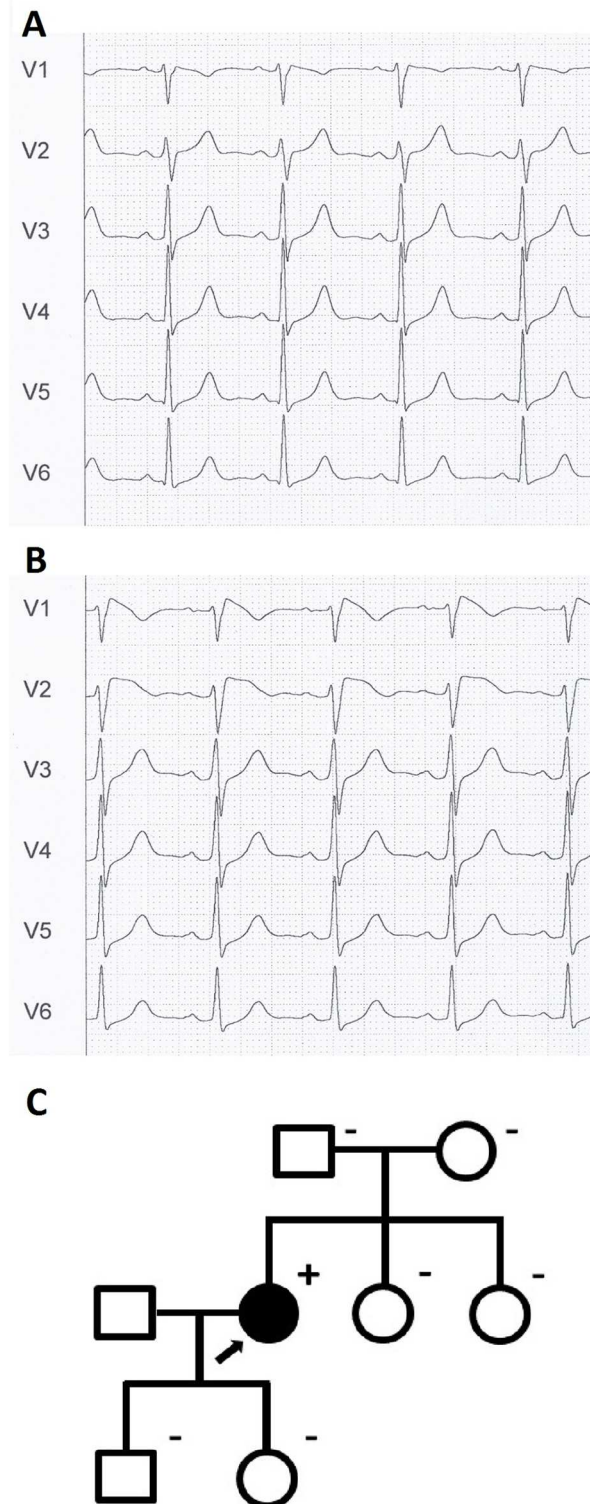


Fig 2. Clinical data from the patient with the Copy Number Variant in *SCN5A*. **A.** Basal electrocardiogram. **B.** Electrocardiogram after flecainide test, showing type I Brugada pattern. **C.** Family pedigree. The proband is indicated by an arrow. Subjects affected and unaffected by BrS are indicated by solid and open symbols, respectively. Genetic status for the *SCN5A* rearrangement is indicated by a superindex (+ or -).

doi:10.1371/journal.pone.0163514.g002

CNVs in minor genes

No large exon duplications or deletions in the BrS-associated minor genes screened were found in any of the 63 genotype-negative BrS patients studied by NGS. Specifically, in 43 patients the analysis involved the genes *CACNA1C*, *CACNB2*, *GPD1L*, *HCN4*, and *PKP2*; and in 20 patients the genes *ABCC9*, *CACNA1C*, *CACNA2D1*, *CACNB2*, *GPD1L*, *HCN4*, *KCND3*, *KCNE1L*, *KCNE3*, *KCNJ8*, *PKP2*, *RANGRF*, *SCN1B*, *SCN2B*, *SLMAP*, and *TRPM4*. NGS raw data have been uploaded to Figshare (<https://dx.doi.org/10.6084/m9.figshare.3564141.v3> and <https://dx.doi.org/10.6084/m9.figshare.3565980.v1>).

Discussion

Large-scale genomic imbalances are a significant contributor to the molecular pathology of a number of different genetic disorders [9,14]. To determine the involvement of such rearrangements to BrS, we performed the largest screening for CNVs in *SCN5A* in genotype-negative BrS patients and assessed, for the first time, their prevalence in BrS-associated minor genes. We also evaluated the clinical value of investigating these large rearrangements in BrS patients as part of the routine molecular testing.

Only one large genomic imbalance was detected after screening *SCN5A* for CNVs in our series of 220 genotype-negative BrS patients (for SNVs and indels in *SCN5A* and in some of them also in BrS-associated minor genes) (0.45%). These results indicate that although such rearrangements in *SCN5A* are not very common among BrS patients, they are found at least in some of the cases who test negative for disease-causing variants in BrS-related genes using conventional sequencing methods. The obtained results are in agreement with published data. A previous study reported a BrS patient (with a concomitant conduction system disease) with a large deletion in *SCN5A* that was considered the underlying cause of the phenotype [10]. However, the other three published series evaluating CNVs in *SCN5A* in BrS patients revealed a frequency of such rearrangements of 0% (cohort sizes $N = 38$, $N = 68$ and $N = 37$) [11–13]. The latter study was published by our group and the cases are included in the present series.

The CNV identified in *SCN5A* in a BrS patient consisted of a large rearrangement involving a duplication of exons 15 to 28 in a mosaic state. According to the signals obtained from MLPA and NGS analysis, the imbalance is probably more complex than a typical tandem duplication. Unfortunately, the rearrangement could not be further characterized since it was not detected in the blood, saliva and skin samples received 11 years later. However, we suggest that the imbalance may be present in heart, being responsible for the observed BrS phenotype as: a) the CNV involves *SCN5A* gene, which is the most significant gene so far described as causing BrS; b) the rearrangement is *de novo* and not transmitted to the offspring, and the index case is the only family member affected by BrS; c) radical variants (including nonsense variants, indels, frameshift variants, and variants affecting splice sites) are a well-known cause of BrS, as they lead to loss of sodium channel function. The rearrangement identified could also result in a complete loss of function of the affected allele, leading to the BrS phenotype. The option of the duplication being inserted in another region of the genome (thus not altering *SCN5A* structure and expression) can not be discarded. However, only 2.5% of clinically relevant duplications are found to be insertional translocations [15].

