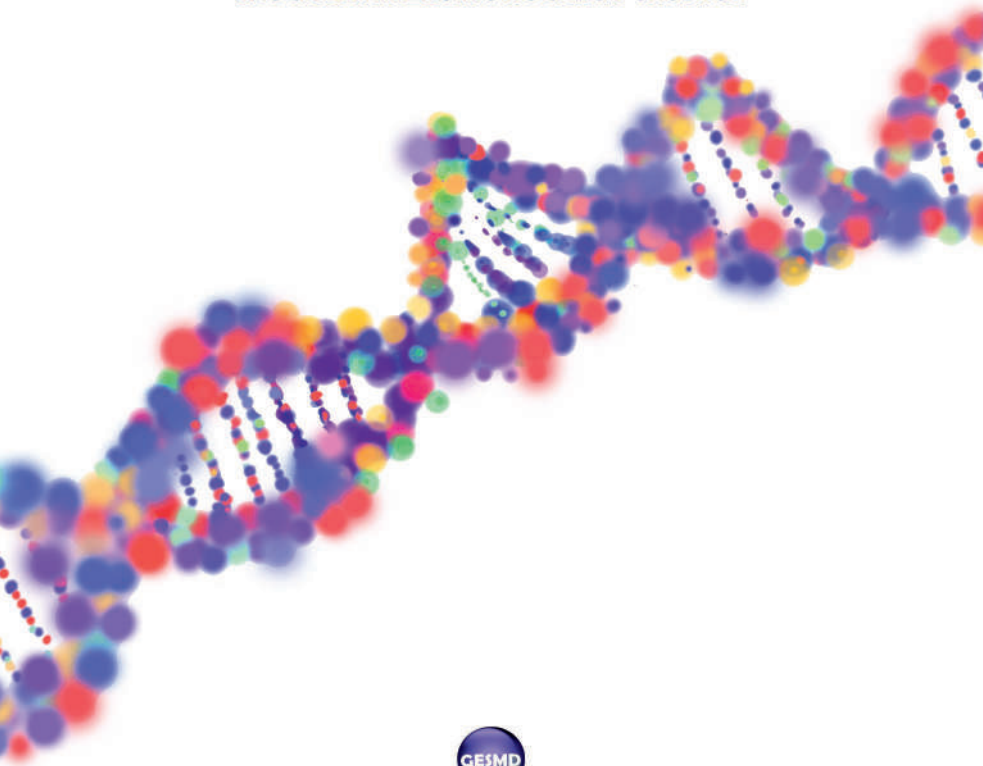


GUÍA DE APLICACIÓN CLÍNICA DE LA SECUENCIACIÓN MASIVA EN SÍNDROMES MIELODISPLÁSICOS Y LEUCEMIA MIELOMONOCÍTICA CRÓNICA



GRUPO ESPAÑOL DE
SÍNDROMES MIELODISPLÁSICOS



GUÍA DE APLICACIÓN CLÍNICA
DE LA SECUENCIACIÓN MASIVA EN
SÍNDROMES MIELODISPLÁSICOS Y
LEUCEMIA MIELOMONOCÍTICA CRÓNICA



GRUPO ESPAÑOL DE
SÍNDROMES MIELODISPLÁSICOS

COORDINADORES

ESPERANZA SUCH
HOSPITAL UNIVERSITARIO Y POLITÉCNICO LA FE (VALENCIA)

FRANCESC SOLÉ
INSTITUT DE RECERCA CONTRA LA LEUCÈMIA JOSEP CARRERAS (BARCELONA)

COLABORADORES

CIMA LAB DIAGNOSTICS (PAMPLONA)
Iria Vázquez, Marta Fernandez-Mercado, María José Larráyo, María José Calasanz

HOSPITAL UNIVERSITARIO Y POLITÉCNICO LA FE (VALENCIA)
Esperanza Such, Mariam Ibáñez, José Cervera, Guillermo Sanz

HOSPITAL UNIVERSITARIO DE SALAMANCA (SALAMANCA)
María Abáigar, Rocío Benito, María Díez Campelo, Jesús María Hernández-Sánchez,
Jesús Hernández-Rivas

HOSPITAL UNIVERSITARI VALL D'HEBRON (BARCELONA)
Barbara Tazón, David Valcárcel

HOSPITAL UNIVERSITARIO 12 OCTUBRE (MADRID)
Inmaculada Rapado, Joaquín Martínez

INSTITUT CATALÀ D'ONCOLOGIA - HOSPITAL UNIVERSITARI GERMANS TRIAS I PUJOL
(BADALONA)
Marta Cabezón, Lurdes Zamora

INSTITUT DE RECERCA CONTRA LA LEUCÈMIA JOSEP CARRERAS (BARCELONA)
Laura Palomo, Francisco Fuster, Pamela Acha, Francesc Solé

NIMGENETICS (MADRID)
Sara Álvarez, Paolo Maietta, Juan Cruz Cigudosa

CONTENIDOS

ABREVIATURAS Y DEFINICIONES 9

INTRODUCCIÓN 16

La secuenciación masiva de los
SMD y LMMC en la práctica clínica 17
Referencias 19

GENES: INCIDENCIA Y PRONÓSTICO 20

Factores de splicing 24
Reguladores epigenéticos 27
Cohesinas 32
Factores de transcripción 33
Transducción de señales (vías de señalización) 36
Reparación del daño celular 37
Variantes germinales de predisposición y
hematopoyesis clonal en edad avanzada 40
Referencias 44

PLATAFORMAS Y CRITERIOS MÍNIMOS 50

Plataformas 51
Diseño del panel de NGS 58
Preparación de librerías de NGS 64
Validación de paneles de NGS 75
Referencias 78

ANÁLISIS Y FILTRADO DE VARIANTES 80

Secuenciación 81
Análisis de las muestras 86
Filtrado de variantes 97
Referencias 100

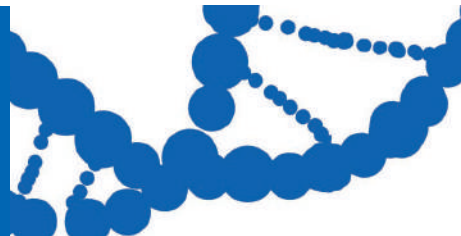
INTERPRETACIÓN DE VARIANTES 102

Bases de datos clínicas 103
Algoritmos de predicción 105
Criterios para la categorización de variantes 107
Clasificación de las variantes 109
Referencias 112

ELABORACIÓN DE INFORMES 114

Información recogida en el informe de NGS 115
Modelo de informe de NGS consensuado por el GESMD 118

RECURSOS WEB 125



ABREVIATURAS Y DEFINICIONES

A

Adaptador: secuencias cortas de DNA (oligonucleótidos) que se añaden a los extremos de cada fragmento de DNA durante el proceso de preparación de las librerías. Estas secuencias son complementarias con la plataforma de secuenciación empleada.

Amplicón: segmento de DNA o RNA que resulta de la amplificación (producción de múltiples copias) de una secuencia de interés.

Alelo: cada una de las formas alternativas que puede tener un gen localizado en una posición específica (*locus*) de un cromosoma determinado.

AREB: anemia refractaria con exceso de blastos.

ARSA-T: anemia refractaria con sideroblastos en anillo en transformación.

AT: hace referencia a las bases nitrogenadas Adenina y Timina.

B

Base calling: es el proceso de asignación de bases (nucleótidos) a cada señal detectada durante una reacción de secuenciación. Típicamente, cada secuenciador tiene integrado su propio *software* (conjunto de programas) con el cual se lleva a cabo este proceso.

.bed: tipo de formato de archivo que contiene coordenadas genómicas, además de información asociada a las mismas (opcional). Más información sobre la estructura y función de los ficheros .bed en Tabla 1 y Tabla 17.

C

Carrera (run): proceso completo de secuenciación que genera información digital de las secuencias de muestras de DNA o RNA. Comprende todos los pasos que ocurren dentro del secuenciador, comúnmente, desde la amplificación hasta la *base calling*.

CASAVA (Consensus Assessment of Sequence And Variation): parte del *software* de análisis de Illumina que se encarga del alineamiento, análisis de variantes y contaje de lecturas.

Clonal Cytopenia of Undetermined Significance (CCUS): citopenias clonales de significado incierto.

Clonal Hematopoiesis of Intermediate potential (CHIP): hematopoyesis clonal de potencial indeterminado.

Cluster: resultado de la amplificación clonal de cada fragmento de la librería. Mediante sucesivas reacciones de amplificación, a partir de un único fragmento se genera un conjunto de fragmentos idénticos (*clusters*) que luego serán secuenciados.

CMML-Specific Prognostic Scoring System (CPSS): *score* pronóstico específico para la leucemia mielomonocítica crónica.

CRDM: citopenia refractaria con displasia multilínea.

Cyclic Reversible Termination (CRT): secuenciación por síntesis mediante terminación reversible cíclica.

D

Delección: tipo de alteración genética que consiste en la pérdida de un fragmento de DNA dentro de un cromosoma.

DNA: ácido desoxirribonucleico.

E

Edge effect: efecto que hace que los *variant callers* detecten variantes en los extremos de las lecturas. Este efecto podría deberse a que pueden quedar restos de adaptadores dentro de las lecturas al finalizar el proceso de secuenciación. Durante el análisis de las secuencias, estos restos de adaptadores son considerados erróneamente como variantes.

F

Flow cell: soporte físico donde ocurre la formación de *clusters* para la posterior secuenciación en los equipos de Illumina. Es un soporte de vidrio cuya superficie está cubierta con oligonucleótidos complementarios a los adaptadores de las librerías que los contienen, permitiendo que éstas se unan y luego se formen los *clusters*.

Frameshift: tipo de variante que genera un desplazamiento o cambio en el marco de lectura. Ocurre por una inserción o delección de un número de nucleótidos que no es múltiplo de tres en una secuencia de DNA.

G

GC: hace referencia a las bases nitrogenadas Guanina y Citosina.

Genome Browser: navegador del genoma. Son herramientas, normalmente online, que ofrecen acceso a la secuencia del genoma de una o varias especies, junto con información asociada al mismo (anotaciones). Entre otras características, permiten visualizar, examinar y consultar datos relacionados con dicho genoma.

gzip: formato de compresión informática usado comúnmente en secuenciación. Codifica la información de uno o varios archivos para que ocupen menos espacio en disco.

H

Hardware: partes físicas de un sistema informático. Se engloban los componentes eléctricos, electrónicos, mecánicos, electromecánicos, sensores, periféricos, cables, cajas y cualquier otro tipo de elemento físico.

Hematopoietic Stem Cells (HSC): células madre hematopoyéticas.

Heterocigota: al referirse a una variante, implica que la misma está presente solo en uno de los dos alelos siendo el otro alelo normal.

Homopolímeros: secuencias que contienen varios nucleótidos iguales.

Hotspot: zona dentro del genoma propensa a ser alterada y en la cual se detectan variantes más frecuentemente. Dicha región puede comprender un solo nucleótido, un codón o un exón.

I

Idiopathic cytopenia of undetermined significance (ICUS): citopenia idiopática de significado incierto.

Indel: alteración genética en la cual se produce una inserción o una delección.

Índice (barcode): secuencias cortas de DNA (oligonucleótidos), específicas para cada muestra, que se añaden a los extremos de cada fragmento de DNA durante el proceso de preparación de las librerías. Gracias a ellos, es posible mezclar (multiplexar) y secuenciar en simultáneo muestras diferentes sin perder la identidad de cada una, ya que permiten identificar cada muestra durante el proceso de secuenciación.

Inserción: tipo de alteración genética que consiste en la ganancia de un fragmento de DNA dentro de un cromosoma.

IPSS-R: *revised international prognostic scoring system*.

L

Lectura (read): cada una de las secuencias que se generan durante el proceso de secuenciación.

Librerías: fragmentos de DNA de una muestra a secuenciar, de un determinado tamaño, que en sus extremos contienen los índices y los adaptadores. En la secuenciación dirigida, estos fragmentos corresponden a las regiones de interés.

LMA: leucemia mieloide aguda.

LMC: leucemia mielomonocítica crónica.

M

Mapear o alinear: proceso de identificar las localizaciones de las lecturas que han sido secuenciadas previamente, usando para ello un genoma de referencia.

Minor Allele Frequency (MAF): mínima frecuencia alélica reportada en las bases de datos poblacionales para una variante. Dependiendo del porcentaje, se puede considerar la variante como un polimorfismo (SNP).

Mismatch: en secuenciación, se refiere a la discrepancia de secuencias entre las lecturas y el genoma de referencia. Se detecta durante el *variant calling*, después de haber alineado la secuencia de la muestra que se está analizando frente al genoma de referencia.

Missense: tipo de variante donde el cambio de un solo nucleótido resulta en el cambio de un codón que produce un aminoácido diferente.

Mtase: metiltransferasa.

Multiplexar: o “hacer un *pool* de muestras”, consiste en agrupar varias muestras en un solo ensayo. El uso de índices posibilita mezclar varias muestras y secuenciarlas en una sola carrera, lo que permite reducir el costo de secuenciación por muestra.

N

NMP: neoplasias mieloproliferativas.