In relation to the mosaic state of the rearrangement, to our knowledge there are no published BrS patients with pathogenic genetic variants (SNVs, indels or CNVs) found in mosaic state. However, mosaicism for CNVs has been nowadays largely reported in both healthy individuals and in association with disease [16–18]. A particularity of our case is that the mosaicism in blood disappeared within a period of eleven years. Cases with genetic analysis being repeated in the same individual after a period of several years are rare, but there are some

publications reporting the follow-up of mosaic cases which have revealed differences in the proportions of mosaicism)[19–24]. Furthermore, some of these studies report cases with complete disappearance of the normal or the abnormal cell line, a phenomenon first described by La Marche et al. in 1967 as “disappearing mosaicism” [25]. Regarding this issue, two previous reports deserve special attention. In 1984, Motegi and Minoda [20] reported three patients with retinoblastoma and 13q14 deletion mosaicism for which a significant decrease in the proportion of the abnormal cell line was observed in peripheral lymphocytes over time. On the other hand, Morales et al. reported in 2007 [23] a newborn with a partial duplication of chromosome 7q and the complementary deletion, in whom the cell line with the deletion completely disappeared in blood after the first year of life. It is important to highlight that although mosaicism may disappear from blood, it could remain in other more stable tissues (such as skin, brain or heart). Blood cells are unfortunately an unstable source of genetic material given multiple rounds of self-renewal during hematopoiesis and, moreover, the diversity of the clonal lineages that give rise to circulating blood cells appears to decrease with age [18]. Considering all these data, in our case we hypothesize that, although disappearing from blood, the abnormal cell line with the duplication of several exons of *SCN5A* may be present in the heart, resulting in the observed BrS phenotype. Our hypothesis is consistent with previous studies, such as those performed in right ventricular outflow tract tachycardia and atrial fibrillation, which showed that certain genetic variants found in cardiac tissue could be completely absent in blood [26–28]. The fact that the *SCN5A* duplication was not detected in skin may be explained by the embryological origin of the tissues under consideration. Whereas blood and cardiac tissue are derived exclusively from the mesoderm, skin is formed from both mesoderm and ectoderm. All these data suggest that a portion of sporadic BrS cases may arise from pathogenic variants found in the heart and not detectable in other tissues. Molecular diagnosis of these cases is important, as if the mosaicism involves the germline there is a high risk for recurrence of the disease. Thus, we believe that further research in this field will prove beneficial for better understanding of the role of mosaicism in BrS, and for determining the appropriate tissue for diagnosis of sporadic BrS cases.

On the other hand, in the present study we explored, for the first time, the presence of CNVs in BrS-minor genes as responsible of the phenotype in 63 genotype-negative BrS patients. Specifically, the minor genes investigated were *ABCC9*, *CACNA1C*, *CACNA2D1*, *CACNB2*, *GPD1L*, *HCN4*, *KCND3*, *KCNE1L*, *KCNE3*, *KCNJ8*, *PKP2*, *RANGRF*, *SCN1B*, *SCN2B*, *SLMAP*, and *TRPM4*. No large imbalances were detected, suggesting that such imbalances in these genes do not probably have a major contribution to BrS. However, further studies with larger cohorts are needed to elucidate the precise involvement of CNVs in these genes in BrS patients.

In most laboratories, the current approach for molecular diagnosis of BrS patients involves exclusively conventional Sanger sequencing of *SCN5A*, which does not enable the screening for CNVs. Although SNVs and indels in *SCN5A* are the main recognized cause of BrS, 20 other genes have been associated with the disease so far, and CNVs may also explain a portion of cases [3–10]. As the identification of the genetic variant causing the BrS phenotype is the only way to offer an accurate genetic counseling to families and to identify at-risk family members (with the ultimate aim of preventing sudden death), we believe that the best approach for a comprehensive study of BrS patients would be targeted NGS using a panel including all known BrS-related genes, and investigating the presence of CNVs as part of the genetic analysis. Under this scenario, all well-known genetic causes of BrS could be explored. However, as the contribution of the known minor genes and CNVs to BrS seems to be low, a significant proportion of individuals with a clinical diagnosis of BrS will still remain without a positive genetic diagnosis. Further efforts need to be done to describe other causes of BrS, which may include

genetic variants in non-coding regions (i.e. promoters and other regulatory regions, introns, and untranslated regions), novel pathogenic alterations in as yet unknown genes and, as previously discussed, cardiac-specific mosaicism.

In conclusion, our results after performing the largest screening for CNVs in *SCN5A* and minor genes in BrS patients reveal that such rearrangements are not a common finding among genotype-negative BrS patients. However, as these rearrangements may underlie a portion of cases and they undergo unnoticed by traditional sequencing, we believe that an appealing alternative to conventional studies in BrS patients would be targeted NGS, including in a single experiment the study of SNVs, indels and CNVs in all the known BrS-related genes.

Acknowledgments

MP-A acknowledges a predoctoral fellowship from Generalitat de Catalunya (2014FI_B 00586); and ES acknowledges a Sara Borrell postdoctoral fellowship from Instituto de Salud Carlos III.

Author Contributions

Conceptualization: IM-S RB.

Data curation: IM-S JM AI.

Formal analysis: IM-S MP-A HR JM.

Funding acquisition: OC RB.

Investigation: IM-S MP-A HR AP-S MC FP SP ES.

Methodology: IM-S RB OC.

Project administration: RB.

Resources: JMP AI GS-B EA SC JB RB.

Software: JM BO CF-C.

Supervision: OC RB.

Validation: IM-S.

Visualization: IM-S.

Writing – original draft: IM-S MP-A HR.

Writing – review & editing: IM-S MP-A OC RB.

References

1. Brugada P, Brugada J Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report. *J Am Coll Cardiol.* 1992; 20: 1391–1396. PMID: [1309182](#)
2. Napolitano C, Priori SG. Brugada syndrome. *Orphanet J Rare Dis.* 2006; 1: 35. PMID: [16972995](#)
3. Sarquella-Brugada G, Campuzano O, Arbelo E, Brugada J, Brugada R. Brugada syndrome: clinical and genetic findings. *Genet Med.* 2016; 18: 3–12. doi: [10.1038/gim.2015.35](#) PMID: [25905440](#)
4. Priori SG, Napolitano C, Gasparini M, Pappone C, Della Bella P, Giordano U, et al. Natural history of Brugada syndrome: insights for risk stratification and management. *Circulation.* 2002; 105: 1342–1347. PMID: [11901046](#)
5. Kapplinger JD, Tester DJ, Alders M, Benito B, Berthet M, Brugada J, et al. An international compendium of mutations in the *SCN5A*-encoded cardiac sodium channel in patients referred for Brugada