Nonsense: tipo de variante en un solo nucleótido que resulta en un codón de *stop* prematuro o en un codón que después de la traducción genera una proteína no funcional.

O

Off-target: secuencias que alinear fuera de la región de interés.

Oligonucleótidos: secuencia corta de DNA o RNA de aproximadamente 50 pares de bases. En secuenciación se utilizan como *primers* en las reacciones de amplificación.

P

Pipeline: conjunto de programas que se usan para realizar un determinado análisis bioinformático. Se suelen recoger los datos que un programa muestra como salida y usarlos como entrada del programa siguiente.

Polymerase Chain Reaction (PCR): reacción en cadena de la polimerasa. Técnica de biología molecular cuyo objetivo es obtener un gran número de copias de un fragmento de DNA particular.

PRC2: complejo represivo de *polycomb 2*.

Predictor *in silico*: programa informático que determina, mediante estudios estadísticos y estructurales, la posibilidad de que una variante genética pueda afectar a la estructura o función de la proteína o al producto resultante que contiene dicha variante.

R

Reguladores epigenéticos: elementos que contribuyen y regulan el conjunto de reacciones químicas y procesos que modifican la actividad del DNA pero sin alterar su secuencia.

S

Sanger: método de secuenciación del DNA, también referido como método clásico de terminación de la cadena. Se basa en la polimerización del DNA y el uso de didesoxinucleótidos que sirven como terminadores de la reacción. Cada didesoxinucleótido (dATP, dCTP, dGTP, dTTP) implica un color de fluoróforo diferente, lo que permite su identificación.

Secuenciación dirigida (*targeted sequencing*): técnica que consiste en la secuenciación de un conjunto de regiones previamente seleccionadas del genoma.

Single-Nucleotide Addition (SNA): secuenciación por síntesis mediante adición de un solo nucleótido.

Single Nucleotide Polymorphism (SNP): variante que difiere en un nucleótido respecto al genoma de referencia (SNV) y que está presente en $\geq 1\%$ de la población (polimorfismo).

Single Nucleotide Variant (SNV): variante consistente en el cambio de un solo nucleótido respecto del genoma de referencia. Una SNV puede ser sinónima o no (*missense*).

SMD: síndromes mielodisplásicos.

Software: conjunto de programas que contiene un aparato electrónico que permite usar los componentes del *hardware* de un modo determinado.

Spliceosoma: complejo de ribonucleoproteínas capaces de llevar a cabo el proceso de *splicing*.

Splicing: proceso co-transcripcional conocido como “corte y empalme”. Una vez que se ha sintetizado la molécula de ARNm, es el proceso mediante el cual se eliminan los intrones, dejando solo los exones para formar una cadena de ARNm maduro que, posteriormente, se transportará al citoplasma para su traducción.

Strand-bias: sesgo en el que se ve favorecida una de las cadenas (sentido o anti-sentido) del DNA, lo que genera una distribución diferente de una respecto a la otra. En este contexto, implica que al secuenciar se encontrarán más lecturas de una cadena que de otra.

U

UCSC Genome BrowserHome:

Unique Molecular Identifier (UMI): secuencias cortas de DNA (oligonucleótidos) que se añaden a los extremos de cada fragmento de DNA durante el proceso de preparación de las librerías. Permite identificar y eliminar durante el análisis los duplicados de PCR generados en la preparación de las librerías.

V

Variant allele frequency (VAF): frecuencia alélica de una variante. En secuenciación, hace referencia a la proporción de lecturas de la región de interés que contienen la secuencia alternativa (variante) respecto de las lecturas que contienen la secuencia de referencia.

Variant caller: programa que permite realizar el *variant calling* a partir de una secuencia de DNA previamente alineada frente al genoma de referencia.

Variant calling: es un paso en el análisis de los resultados de secuenciación que consiste en la detección de aquellas variantes (cambios) presentes en la secuencia de DNA que se está analizando respecto de la secuencia del genoma de referencia.

Variante sinónima: cambio en la secuencia nucleotídica que después del proceso de traducción no deriva en el cambio de la secuencia de aminoácidos de la cadena polipeptídica, por lo que no se generan cambios estructurales ni funcionales de la proteína.

W

Whole Genome Sequencing (WGS) - secuenciación del genoma: técnica que consiste en la determinación de la secuencia del genoma completo de un organismo incluyendo, además del DNA cromosómico, el DNA mitocondrial.

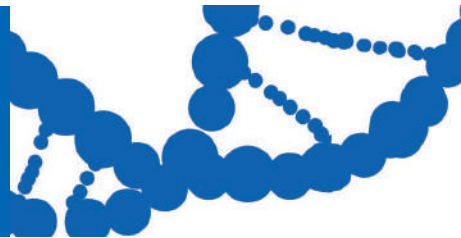
Whole Exome Sequencing (WES) - secuenciación del exoma: técnica que consiste en la determinación de la secuencia del exoma, es decir, solo la parte codificante del DNA genómico.



INTRODUCCIÓN

GUÍA DE APLICACIÓN CLÍNICA DE LA SECUENCIACIÓN MASIVA EN SÍNDROMES MIELODISPLÁSICOS Y LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:
Esperanza Such y Francesc Solé.



Los síndromes mielodisplásicos (SMD) son por naturaleza heterogéneos y diversos, como resultado de su compleja fisiopatología, en la que se reconocen desde fenómenos de disregulación epigenética a profundas alteraciones mutacionales. Es bien sabido que se trata de un complejo grupo de neoplasias hematológicas de las células progenitoras hematopoyéticas que comparten como características comunes la presencia de citopenias y una tendencia variable a evolucionar a leucemia mieloblástica aguda⁽¹⁾, pero también son una de las neoplasias hematológicas más frecuentes en personas de edad avanzada, con una mediana de presentación de 70 años, y un 25% de los pacientes diagnosticados con más de 80 años⁽¹⁾.

La leucemia mielomonocítica crónica (LMMC) es un trastorno hematológico clonal poco frecuente que suele afectar a pacientes de edad avanzada y que tiene una expresión clínica y morfológica muy variable. Se caracteriza por un incremento de células de estirpe monocítica tanto en médula ósea como en sangre periférica. Su forma de presentación puede oscilar desde un cuadro de hematopoyesis ineficaz con una médula ósea hiperclular, citopenias en sangre periférica y rasgos displásicos en una o más de las líneas hematopoyéticas, hasta en un cuadro medular mieloproliferativo, por ello, se engloba dentro de las neoplasias mixtas mielodisplásicas/mieloproliferativas⁽¹⁾.

En la última década, la secuenciación del DNA ha revolucionado nuestro entendimiento de la patogénesis de distintas enfermedades con base genética como son los SMD y la LMMC, estableciendo que estas enfermedades surgen tras la adquisición seriada de mutaciones somáticas en genes implicados en numerosas funciones y mecanismos. Esta acumulación secuencial de mutaciones rige la evolución de la enfermedad desde su inicio, como la hematopoyesis clonal asintomática, a estadios establecidos como SMD o LMMC y, en última instancia, podría provocar la evolución hacia una LMA⁽²⁾.

La aplicación de las técnicas de secuenciación masiva en el estudio de los SMD y la LMMC, ha permitido la identificación de mutaciones genéticas recurrentes en más de 40 genes. En el momento actual, se estima que la mayoría de los pacientes (80-90%) presenta al diagnóstico al menos una mutación en algún gen. Esta información es relevante ya que los marcadores moleculares

pueden contribuir a una mejor caracterización y estratificación del riesgo de estos pacientes⁽³⁾. Los genes más frecuentemente mutados en los SMD y la LMMC son *TET2*, *SF3B1*, *ASXL1*, *DNMT3A*, *SRSF2*, *RUNX1*, *TP53*, *U2AF1*, *EZH2*, *ZRSR2*, *STAG2*, *CBL*, *NRAS*, *JAK2*, *SETBP1*, *IDH1*, *IDH2* y *ETV6*⁽³⁻⁷⁾. A diferencia de otras patologías, como por ejemplo las neoplasias mieloproliferativas crónicas (NMP) con mutaciones específicas en los genes *JAK2*, *CALR* y *MPL*, no hay ninguna mutación que sea dominante y específica de los SMD⁽⁸⁾ y la LMMC. Sin embargo, algunas de las mutaciones que presentan han demostrado tener valor diagnóstico, al estar asociadas con subtipos concretos de SMD; o valor pronóstico, ya que se han asociado con una supervivencia más corta y un mayor riesgo de transformación a LMA⁽⁵⁾; o bien, actúan como biomarcadores de respuesta a determinados tratamientos⁽⁹⁾. Se han descrito, por ejemplo, mutaciones en *TET2* en el 20-25% de los SMD⁽⁶⁾ y el 45-60% de las LMMC⁽⁷⁾ y mutaciones en *SF3B1* en el 20-30% de los SMD alcanzando una incidencia del 80% de los SMD que presentan sideroblastos en anillo (SA)⁽⁸⁾. Además, algunas mutaciones menos frecuentes se han asociado con características clínicas adversas tales como los cariotipos complejos (con una alta frecuencia de mutaciones en *TP53*), el exceso de blastos en médula ósea (mayor frecuencia de mutaciones en *RUNX1*, *NRAS* y *TP53*) y la trombocitopenia severa (presentan mutaciones en *RUNX1*, *NRAS* y *TP53*), o se ha visto que predicen una supervivencia más corta y un mayor riesgo a desarrollar una LMA (*ASXL1*, *ETV6*, *EZH2*, *RUNX1*, *TP53*)⁽¹⁰⁾. Por ello, actualmente podemos entender que estas mutaciones están relacionadas con características fisiopatológicas de los SMD y LMMC (SA, citopenias, monocitosis y alteraciones cromosómicas), y que podrían, en parte, estar contribuyendo a la heterogeneidad clínica característica de estas enfermedades y, por tanto, influir en el pronóstico de los pacientes.

En el momento actual, la caracterización genómica está complementando las clasificaciones diagnósticas existentes consiguiendo una mejora en la estratificación de los pacientes. Como consecuencia de ello, muchos centros han introducido los estudios genómicos mediante secuenciación masiva en su práctica clínica diaria, en un esfuerzo por individualizar el diagnóstico, pronóstico y tratamiento de los pacientes con SMD y LMMC. En esta guía se detallan los genes que se recomienda evaluar en todos los pacientes con sospecha diagnóstica de SMD y LMMC, según el Grupo Español de Síndromes Mielodisplásicos (GESMD), y se establecen los criterios mínimos de estandarización y calidad que se deben emplear para un correcto diagnóstico molecular de los SMD.

REFERENCIAS

1. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
2. Kennedy JA, Ebert BL. Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. *J Clin Oncol*. 2017;35(9):968-974. doi:10.1200/JCO.2016.71.0806.
3. Haferlach T, Nagata Y, Grossmann V, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*. 2014;28(2):241-247.
4. Bejar R, Stevenson K, Abdel-Wahab O, et al. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*. 2011;364(26):2496-2506.
5. Bejar R, Stevenson KE, Caughey BA, et al. Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. *J Clin Oncol*. 2012;30(27):3376-3382.
6. Papaemmanuil E, Gerstung M, Malcovati L, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616-27; quiz 3699.
7. Elena C, Galli A, Such E, et al. Integrating clinical features and genetic lesions in the risk assessment of patients with chronic myelomonocytic leukemia. *Blood*. 2016;128(10):1408-1417.
8. Papaemmanuil E, Cazzola M, Boultonwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384-1395.
9. Bejar R, Stevenson KE, Caughey B, et al. Somatic mutations predict poor outcome in patients with myelodysplastic syndrome after hematopoietic stem-cell transplantation. *J Clin Oncol*. 2014;32(25):2691-2698.
10. Kennedy JA, Ebert BL. Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. *J Clin Oncol*. 2017 Mar 20;35(9):968-974.

GENES: INCIDENCIA Y PRONÓSTICO

GUÍA DE APLICACIÓN CLÍNICA DE
LA SECUENCIACIÓN MASIVA EN
SÍNDROMES MIELODISPLÁSICOS Y
LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:

Lurdes Zamora, Mariam Ibáñez, Laura Palomo, María Abáigar, Iria Vázquez, Bárbara Tazón-Vega, Pamela Acha, Rocío Benito, María José Larráyo, Marta Cabezón, Juan Cruz Cigudosa, Francesc Solé y Esperanza Such.

GENES: INCIDENCIA Y PRONÓSTICO

Como se ha mencionado anteriormente, se han identificado mutaciones en más de 40 genes implicados en el desarrollo de los SMD y la LMMC, aunque ninguna sea dominante y específica de estas enfermedades. Estos genes pueden agruparse en diferentes categorías funcionales según el proceso celular en el que se encuentran involucrados: factores de *splicing*, reguladores epigenéticos, componentes de cohesinas, factores de transcripción, reguladores de la respuesta a las lesiones provocadas en el DNA y moléculas implicadas en la transducción de señales entre otros⁽¹⁾ (Figura 1, Tabla 1). Los genes recomendados por el GESMD que son necesarios estudiar para el diagnóstico diferencial y manejo de los SMD se detallan en la Tabla 1.