- syndrome genetic testing. *Heart Rhythm*. 2010; 7: 33–46. doi: [10.1016/j.hrthm.2009.09.069](https://doi.org/10.1016/j.hrthm.2009.09.069) PMID: [20129283](https://pubmed.ncbi.nlm.nih.gov/20129283/)
6. Hennessey JA, Marcou CA, Wang C, Wei EQ, Tester DJ, Torchio M, et al. FGF12 is a candidate Brugada syndrome locus. *Heart Rhythm*. 2013; 10: 1886–1894. doi: [10.1016/j.hrthm.2013.09.064](https://doi.org/10.1016/j.hrthm.2013.09.064) PMID: [24096171](https://pubmed.ncbi.nlm.nih.gov/24096171/)
 7. Wang Q, Ohno S, Ding WG, Fukuyama M, Miyamoto A, Itoh H, et al. Gain-of-function KCNH2 mutations in patients with Brugada syndrome. *J Cardiovasc Electrophysiol*. 2014; 25: 522–530. doi: [10.1111/jce.12361](https://doi.org/10.1111/jce.12361) PMID: [24400717](https://pubmed.ncbi.nlm.nih.gov/24400717/)
 8. Priori SG, Blomstrom-Lundqvist C, Mazzanti A, Blom N, Borggrefe M, Camm J, et al. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). *Eur Heart J*. 2015; 36: 2793–2867. doi: [10.1093/eurheartj/ehv316](https://doi.org/10.1093/eurheartj/ehv316) PMID: [26320108](https://pubmed.ncbi.nlm.nih.gov/26320108/)
 9. Campuzano O, Sarquella-Brugada G, Mademont-Soler I, Allegue C, Cesar S, Ferrer-Costa C, et al. Identification of Genetic Alterations, as Causative Genetic Defects in Long QT Syndrome, Using Next Generation Sequencing Technology. *PLoS One*. 2014; 9: e114894. doi: [10.1371/journal.pone.0114894](https://doi.org/10.1371/journal.pone.0114894) PMID: [25494010](https://pubmed.ncbi.nlm.nih.gov/25494010/)
 10. Eastaugh LJ, James PA, Phelan DG, Davis AM. Brugada syndrome caused by a large deletion in SCN5A only detected by multiplex ligation-dependent probe amplification. *J Cardiovasc Electrophysiol*. 2011; 22: 1073–1076. doi: [10.1111/j.1540-8167.2010.02003.x](https://doi.org/10.1111/j.1540-8167.2010.02003.x) PMID: [21288276](https://pubmed.ncbi.nlm.nih.gov/21288276/)
 11. Koopmann TT, Beekman L, Alders M, Meregalli PG, Mannens MM, Moorman AF, et al. Exclusion of multiple candidate genes and large genomic rearrangements in SCN5A in a Dutch Brugada syndrome cohort. *Heart Rhythm*. 2007; 4: 752–755. PMID: [17556197](https://pubmed.ncbi.nlm.nih.gov/17556197/)
 12. Garcia-Molina E, Lacunza J, Ruiz-Espejo F, Sabater M, Garcia-Alberola A, Gimeno JR, et al. A study of the SCN5A gene in a cohort of 76 patients with Brugada syndrome. *Clin Genet*. 2013; 83: 530–538. doi: [10.1111/cge.12017](https://doi.org/10.1111/cge.12017) PMID: [22984773](https://pubmed.ncbi.nlm.nih.gov/22984773/)
 13. Selga E, Campuzano O, Pinsach-Abuin ML, Perez-Serra A, Mademont-Soler I, Riuro H, et al. Comprehensive Genetic Characterization of a Spanish Brugada Syndrome Cohort. *PLoS One*. 2015; 10: e0132888. doi: [10.1371/journal.pone.0132888](https://doi.org/10.1371/journal.pone.0132888) PMID: [26173111](https://pubmed.ncbi.nlm.nih.gov/26173111/)
 14. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet*. 2010; 86: 749–764. doi: [10.1016/j.ajhg.2010.04.006](https://doi.org/10.1016/j.ajhg.2010.04.006) PMID: [20466091](https://pubmed.ncbi.nlm.nih.gov/20466091/)
 15. Newman S, Hermetz KE, Wechselblatt B, Rudd MK. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am J Hum Genet*. 2015; 96: 208–220. doi: [10.1016/j.ajhg.2014.12.017](https://doi.org/10.1016/j.ajhg.2014.12.017) PMID: [25640679](https://pubmed.ncbi.nlm.nih.gov/25640679/)
 16. Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, Sandgren J, et al. Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat*. 2008; 29: 1118–1124. doi: [10.1002/humu.20815](https://doi.org/10.1002/humu.20815) PMID: [18570184](https://pubmed.ncbi.nlm.nih.gov/18570184/)
 17. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci U S A*. 2012; 109: 18018–18023. doi: [10.1073/pnas.1213736109](https://doi.org/10.1073/pnas.1213736109) PMID: [23043118](https://pubmed.ncbi.nlm.nih.gov/23043118/)
 18. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet*. 2015; 31: 382–392. doi: [10.1016/j.tig.2015.03.013](https://doi.org/10.1016/j.tig.2015.03.013) PMID: [25910407](https://pubmed.ncbi.nlm.nih.gov/25910407/)
 19. Holmgren G, Jagell S, Lagerkvist B, Nordenson I. A pair of siblings with diastrophic dysplasia and E trisomy mosaicism. *Hum Hered*. 1984; 34: 266–268. PMID: [6479995](https://pubmed.ncbi.nlm.nih.gov/6479995/)
 20. Motegi T, Minoda K. A decreasing tendency for cytogenetic abnormality in peripheral lymphocytes of retinoblastoma patients with 13q14 deletion mosaicism. *Hum Genet*. 1984; 66: 186–189. PMID: [6714979](https://pubmed.ncbi.nlm.nih.gov/6714979/)
 21. McCorquodale MM, Bowdle FC. Two pregnancies and the loss of the 46,XX cell line in a 45,X/46,XX Turner mosaic patient. *Fertil Steril*. 1985; 43: 229–233. PMID: [3967782](https://pubmed.ncbi.nlm.nih.gov/3967782/)
 22. Priest JH, Rust JM, Fernhoff PM. Tissue specificity and stability of mosaicism in Pallister-Killian +1 (12p) syndrome: relevance for prenatal diagnosis. *Am J Med Genet*. 1992; 42: 820–824. doi: [10.1002/ajmg.1320420615](https://doi.org/10.1002/ajmg.1320420615) PMID: [1554021](https://pubmed.ncbi.nlm.nih.gov/1554021/)
 23. Morales C, Madrigal I, Esque T, de la Fuente JE, Rodriguez JM, Margarit E, et al. Duplication/deletion mosaicism of the 7q(21.1 → 31.3) region. *Am J Med Genet A*. 2007; 143A: 179–183. doi: [10.1002/ajmg.a.31570](https://doi.org/10.1002/ajmg.a.31570) PMID: [17163539](https://pubmed.ncbi.nlm.nih.gov/17163539/)

24. Gravholt CH, Friedrich U, Nielsen J. Chromosomal mosaicism: a follow-up study of 39 unselected children found at birth. *Hum Genet.* 1991; 88: 49–52. PMID: [1959925](#)
25. La Marche PH, Heisler AB, Kronemer NS. Disappearing mosaicism. Suggested mechanism is growth advantage of normal over abnormal cell population. *R I Med J.* 1967; 50: 184–189. PMID: [5231789](#)
26. Lerman BB, Dong B, Stein KM, Markowitz SM, Linden J, Catanzaro DF. Right ventricular outflow tract tachycardia due to a somatic cell mutation in G protein subunit α_2 . *J Clin Invest.* 1998; 101: 2862–2868. doi: [10.1172/JCI1582](#) PMID: [9637720](#)
27. Gollob MH, Jones DL, Krahn AD, Danis L, Gong XQ, Shao Q, et al. Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *N Engl J Med.* 2006; 354: 2677–2688. doi: [10.1056/NEJMoa052800](#) PMID: [16790700](#)
28. Thibodeau IL, Xu J, Li Q, Liu G, Lam K, Veinot JP, et al. Paradigm of genetic mosaicism and lone atrial fibrillation: physiological characterization of a connexin 43-deletion mutant identified from atrial tissue. *Circulation.* 2010; 122: 236–244. doi: [10.1161/CIRCULATIONAHA.110.961227](#) PMID: [20606116](#)

ANNEX 5

RESEARCH ARTICLE

Identification of Genetic Alterations, as Causative Genetic Defects in Long QT Syndrome, Using Next Generation Sequencing Technology

Oscar Campuzano^{1*}, Georgia Sarquella-Brugada^{2*}, Irene Mademont-Soler¹, Catarina Allegue¹, Sergi Cesar², Carles Ferrer-Costa³, Monica Coll¹, Jesus Mates¹, Anna Iglesias¹, Josep Brugada², Ramon Brugada^{1,4*}

1. Cardiovascular Genetics Center, University of Girona-IdIBGi, Girona, Spain, 2. Arrhythmia Unit, Hospital Sant Joan de Déu, University of Barcelona, Barcelona, Spain, 3. Gendiag SL, Barcelona, Spain, 4. Cardiology Service, Hospital Josep Trueta, Girona, Spain

*ramon@brugada.org

These authors contributed equally to this work.



CrossMark
click for updates

OPEN ACCESS

Citation: Campuzano O, Sarquella-Brugada G, Mademont-Soler I, Allegue C, Cesar S, et al. (2014) Identification of Genetic Alterations, as Causative Genetic Defects in Long QT Syndrome, Using Next Generation Sequencing Technology. PLoS ONE 9(12): e114894. doi:10.1371/journal.pone.0114894

Editor: Bernard Attali, Sackler Medical School, Tel Aviv University, Israel

Received: May 20, 2014

Accepted: November 15, 2014

Published: December 10, 2014

Copyright: © 2014 Campuzano et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are contained within the paper.

Funding: This study has been funded by “La Caixa” Foundation and Fundació Privada Daniel Bravo Andreu. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Dr. Ramon Brugada is consultant of Ferrer-inCode. Dr. Ferrer-Costa is employed by Gendiag SL. This does not alter the authors’ adherence to PLOS ONE policies on sharing data and materials. The other authors have no conflicts of interest to disclose.

Abstract

Background: Long QT Syndrome is an inherited channelopathy leading to sudden cardiac death due to ventricular arrhythmias. Despite that several genes have been associated with the disease, nearly 20% of cases remain without an identified genetic cause. Other genetic alterations such as copy number variations have been recently related to Long QT Syndrome. Our aim was to take advantage of current genetic technologies in a family affected by Long QT Syndrome in order to identify the cause of the disease.

Methods: Complete clinical evaluation was performed in all family members. In the index case, a Next Generation Sequencing custom-built panel, including 55 sudden cardiac death-related genes, was used both for detection of sequence and copy number variants. Next Generation Sequencing variants were confirmed by Sanger method. Copy number variations variants were confirmed by Multiplex Ligation dependent Probe Amplification method and at the mRNA level. Confirmed variants and copy number variations identified in the index case were also analyzed in relatives.

Results: In the index case, Next Generation Sequencing revealed a novel variant in *TTN* and a large deletion in *KCNQ1*, involving exons 7 and 8. Both variants were confirmed by alternative techniques. The mother and the brother of the index case were also affected by Long QT Syndrome, and family cosegregation was observed for the *KCNQ1* deletion, but not for the *TTN* variant.

Conclusions: Next Generation Sequencing technology allows a comprehensive genetic analysis of arrhythmogenic diseases. We report a copy number variation identified using Next Generation Sequencing analysis in Long QT Syndrome. Clinical and familiar correlation is crucial to elucidate the role of genetic variants identified to distinguish the pathogenic ones from genetic noise.

Introduction

The long QT syndrome (LQTS) is an inherited cardiac disorder characterized by prolonged QT interval on the surface electrocardiogram (ECG). It affects 1/2500 individuals, causing lethal ventricular tachycardias (VT), *torsades de pointes* (TdP) and sudden cardiac death (SCD) [1]. These events can be triggered by physical or emotional stress, but in some individuals they may occur during periods of sleep or rest. However, there is important phenotypic heterogeneity [2].

Genetic studies have shown that LQTS is caused by pathogenic mutations in 15 genes encoding cardiac ion channels or membrane adaptors (*KCNQ1*, *KCNH2*, *SCN5A*, *ANK2*, *KCNE1*, *KCNE2*, *KCNJ2*, *CACNA1C*, *CAV3*, *SCN4B*, *AKAP9*, *SNTA1*, *RYR2*, *KCNJ5* and *SCN1B*) [3]. Pathogenic mutations identified in the *KCNQ1* and *KCNH2* genes as well as the sodium channel, encoded by *SCN5A*, are responsible for nearly 80% of all clinically diagnosed cases. All the other genes together explain less than 5% of LQTS cases. Recently, large intragenic deletions and duplications have been reported in LQTS families, suggesting that the cause of disease in some patients could be the presence of copy number variants (CNVs) affecting the major genes for LQTS. Detection rate for CNVs among LQTS patients, mutation-negative by traditional analysis, seem to be around 2–11.5% [4–7]. Other unknown genetic causes might be responsible for the remaining LQTS cases, such as mutations in non-coding regions and novel mutations in as yet unknown genes [6, 8].

Currently, most genetic studies focus on the analysis of the main genes associated with LQTS, following current clinical guidelines for LQTS [9]. All these studies use conventional Sanger sequencing. Because of its high cost, a comprehensive genetic analysis has not regularly been performed in LQTS for all genes. In recent years, Next-Generation Sequencing (NGS) has emerged as a revolutionary technology which enables the generation of high amount of genetic data [10]. This massive amount of information has triggered the development of potent bioinformatic tools to help interpret potential causality implications [11, 12].

The goal of our study was to identify the genetic alteration that could explain the LQTS in our family. Because of substantial percentage of LQTS cases without genetic diagnose after screening of all known LQTS genes, we used a NGS custom panel to screen the main genes associated with SCD.

Materials and Methods

Clinical evaluation

All relatives included in our study were clinically evaluated at our Pediatric Arrhythmia Unit. Complete clinical evaluation, including electrocardiogram (ECG), transthoracic echocardiogram (ECHO), 24-hour ECG Holter recording and exercise test was performed in index case and all relatives. This study was approved by the Ethics Committee of Hospital Josep Trueta (Girona, Spain) and conforms to the principles outlined in the Declaration of Helsinki. All individuals signed a written informed consent to participate in the study. Informed consent of all patients was obtained in accordance with international review board guidelines of Hospital Josep Trueta and Universitat of Girona (Girona, Spain).

DNA sample

Genomic DNA was extracted with Chemagic MSM I from whole blood (Chemagic human blood). DNA samples were checked in order to assure quality and quantify before processing to get the 3µg needed for the NGS strategy. DNA integrity was assessed on a 0,8% agarose gel. Spectrophotometric measurements are also performed to assess quality ratios of absorbance; dsDNA concentration is determined by fluorometry (Qubit, Life Technologies). DNA sample was fragmented by Bioruptor (Diagenode). Library preparation was performed according to the manufacturer's instructions (SureSelect XT Custom 0.5–2.9 Mb library, Agilent Technologies, Inc). After capture, the indexed library was sequenced in a six-sample pool cartridge. Sequencing process was developed on MiSeq System (Illumina) using 2 × 150 bp reads length.

Custom Resequencing panel

We selected 55 genes, the most prevalent involved in SCD-related pathologies, according to available scientific literature. The genomic coordinates corresponding to these 55 genes ([Table 1](#)) were designed using the tool eArray (Agilent Technologies, Inc.). All the isoforms described at the UCSC browser were included at the design. The final size was 432,512 kbp of encoding regions and UTR boundaries. The coordinates of the sequence data is based on NCBI build 37 (UCSC hg19).

Bioinformatics

The secondary bioinformatic analysis of the data obtained includes a first step trimming of the FAST-Q files. The trimmed reads are then mapped with GEM II and output is joined and sorted and uniquely and properly mapping read pairs are selected. Finally, variant call over the cleaned BAM file is performed with SAMtools v.1.18, GATK v2.4 to generate the first raw VCF files. Variants are annotated with dbSNP IDs, Exome Variant Server and the 1000 Genomes browser, in-home database IDs and Ensembl information, if available.