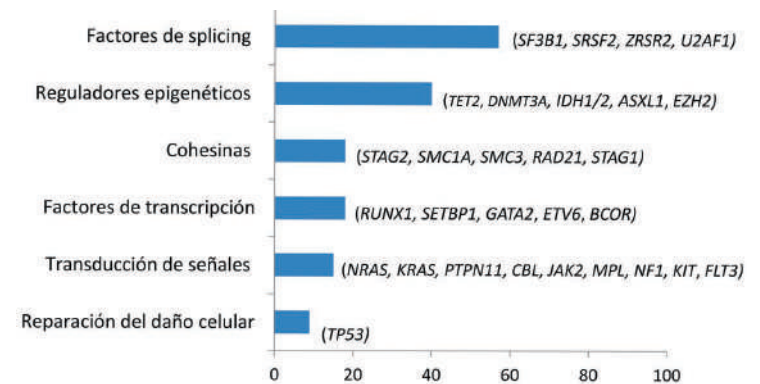


Figura 1. Distribución de los genes más frecuentemente mutados en los pacientes con SMD y LMMC en base a su categoría funcional (modificada de Kennedy and Ebert, 2017)⁽¹⁾.

Tabla 1. Descripción de los genes recomendados por el GESMD que es necesario estudiar para el manejo de los pacientes con SMD y LMMC (modificada de Greenberg et al., 2017)⁽²⁾.

GEN	REGIÓN ESTUDIADA MUTACIONES MÁS FRECUENTES	TIPO DE MUTACIÓN	INCIDENCIA	
			SMD	LMMC
Frecuentes en síndromes mielodisplásicos				
			SMD	LMMC
ASXL1	Exones 10-12 Codones: todos	<i>Nonsense, frameshift</i>	5-25%	40-50%
CSNK1A1	Exones 3 y 4 Codones: E98, H134, D140	<i>Missense</i>	5-10%	<1%
DNMT3A	Gen completo Codón: R882	<i>Todos</i>	12-18%	2-10%
EZH2	Gen completo Codones: todos	<i>Nonsense, frameshift</i>	5-10%	5-12%
IDH1	Exón 4 Codón: R132	<i>Missense</i>	<5%	<1%
IDH2	Exón 4 Codones: R140 y R172	<i>Missense</i>	<5%	5-10%
JAK2	Gen completo Codón: V617F	<i>Missense</i>	<5%	2-10%
KRAS	Exones 2 y 3 Codones: G12, G13, Q61 y G146	<i>Missense</i>	5-10%	10-20%
NRAS	Exones 2 y 3 Codones: G12, G13 y Q61	<i>Missense</i>	5-10%	10-20%
RUNX1	Gen completo Codones: todos	<i>Nonsense, frameshift</i>	10-15%	10-30%
SETBP1	Exón 4 Codones: 858-870	<i>Missense</i>	<5%	5-10%
SF3B1	Exones 10-16 Codones: 622-781	<i>Missense</i>	20-30% 80% con SA	5-10%
SRSF2	Gen completo Codón: P95	<i>Missense</i>	10-15%	30-50%
STAG2	Gen completo Codones: todos	<i>Nonsense, frameshift, splicing</i>	5-10%	5-10%
TET2	Gen completo Codones: 1134-1444 ó 1842-1921	<i>Todos</i>	20-25%	45-60%
TP53	Gen completo Codones: todos (excepto P47S y P72R)	<i>Todos</i>	8-12%	<5%
U2AF1	Exones 2-6 Codones: S34 y Q157	<i>Missense</i>	8-12%	5-10%
ZRSR2	Gen completo Codones: todos	<i>Nonsense, frameshift</i>	5-10%	5-10%

Continúa...

...Tabla 1 Continúa

GEN	REGIÓN ESTUDIADA MUTACIONES MÁS FRECUENTES	TIPO DE MUTACIÓN	INCIDENCIA	
			SMD	LMMC
Frecuentes en neoplasias mieloproliferativas				
			SMD	LMMC
CALR	Exón 9 Codones: todos	<i>Frameshift</i>	<1%	<1%
CBL	Exones 8 y 9 Codones: 366-420	<i>Missense</i>	<5%	8-18%
CSF3R	Gen completo Codones: I598 y T618	<i>Missense</i>	<1%	3-4%
GATA2	Exones 2 y 6 Codones: todos	<i>Missense, frameshift</i>	<5%	<1%
MPL	Gen completo Codones S505 y W515	<i>Missense</i>	<1%	<1%
NF1	Gen completo Codones: todos	<i>Nonsense, frameshift, splicing</i>	<5%	<5%
PTPN11	Exones 3 y 7 Codones: todos	<i>Missense</i>	<1%	4%
Frecuentes en leucemia mieloide aguda				
			SMD	LMMC
BCOR	Gen completo Codón N1425	<i>Todos</i>	<5%	<5%
BCORL1	Gen completo Codones: todos	<i>Nonsense, frameshift</i>	<1%	<1%
CEBPA	Gen completo Codones: todos	<i>Missense, frameshift</i>	<5%	<5%
ETV6	Gen completo Codones: todos	<i>Nonsense, frameshift</i>	<5%	<1%
FLT3	Exones 13-15 y 20 Codones: FLT3-ITD y D835	<i>Missense, frameshift</i>	<5%	<5%
KIT	Exones 2, 8-11, 13 y 17 Codones: todos	<i>Missense, frameshift</i>	<3%	<1%
NPM1	Exones 10 y 11 Codones 956-959	<i>Frameshift</i>	<5%	<5%
WT1	Exones 7 y 9 Codones: todos	<i>Missense, frameshift</i>	<3%	<3%

LMMC: leucemia mielomonocítica crónica; SA: sideroblastos en anillo; SMD: síndromes mielodisplásicos.

FACTORES DE SPLICING

El *spliceosoma* es un complejo proteico y dinámico que participa en la eliminación de intrones (*splicing*) para generar un mRNA maduro. Son muchas las subunidades y moléculas que participan en este proceso de escisión de intrones para asegurar que se produce de forma adecuada. Las mutaciones en los genes que codifican este complejo provocarían una alteración en los patrones de *splicing*, dando lugar a la incorporación inadecuada de intrones o exones en el RNA específico, pudiendo así modificar el producto proteico final⁽³⁾. Esto podría alterar aún más la capacidad de la célula para diferenciarse en células sanguíneas completamente maduras dando lugar a un fenotipo displásico⁽⁴⁾.