Table 1. List of the 55 SCD-related genes included in our panel and its association with the disease.

DISEASE	GENES
Brugada Syndrome	CACNA1C, CACNB2, GPD1L, HCN4, SCN5A
Long QT Syndrome	ANK2, CACNA1C, CAV3, KCNE1, KCNE2, KCNH2, KCNJ2, KCNQ1, RYR2, SCN4B, SCN5A
Short QT Syndrome	CACNA1C, CACNB2, KCNH2, KCNJ2, KCNQ1
Catecholaminergic Polymorphic Ventricular Tachycardia	CASQ2, KCNJ2, RYR2
Hypertrophic Cardiomyopathy	ACTC1, ACTN2, CAV3, CSRP3, GLA, JPH2, LAMP2, LDB3, MYBPC3, MYH6, MYH7, MYL2, MYL3, MYOZ2, PDLIM3, PLN, PRKAG2, RYR2, TCAP, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL
Dilated Cardiomyopathy	ACTC1, ACTN2, CAV3, CRYAB, CSRP3, DES, DMD, DSC2, DSG2, DSP, EMD, LAMP2, LDB3, LMNA, MYBPC3, MYH6, MYH7, PKP2, PLN, SCN5A, SGCD, TAZ, TCAP, TNNC1, TNNI3, TNNT2, TPM1, TTN, VCL
Arrhythmogenic Right Ventricular Cardiomyopathy	DES, DSC2, DSG2, DSP, JUP, LMNA, PKP2, PLN, TGFB3, TTN

doi:10.1371/journal.pone.0114894.t001

Tertiary analysis is then developed. For each genetic variation identified, allelic frequency was consulted in Exome Variant Server -EVS- (<http://evs.gs.washington.edu/EVS/>) and 1000 genomes database (<http://www.1000genomes.org/>). In addition, Human Gene Mutation Database -HGMD- (<http://www.hgmd.cf.ac.uk/ac/index.php>) was also consulted to identify pathogenic mutations previously reported. *In silico* pathogenicity of novel genetic variations were consulted in CONDEL software (CONsensus DELeteriousness scores of *missense* SNVs) (<http://bg.upf.edu/condel/>), and PROVEAN (Protein Variation Effect Analyzer) (<http://provean.jcvi.org/index.php>). Alignment among species was also performed for these novel variations using UniProt database (<http://www.uniprot.org/>).

Regarding CNV identification using NGS data, a new methodology was developed. Our approach focused on capturing significant differences between expected normalized coverage and obtained normalized coverage for a given sample in the region of interest. We normalized the raw coverage by the amount of DNA yielded for each sample in the MiSeq run. The log₂ ratio data between samples was evaluated. Detection of losses and gains were based on those genomic coordinates with a log₂ ratio near the stringent ratio cut-offs for duplication or deletion (less than -1.0 or greater than 0.6, respectively). Several samples were analyzed to corroborate similar levels of coverage between samples.

Genetic confirmation

Sanger sequencing

Non-common (Minor Allele Frequency –MAF- <1%) genetic variants were confirmed by Sanger method. First, polymerase chain reaction (PCR) was performed. PCR products were purified using ExoSAP-IT (USB Corporation, Cleveland, OH, USA), and the analysis of the exonic and intron-exon regions was performed by direct sequencing (Genetic Analyzer 3130XL, Applied Biosystems,

CA, USA) with posterior SeqScape Software v2.5 (Life Technologies) analysis comparing obtained results with the reference sequence from hg19. Each sample underwent a genetic study of corresponding genes (NCBI -National Center for Biotechnology Information-, <http://www.ncbi.nlm.nih.gov/>) (*TTN* NM_133378). Familial cosegregation of rare genetic variants was also performed using Sanger technology.

Multiplex Ligation dependent Probe Amplification

The CNV detected by NGS was confirmed by Multiplex Ligation dependent Probe Amplification (MLPA), using the probemix SALSA MLPA P114-B2 Long-QT (MRC-Holland, Amsterdam, the Netherlands). This kit contains 17 probes for the *KCNQ1* gene, 16 probes for *KCNH2*, 4 probes for *KCNE1* and 2 probes for *KCNE2*. The MLPA DNA detection and quantification were carried out according to the manufacturer's protocol (MRC-Holland, Amsterdam, The Netherlands). After the multiplex PCR reaction, electrophoresis was performed using the ABI3130xl Genetic Analyzer (Applied Biosystems, CA, USA). Data was collected and analysed with Coffalyser. Net software (MRC-Holland). Significantly (>30%) decreased or increased signals in the patient sample relative to controls were considered as deletions or duplications, respectively. Familial cosegregation of CNVs was also performed using MLPA.

Sequencing of cDNA

The deletion of exons 7 and 8 of *KCNQ1* was also confirmed at the mRNA level in both the brother and the mother of the proband (index case refused analysis, and the healthy father was analysed as a control). Total RNA was isolated with the QIAamp RNA Blood Mini Kit and converted to cDNA with the QuantiTect Reverse Transcription Kit (both from Qiagen, California, USA). Afterwards, amplicon spanning from exon 6 to 9 of *KCNQ1* of the cDNA was generated by PCR using the primers 5'ACCCTGTACATCGGCTTCC3' and 5'GGGTGACAGCAGAGTGTGG3'. PCR products were purified and sequenced (with the same primers) according to the abovementioned protocol for Sanger sequencing.

Results

Clinical

The proband (female, 14 years old) was seen in our Paediatric Arrhythmia Unit for abnormal ECG performed in pre-exercise screening. She was asymptomatic for the cardiac point of view. Baseline ECG showed a corrected QT interval (QTc) using Bazhett formula of 500 ms ([Fig. 1A](#)). She was on no medication and had no ionic alteration which could explain the prolonged QT. Echocardiography was normal. 24-hour ECG Holter showed no arrhythmic events, and exercise test showed long QT interval.