Se han descrito mutaciones frecuentes, tanto en los SMD como en la LMMC, en varios genes que codifican algunas de las subunidades proteicas que forman el complejo del *spliceosoma*, como *SF3B1*, *U2AF1* (*U2AF35*), *SRSF2* y *ZRSR2*. También se han documentado mutaciones en otros factores de *splicing* como *PRPF40B*, *SF3A1*, *SF1* y *U2AF2* (*U2AF65*), aunque con una frecuencia muy baja. Tomadas en conjunto, las mutaciones más frecuentes en los SMD son las que afectan a genes del *splicing* (45-85%)⁽³⁾.

Las mutaciones descritas en estos genes son mayoritariamente heterocigóticas, de tipo *missense*, e implican sustituciones de un único aminoácido, se localizan en regiones *hotspot* definidas⁽³⁻⁵⁾ (Tabla 1) y generalmente son mutuamente excluyentes, sugiriendo que las células no pueden tolerar dos mutaciones en el mismo gen complejo, ya que estos cambios tienen un papel redundante en la patogénesis de la enfermedad⁽³⁾.

SF3B1 (2q33.1)

El gen *SF3B1* codifica la subunidad 1 del complejo de *splicing* 3b, que es un componente central de la snRNP U2 que reconoce el sitio 3' de ensamblaje en las uniones intrón-exón.

Los pacientes con SMD y mutaciones en *SF3B1* presentan mayor número de sideroblastos en anillo y, a menudo, tienen un recuento de plaquetas más alto, menos citopenias, y menor porcentaje de blastos comparado con los pacientes sin mutaciones en este gen⁽⁶⁾.

Se encuentra mutado en el 60-80% de los pacientes con SMD con sideroblastos en anillo y en el 20-30% de los SMD en general⁽⁶⁻⁷⁾. La mayoría de las mutaciones se agrupan en los exones 14 y 15 (aminoácidos 622-700), siendo la más frecuente la mutación K700E⁽⁶⁾ (Tabla 1).

De todas las alteraciones descritas en el *spliceosoma*, solo las mutaciones en *SF3B1* han sido relacionadas con un pronóstico favorable en la mayoría de los estudios realizados, asociándose, además, con la presencia de sideroblastos en anillo⁽⁶⁻⁶⁾. Múltiples estudios han descrito que los pacientes con ARSA con *SF3B1*-mutado podrían tener mejor pronóstico que el mismo subtipo de pacientes sin mutaciones de *SF3B1*⁽⁶⁾. Las mutaciones en *SF3B1* están asociadas a una mejor supervivencia global y una menor probabilidad de transformación a LMA⁽⁶⁾ (Tabla 2). Por otro lado, se ha descrito la frecuente coexistencia de mutaciones de *SF3B1* con mutaciones en el gen *DNMT3A*, y algunos estudios han asociado esta combinación a un peor curso clínico de la enfermedad⁽⁸⁾.

SRSF2 (17q25.1)

El gen *SRSF2* codifica para un miembro de la familia *serine/arginine-rich*, crítica para el ensamblaje de pre-mRNA. Se encarga de marcar los límites entre exón/intrón, y presenta la capacidad de redefinir los sitios de *splicing*, lo que lleva a la formación de transcritos alternativos⁽⁹⁾.

Se han descrito mutaciones en *SRSF2*, mayoritariamente en el *hotspot* P95 del exón 1, (Tabla 1) en aproximadamente el 10-15% de los SMD, siendo más frecuentes en los subtipos con displasia multilínea y exceso de blastos⁽⁷⁾, y se han asociado a neutropenia y pronunciada trombocitopenia. Además, el 85% de los pacientes con SMD que tiene mutado *SRSF2* también tienen al menos una mutación adicional en genes como *RUNX1*, *IDH1*, *IDH2*, y *ASXL1*⁽¹⁰⁾.

Estas mutaciones son especialmente prevalentes en la LMMC, detectándose en un 30-50% de los pacientes y suelen coexistir con mutaciones en *TET2*⁽¹¹⁾.

El papel de las mutaciones en *SRSF2* en los SMD no está bien definido y tienen un valor pronóstico controvertido: algunos estudios las han asociado con supervivencias más cortas y mayor riesgo de transformación a LMA⁽¹⁰⁻¹²⁾ sobre todo en la LMMC⁽¹¹⁾, mientras que otros no han encontrado ningún impacto pronóstico⁽¹³⁾ (Tabla 2).

ZRSR2 (Xp22.1)

ZRSR2 codifica para una proteína asociada con el factor auxiliar de U2 heterodímero. Las mutaciones en este gen ocurren en el 5-10% de los SMD y el 5-10% de la LMMC, se distribuyen a lo largo de toda la proteína, y se han asociado a neutropenia aislada⁽⁹⁾. Se ha descrito la frecuente coexistencia de mutaciones de *ZRSR2* con mutaciones del gen *TET2*⁽¹²⁾ (Tabla 2).

U2AF1 (21q22.3)

El gen *U2AF1* (*U2AF35*) codifica para la proteína del factor auxiliar U2, que reconoce el aceptor de dinucleótido AG de ensamblaje en el extremo 3' de los intrones durante el ensamblaje del RNA⁽¹⁴⁾. El análisis de los patrones de ensamblaje en 35 genes diana de *U2AF1* ha revelado que estas mutaciones están asociadas con la desregulación del ciclo celular y el procesamiento de RNA⁽¹⁴⁾.

Existe una mutación *missense* recurrente en *U2AF1* que afecta a la serina del codón 34 (S34) y tiene lugar en el 5-10% de los SMD de *novo*⁽¹⁵⁾ (Tabla 1). Estudios recientes sugieren que la presencia de mutaciones en *U2AF1* puede ser mayor (~20%) en aquellos pacientes con SMD con delección 20q⁽¹⁶⁾. Además, se ha observado también concurrencia de mutaciones con *U2AF1* y con *ASXL1* y *DNMT3A* riesgo⁽¹⁰⁾.

Las mutaciones en *U2AF1* se han relacionado con una menor supervivencia y con un menor tiempo de transformación a LMA en pacientes jóvenes, particularmente en los SMD de bajo riesgo⁽¹⁷⁾ (Tabla 2).

REGULADORES EPIGENÉTICOS

Los mecanismos epigenéticos son aquellos procesos que modifican la actividad del DNA sin alterar su secuencia. Son aquellos involucrados en la metilación del DNA (normalmente de las islas CpG) y los que modifican la función de las histonas, también denominados modificadores de cromatina (por metilación, acetilación y otras modificaciones de residuos).

En los SMD, las alteraciones epigenéticas frecuentemente causan hipermetilación de genes que controlan la proliferación, adhesión y los cambios específicos de la enfermedad en las células madre hematopoyéticas (HSC). Las mutaciones en *DNMT3A*, *TET2* e *IDH1/2* han sido consideradas como *drivers* en la patogénesis de los SMD alterando la metilación del DNA en estos pacientes y siendo, por tanto, potenciales marcadores terapéuticos⁽¹⁸⁻¹⁹⁾.