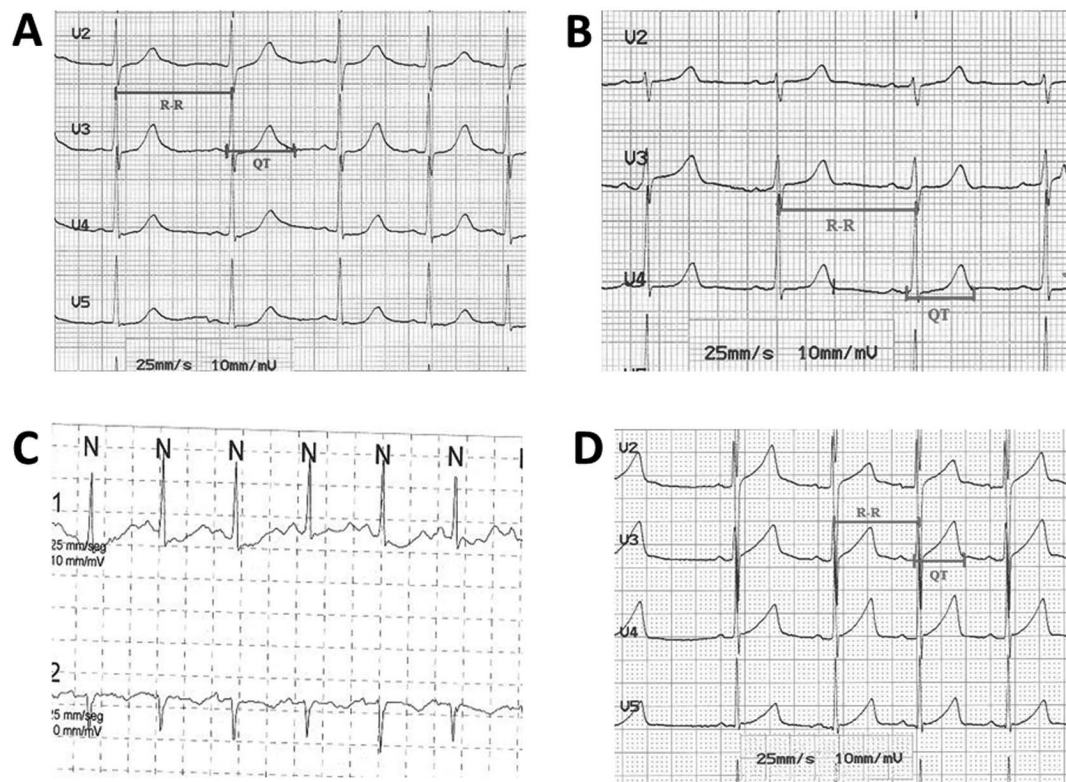


Fig. 1. ECG of family members. (A) Twelve-lead ECG of index case. The ECG shows QTc of 500 ms. (B) Twelve-lead ECG of mother's index case. The ECG shows a normal QTc, and (C) a LQT during tachycardia registered by Holter. (D) Twelve-lead ECG of brother's index case. The ECG shows QTc of 485 ms.

doi:10.1371/journal.pone.0114894.g001

Both parents were studied. The proband's father had a normal ECG, 24-hour ECG Holter and exercise test. The proband's mother had a normal QTc interval at baseline ECG but with paradoxical response to tachycardisation (Fig. 1B, 1C). The 10 year-old brother had prolonged QTc (485 ms) interval at baseline ECG (Fig. 1D). Affected patients were treated with beta-blockers.

NGS analysis

We analyzed 55 genes previously associated with SCD (Table 1). After the NGS process and the application of bioinformatics pipeline, the call rate ranged from 99,6% to 98,92% at 20x and 100x respectively in this sample. We selected the Non Synonymous (NS) variants with a MAF<1% in the EVS for its conventional Sanger sequencing confirmation. Only one single nucleotide variant (SNV) was confirmed in the index case, the *TTN* gene (p.R20729G). This novel variant is consequence of a nucleotide change of A to G (c.62185A>G). The genetic variation was not previously identified in locus specific databases, considered therefore a novel GVUS. It was predicted *in silico* as pathogenic in all databases consulted. In addition, alignment between species showed a high level of

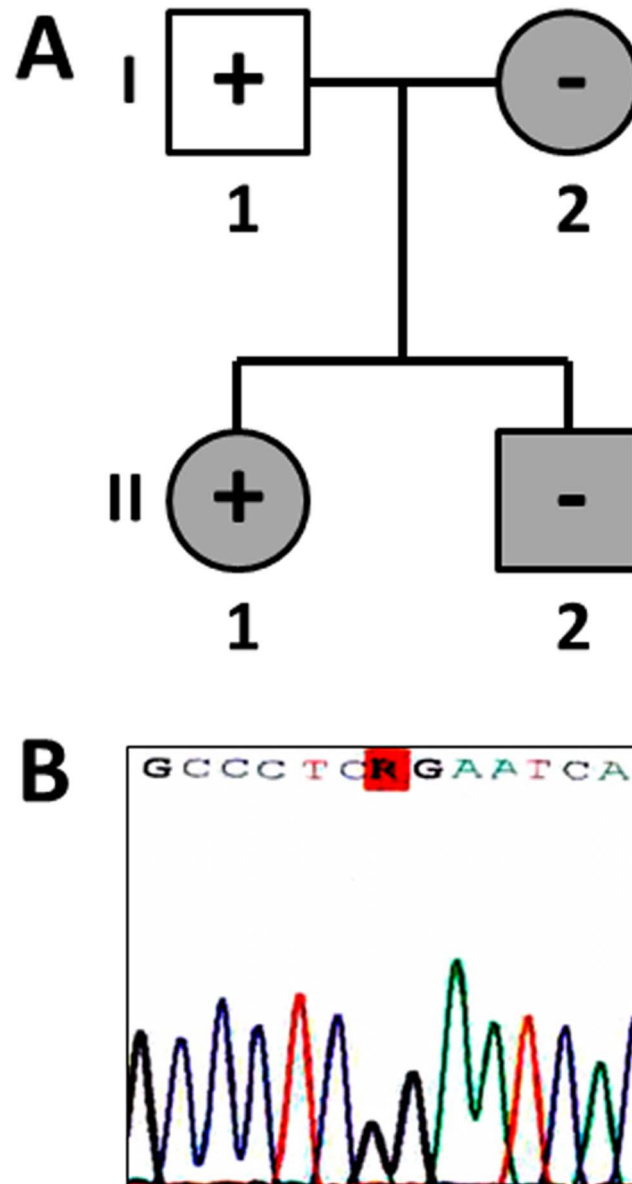


Fig. 2. Pedigree and electropherogram. (A) Index case is II.1. White round/squares indicate healthy status after clinical evaluation. Grey round/squares indicate LQTS after clinical evaluation. Plus sign indicates carrier of genetic variation. Minus sign indicates non-carrier of the genetic variation. (B) Electropherogram of the genetic variation identified (p.R20729G_TTN).

doi:10.1371/journal.pone.0114894.g002

conservation. However, family segregation showed that only the index case’s father carried the same genetic variation (Fig. 2).

On the other hand, NGS analysis revealed a deletion of exons 7 and 8 in the *KCNQ1* gene (Fig. 3). The raw coverage normalization showed that pooled samples were comparable in terms of coverage and no major biases between samples were found (average normalized coverage is 6.7 with sd 0.11 yielding a cv of 1.7%; average sd of normalized coverage is .60 with sd 0.02 yielding a cv of