El 40% de los pacientes con SMD presentan alteraciones en estos genes responsables del proceso de metilación del DNA: *TET2* (15-27%), *DNMT3A* (3-13%) e *IDH1/2* (4-12%)⁽¹⁾. En las LMMC, las mutaciones de *TET2* son más frecuentes que en los SMD, llegando a detectarse en un 40-60% de los pacientes⁽¹⁸⁻¹⁹⁾.

En estos genes se han descrito un amplio rango de mutaciones *missense* (con cambio de aminoácido), *nonsense* (sin sentido), *frameshift* (que alteran el marco de lectura) y en las regiones de *splicing*. Algunas de ellas están localizadas en regiones *hotspot* mientras que otras ocurren a lo largo de todo el gen⁽¹⁸⁻¹⁹⁾.

Otro mecanismo de regulación epigenética es la modificación covalente de las histonas. La acetilación, regulada por histonas acetiltransferasas y deacetilasas, cambia la carga de la cola de la histona haciendo que la arquitectura de la cromatina se relaje (acetilada) o se compacte (deacetilada), promoviendo y reprimiendo la expresión génica respectivamente. Los genes reguladores de las marcas de cromatina que suelen estar mutados en los SMD y la LMMC son *ASXL1* y *EZH2*⁽¹⁸⁾.

Las mutaciones presentes en *ASXL1* y *EZH2* modifican las colas de las histonas y, en consecuencia, provocan una pérdida de función de la proteína⁽¹⁸⁾. Estos genes están mutados en un 5-25% (*ASXL1*) y 5-10% (*EZH2*) de los casos con SMD, siendo más

frecuentes en los subtipos de SMD de las citopenias refractaria con displasia multilineal (CRDM) y de las anemias refractarias con exceso de blastos (AREB). En el caso de la LMMC, la incidencia de mutaciones de *ASXL1* pueden alcanzar el 40-50% de los casos⁽²⁰⁾.

TET2 (4q24)

TET2 (Tet metilcitosina dioxigenasa 2, también conocido como *ten-eleven translocation 2*) es un gen que codifica para una proteína involucrada en la regulación epigenética de la mielopoiesis⁽²¹⁾. La proteína *TET2* cataliza la conversión de 5-metilcitosina en 5-hidroximetilcitosina, hecho que conduce finalmente a desmetilación de dicha citosina, alterando así esta marca regulatoria y está involucrada en la regulación de la metilación del DNA y la expresión génica, inhibiendo la autorenovación.

TET2 es un gen supresor tumoral y, por eso, en cáncer una pérdida de función de *TET2* (que puede suceder vía mutación o delección de *TET2*) puede provocar transformaciones mieloides o linfoides⁽²¹⁾. Las mutaciones de *TET2* (*missense/nonsense/frameshift*) se localizan a lo largo de todo el gen y conllevan una pérdida de función de la capacidad catalítica y, por tanto, un incremento de la metilación. Este incremento en la metilación del DNA provocará un silenciamiento de aquellos genes que tengan la región promotora metilada⁽²¹⁾.

Las mutaciones de *TET2* son frecuentes en muchas enfermedades hematológicas. En los SMD y la LMMC se han descrito en un 20% y 45-60%, respectivamente⁽¹⁾. Se asocian a edad avanzada, hematopoyesis clonal y cariotipo normal, lo que sugiere que la mutación *TET2* es un factor asociado al envejecimiento de las células hematopoyéticas o una mutación iniciadora de procesos malignos⁽¹³⁻²¹⁾. Las mutaciones en *TET2* son mutuamente excluyentes con las mutaciones en los genes *IDH1* e *IDH2*⁽⁴¹⁾, pero pueden coexistir con mutaciones en *EZH2*, con una incidencia de hasta en el 35%⁽²²⁾ y con mutaciones en *SRSF2*, principalmente en LMMC⁽⁶⁾.

Las mutaciones en *TET2* no parecen ser un factor con valor pronóstico de la enfermedad; sí se ha descrito en algunas de las series estudiadas que se correlacionan con la respuesta a agentes hipometilantes, como la 5-azacitidina⁽²³⁾. Paradójicamente, los pacientes con *TET2* mutado tendrían una tasa de respuesta del 82% a azacitidina

en comparación con el 45% de los pacientes sin mutaciones en *TET2*. Por el contrario, no se observó correlación con la duración de la respuesta y la supervivencia global⁽²³⁾. Además, la respuesta fue aún más pronunciada si se excluyen los pacientes con *ASXL1* mutado y aquellos con mutaciones *TET2* de baja carga alélica⁽²⁴⁾. Por otro lado, las mutaciones de *TET2* parecen estar asociadas a una supervivencia global inferior tras un trasplante hematopoyético⁽²⁴⁾. Esto podría indicar que la adquisición de *ASXL1* como una mutación secundaria podría conducir a una enfermedad más progresiva, que posteriormente podría provocar una resistencia al tratamiento hipometilante (Tabla 2).

DNMT3A (2p23)

El gen *DNMT3A*, DNA metiltransferasa 3 alfa, codifica para una proteína involucrada en la regulación epigenética.

Las mutaciones de *DNMT3A* se presentan con una frecuencia del 12-18% de los SMD⁽¹⁾ con una mayor incidencia en pacientes de edad avanzada. Estas mutaciones heterocigóticas y mayoritariamente de tipo *missense* o *nonsense*, frecuentemente se concentran en el dominio metiltransferasa (MTase) aunque se pueden dar a lo largo de todo el gen. La mutación más frecuente se produce en el aminoácido R882, y se cree que conlleva una pérdida de función⁽²⁵⁾. Las mutaciones de *DNMT3A* se han observado también en otras neoplasias no hematológicas.

La presencia de mutaciones de *DNMT3A* en casi todas las células de médula ósea de pacientes con SMD, independientemente del porcentaje de blastos, ha llevado a la hipótesis de que se trata de un evento genético temprano en el proceso de iniciación de la enfermedad que podría dar lugar a una ventaja clonal de las células que contienen esta mutación⁽²⁵⁾.

Las mutaciones de *DNMT3A* en los SMD están asociadas a un curso clínico más agresivo, con progresiones más rápidas a LMA, supervivencia global más corta⁽²⁵⁾ y menor supervivencia tras trasplante⁽²⁴⁾. En los SMD-SA las mutaciones de *DNMT3A* a menudo coexisten con mutaciones en *SF3B1*, y en estos casos se ha observado una asociación con un peor pronóstico y una mayor transformación a LMA⁽⁸⁾. Finalmente, principalmente en estudios en LMA, se ha asociado la presencia de mutaciones en

DNMT3A con una mejor respuesta a tratamientos hipometilantes (especialmente a decitabina)^(24,26) (Tabla 2).