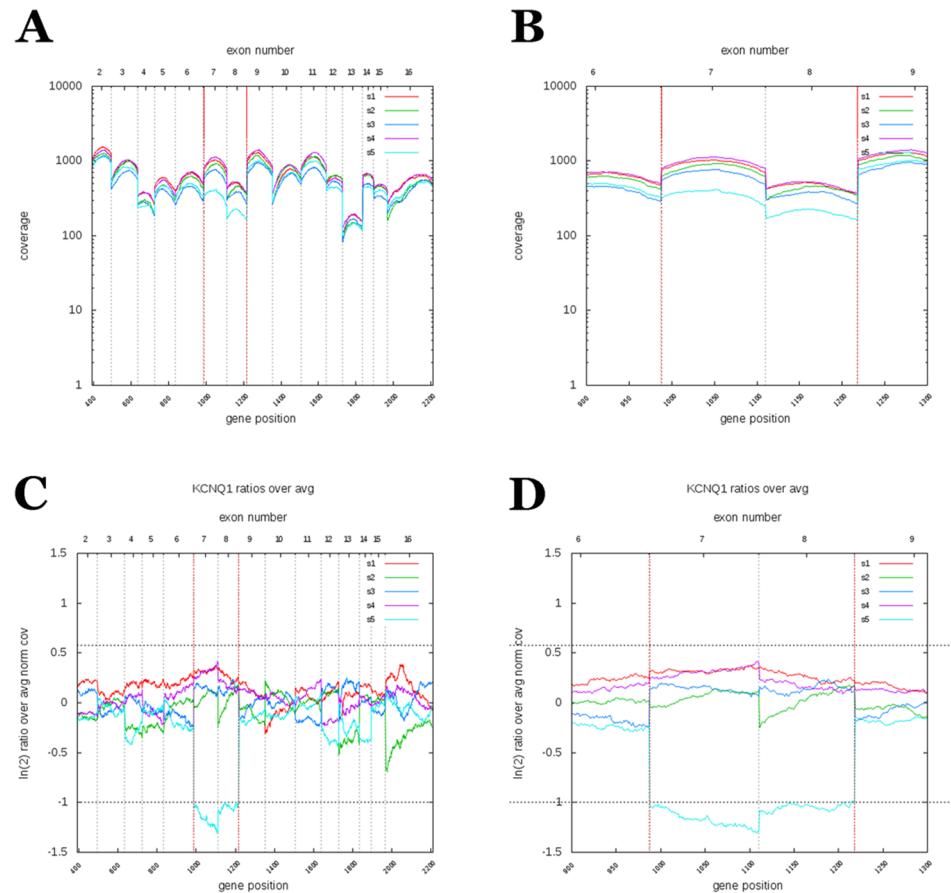


Fig. 3. NGS data showing CNV in the *KCNQ1* gene. (A) Coverage of all exons in the *KCNQ1* gene of several samples. (B) Detail coverage of exons 7 and 8 in several samples. (C) Normalized raw coverage of exons 7 and 8 showing a deletion in comparison to all other exons of the same gene. (D) In detail, normalized raw coverage of exons 7 and 8.

doi:10.1371/journal.pone.0114894.g003

4.1%). Then, the analysis of corrected log₂ ratio coverage by genomic position for each sample was performed. The corrected log₂ ratios fit a Gaussian distribution. A baseline from all pool was inferred and each sample was compared with this prediction. The deviated exons from this baseline were labelled as duplications or deletions. The analysis showed an intense signal over these two exons with more than 6 standard deviations from the mean (log₂ mean ratio for this signal is $-1,1 \pm 0,09$ sd). This CNV alteration was confirmed by MLPA (Fig. 4). Family segregation studies revealed that the brother and the mother of the proband (both affected by LQTS) shared the same CNV, while the father’s MLPA pattern was normal. The deletion of exons 7 and 8 of *KCNQ1* was also confirmed in the brother and the mother of the proband at the mRNA level (Fig. 5).

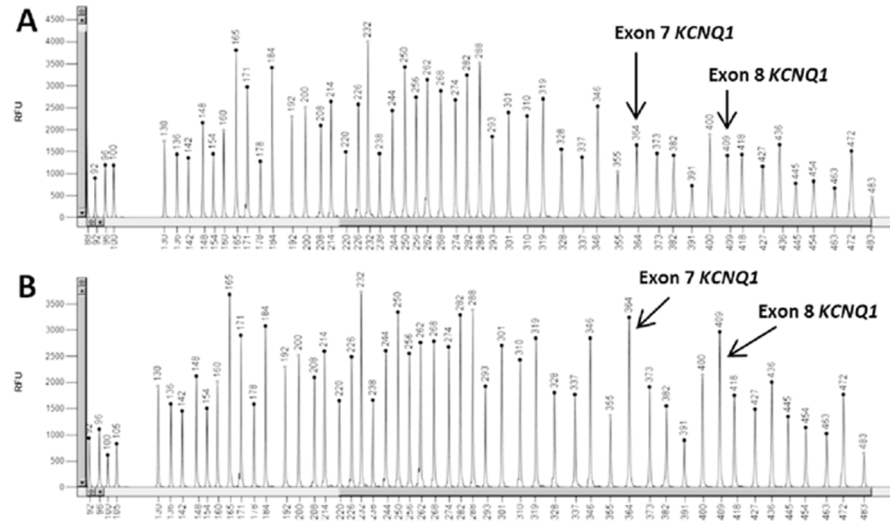


Fig. 4. MLPA capillary electrophoresis pattern. (A) index case and (B) her healthy father, both analysed with SALSA MLPA probemix P114-B2 Long QT. Comparing both profiles, the patient's deletion of exons 7 and 8 of the *KCNQ1* gene can be appreciated.

doi:10.1371/journal.pone.0114894.g004

Discussion

The LQTS is a SCD-related channelopathy of genetic origin. According to current guidelines, when there is a suspicion of LQTS, the genetic analysis using Sanger technology of the three main genes associated with the disease is recommended. It is established that this was a cost-effective approach, until recently, with the advent of NGS technology, which makes the analysis, faster, more extensive and cost effective. NGS data could also be used to analyse CNV alterations, though pipeline bioinformatics analyses are not yet well developed. Thus, to date, few

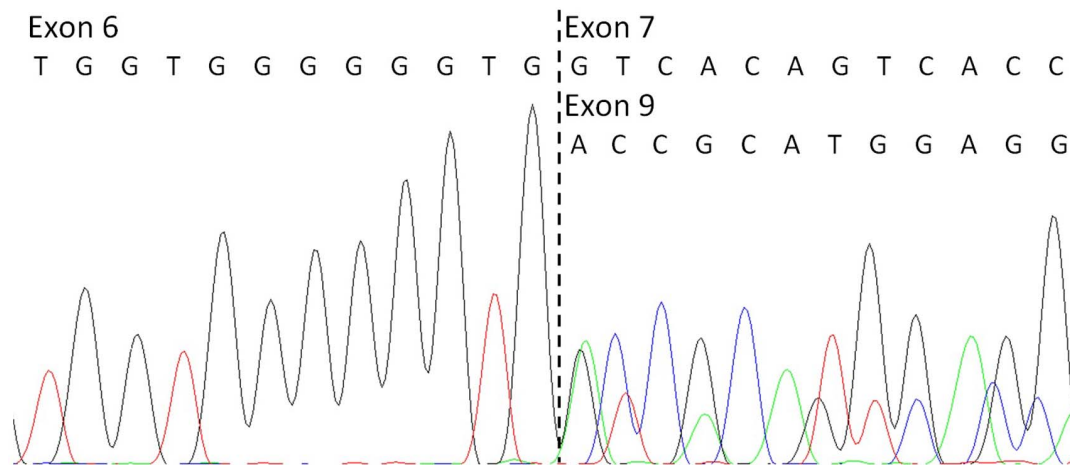


Fig. 5. Partial electropherogram of the sequence of the cDNA of the proband's affected brother. It confirms the deletion of exons 7 and 8 of *KCNQ1* at mRNA level.

doi:10.1371/journal.pone.0114894.g005

reports showing CNV in LQTS families have been published. We performed a thorough analysis covering all exons, and utilizing normalized data. Our novel approach revealed a deletion of exons 7 and 8 in the *KCNQ1* gene. After deep analysis of the protein structure, the deletion was considered as probably pathogenic. In our family, there was complete cosegregation of LQTS phenotype with the *KCNQ1* deletion, and also complete penetrance. This CNV, confirmed by MLPA method and at mRNA level, is considered extremely rare, as overlapping deletions have only been described in one work based on Asian population, and with a frequency of 0.12% [13]. A very similar CNV was previously reported by Barc et al. [6] in a patient with LQTS and without single nucleotide mutations in genes *KCNQ1*, *KCNH2* and *SCN5A*. In that family, the deletion was also identified in the proband's father, who had an undetermined phenotype. This fact may be due to the incomplete penetrance often observed in LQTS families. Other CNVs within or including the *KCNQ1* gene have also been described in LQTS patients [4–6, 14]. All together, these results suggest the deletion of exons 7 and 8 in gene *KCNQ1* may be the cause of the LQTS in our family. CNVs in the *KCNH2* gene have also been reported in association with LQTS [4, 6, 7, 15–17]. Considering previously published series CNVs in *KCNQ1* and *KCNH2*, account for 2–11.5% of LQTS cases [4–7]. This percentage seems to be higher than the frequency of single nucleotide pathogenic variants in minor genes related to LQTS.

In addition, after NGS analysis, we identified a novel genetic variation in titin protein (p.R20729G_ *TTN*) not reported in international databases, so far. The *TTN* variant was predicted as pathogenic by *in silico* tools, alignment showed high conservation between species, and aminoacid change confirms a substitution of R (Arg –polar with positive charge-) to G (Gly –polar without charge-). All these facts suggest a potentially pathogenic role. Genetic studies using NGS technology reveals much higher prevalence of previously *TTN*-associated variants, disputing their possible causality [18]. Hence, recent studies recommend the use of several genetic tools in order to clarify its role in causing the disease, especially for clinical diagnosis [11, 12]. Though no clinical association between any structural gene and LQTS has been yet identified to our knowledge. Especially important was the fact that the variation did not segregate with the affected family members; two LQTS affected members did not carry the genetic variation. This fact confirmed that this novel variation could be discarded as a potential cause of LQTS, at least in our family. This reinforces the importance of family segregation in clinical genetics. If not available, the role of a GVUS in causing disease should be taken with great caution.

Our index case and family members diagnosed by LQTS were placed under beta-blockers, recommended exercise restriction, and provided with a list of QT prolonging drugs list, following current guidelines [9]. In these recommendations, genetic analysis is considered one of the parameters to consider in clinical diagnosis, only when a pathogenic mutation has been identified.

In summary, in familial LQTS, despite that current clinical guidelines recommend genetic analysis restricted to the main genes associated with LQTS, we

provide the evidence that NGS technology can be used efficiently to analyse the rest of the genes associated with the disease. Phenotype interpretation of all these variants remains as the main challenge for its clinical translation. Despite several bioinformatic tools helps to clarify the role of genetic variants, we consider that family segregation should be the first item to be considered and analysed. Multidisciplinary teams including cardiologist and geneticist specialized in SCD related pathologies are crucial to perform an accurate clinical interpretation of all genetic data obtained, and provide helpful genetic counselling.

Author Contributions

Conceived and designed the experiments: OC GSB CA CFC MC JM JB RB. Performed the experiments: OC GSB CA IM SC MC JM AI. Analyzed the data: OC GSB CA IM SC CFC MC JM AI JB RB. Contributed reagents/materials/analysis tools: OC GSB CFC MC JM JB RB. Contributed to the writing of the manuscript: OC GSB CA IM SC CFC MC JM AI JB RB.

References

1. **Schwartz PJ, Stramba-Badiale M, Crotti L, Pedrazzini M, Besana A, et al.** (2009) Prevalence of the congenital long-QT syndrome. *Circulation* 120: 1761–1767.
2. **Schwartz PJ, Ackerman MJ** (2013) The long QT syndrome: a transatlantic clinical approach to diagnosis and therapy. *Eur Heart J* 34: 3109–3116.
3. **Campuzano O, Beltran-Alvarez P, Iglesias A, Scornik F, Pérez G, et al.** (2010) Genetics and cardiac channelopathies. *Genet Med* 12: 260–267.
4. **Eddy CA, MacCormick JM, Chung SK, Crawford JR, Love DR, et al.** (2008) Identification of large gene deletions and duplications in KCNQ1 and KCNH2 in patients with long QT syndrome. *Heart Rhythm* 5: 1275–1281.
5. **Tester DJ, Benton AJ, Train L, Deal B, Baudhuin LM, et al.** (2010) Prevalence and spectrum of large deletions or duplications in the major long QT syndrome-susceptibility genes and implications for long QT syndrome genetic testing. *Am J Cardiol* 106: 1124–1128.
6. **Barc J, Briec F, Schmitt S, Kyndt F, Le Cunff M, et al.** (2011) Screening for copy number variation in genes associated with the long QT syndrome: clinical relevance. *J Am Coll Cardiol* 57: 40–47.
7. **Stattin EL, Bostrom IM, Winbo A, Cederquist K, Jonasson J, et al.** (2012) Founder mutations characterise the mutation panorama in 200 Swedish index cases referred for Long QT syndrome genetic testing. *BMC cardiovascular disorders* 12: 95.
8. **Darbar D** (2006) Screening for genomic alterations in congenital long QT syndrome. *Heart Rhythm* 3: 56–57.
9. **Priori SG, Wilde AA, Horie M, Cho Y, Behr ER, et al.** (2013) HRS/EHRA/APHRS expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes: document endorsed by HRS, EHRA, and APHRS in May 2013 and by ACCF, AHA, PACES, and AEPC in June 2013. *Heart Rhythm* 10: 1932–1963.
10. **Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER** (2013) The next-generation sequencing revolution and its impact on genomics. *Cell* 155: 27–38.
11. **Facio FM, Lee K, O'Daniel JM** (2013) A Genetic Counselor's Guide to Using Next-Generation Sequencing in Clinical Practice. *J Genet Couns* 23: 455–62.
12. **Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, et al.** (2013) A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet* 84: 453–463.

13. **Xu H, Poh WT, Sim X, Ong RT, Suo C, et al.** (2011) SgD-CNV, a database for common and rare copy number variants in three Asian populations. *Hum Mutat* 32: 1341–1349.
14. **Gurrieri F, Zollino M, Oliva A, Pascali V, Orteschi D, et al.** (2013) Mild Beckwith-Wiedemann and severe long-QT syndrome due to deletion of the imprinting center 2 on chromosome 11p. *Eur J Hum Genet* 21: 965–969.
15. **Bisgaard AM, Rackauskaite G, Thelle T, Kirchhoff M, Bryndorf T** (2006) Twins with mental retardation and an interstitial deletion 7q34q36.2 leading to the diagnosis of long QT syndrome. *American journal of medical genetics. Part A* 140: 644–648.
16. **Koopmann TT, Alders M, Jongbloed RJ, Guerrero S, Mannens MM, et al.** (2006) Long QT syndrome caused by a large duplication in the KCNH2 (HERG) gene undetectable by current polymerase chain reaction-based exon-scanning methodologies. *Heart Rhythm* 3: 52–55.
17. **Caselli R, Mencarelli MA, Papa FT, Ariani F, Longo I, et al.** (2008) Delineation of the phenotype associated with 7q36.1q36.2 deletion: long QT syndrome, renal hypoplasia and mental retardation. *American journal of medical genetics. Part A* 146A: 1195–1199.
18. **Refsgaard L, Holst AG, Sadjadieh G, Ariani F, Longo I, et al.** (2012) High prevalence of genetic variants previously associated with LQT syndrome in new exome data. *Eur J Hum Genet* 20: 905–908.