IDH1/2 (2q33.3/15q26.1)

En condiciones normales, las proteínas *IDH1* (citoplásmica) e *IDH2* (mitocondrial) convierten el isocitrato en alfa-cetoglutarato, que constituye uno de los pasos del ciclo de Krebs, una vía metabólica muy importante que afecta otros muchos procesos bioquímicos. Las mutaciones en estos genes provocan la conversión del isocitrato en 2-hidroxiglutarato. La acumulación de este oncometabolito produce un bloqueo de la diferenciación celular, inhibiendo la actividad de la histona modificadora de cromatina y las demetilinas del DNA. La inhibición de estos reguladores epigenéticos provoca una “firma hipermetilada” que altera la expresión, y las células pierden la capacidad de progresar desde los progenitores inmaduros a células completamente diferenciadas.

Las mutaciones en *IDH1* e *IDH2* se encuentran en 10-15% de los casos con LMA y en 5-10% de los casos con SMD (siendo más frecuentes en los subtipos AREB-2 (23%), AREB-1 (14%), e inexistentes en los SMD con del(5q) aislada⁽⁴¹⁾. Generalmente, las mutaciones en *IDH1* e *IDH2* se producen en el exón 4 y afectan a tres residuos específicos de arginina: R132 "*IDH1*", R140 y R172 "*IDH2*"⁽²⁷⁾. Clínicamente, las mutaciones en *IDH* se pueden ver más frecuentemente en aquellos pacientes mayores con mayor recuento de plaquetas⁽²⁷⁾.

La relevancia pronóstica de las mutaciones de *IDH* depende de la mutación identificada. Las mutaciones en el codón 132 de *IDH1* (R132) y en el codón 172 de *IDH2* (R172) parecen estar asociadas con supervivencias globales más cortas; sin embargo, la mutación R140 en *IDH2* no se ha vinculado a mal pronóstico en SMD⁽²⁸⁾. En análisis multivariante utilizando *IDH*, dependencia transfusional, y las variables del IPSS-R, se demostró que la presencia de mutaciones en *IDH1*, pero no *IDH2*, era la única variable asociada con una inferior supervivencia libre de leucemia⁽²⁸⁾. En los SMD de bajo riesgo, las mutaciones de *IDH2*, son un factor pronóstico adverso, en términos de supervivencia global, cuando coocurren con mutaciones en *DNMT3A*, *ASXL1* y *SRSF2*⁽²⁷⁾. Actualmente se están utilizando pequeñas moléculas inhibitoras de *IDH* mutado en ensayos clínicos para pacientes con LMA que portan estas mutaciones⁽²⁸⁾ (Tabla 2).

ASXL1 (20q11.1)

ASXL1 codifica para una proteína de unión a la cromatina involucrada en la regulación epigenética de las células hematopoyéticas, y juega un papel importante reclutando al “complejo represivo del *polycomb 2*” (*PRC2*). Este complejo normalmente reprime la expresión de varios genes relacionados con leucemogénesis. Por este motivo, la pérdida de función de *ASXL1* altera la diferenciación mieloide y conduce a leucemogénesis⁽²⁹⁾.

Las mutaciones en *ASXL1* son de tipo *nonsense* e inserciones/deleciones que provocan un cambio en el marco de lectura en el extremo 5' del exón 12, lo que resulta en la pérdida de la expresión de la proteína *ASXL1*⁽²⁹⁾.

ASXL1 se encuentra mutado en 5-25% de los SMD y 40-50% en LMMC^(1,29). Las mutaciones en este gen han sido relacionadas con un peor pronóstico en un amplio espectro de neoplasias de origen mieloide⁽²⁰⁾. En los SMD, tanto de bajo como de alto riesgo, se asocia a una menor supervivencia independientemente de otros parámetros clínicos como la edad, la citogenética y las citopenias^(20,29). Las mutaciones en *ASXL1* se detectan en pacientes que se transforman a LMA, lo que sugiere un posible papel en la progresión de la enfermedad. Las mutaciones en este gen constituyen un marcador pronóstico adverso independiente, cuya presencia incluso parece anular el buen pronóstico de los genes asociados a pronóstico favorable^(8,29) (Tabla 2).

En LMMC, las mutaciones de cuatro genes (*ASXL1*, *RUNX1*, *NRAS* y *SETBP1*), junto con las alteraciones cromosómicas, la dependencia transfusional, el recuento de leucocitos y el porcentaje de blastos en médula ósea permite definir un *score* pronóstico molecular (*CMML-specific prognostic scoring system*, CPSS-Mol) que identifica 4 grupos de riesgo con diferente supervivencia global y incidencia acumulada de progresión leucémica⁽³⁰⁾ (Tabla 2).

EZH2 (7q35-36)

EZH2 (*enhancer de Zeste Homolog 2*) es un gen que codifica para una proteína histona metiltransferasa de *PRC2*. Como se ha mencionado anteriormente, *PRC2* suele ejercer una función represora de la expresión de diversos genes relacionados con leucemogénesis.

Las mutaciones en *EZH2* han sido descritas en SMD, linfomas, cáncer colorrectal y cáncer de endometrio. En los SMD, estas mutaciones a menudo involucran deleciones o translocaciones, y la opinión más generalizada apoya que estas alteraciones están asociadas a una pérdida de función de la proteína. Se han descrito mutaciones en *EZH2* en aproximadamente el 5-12% de los SMD y la LMMC y se ven raramente en LMA (1-2%), sugiriendo que estas mutaciones pueden ser un indicador importante de progresión de la enfermedad en SMD⁽²²⁾. Suelen asociarse a UPD7q (*uniparental disomy*) del 7q⁽³¹⁾.

Las mutaciones de *EZH2* pueden concurrir con mutaciones en *TET2*. En modelos preclínicos se ha descrito que la asociación de ambas mutaciones acelera la aparición de la enfermedad⁽²²⁾.

Las mutaciones de *EZH2* son un biomarcador pronóstico asociado a supervivencias globales más cortas en SMD de bajo y alto riesgo⁽⁷⁾ y en LMMC-2⁽³²⁾. Recientemente se ha descrito que los pacientes del grupo de bajo riesgo con mutaciones en *EZH2*, tendrían peor pronóstico y se podrían beneficiar de un inicio precoz del tratamiento⁽³²⁾. También se le ha asociado un mayor riesgo de transformación leucémica, sobre todo en los SMD de riesgo intermedio-2 y alto riesgo. Las mutaciones de *EZH2* pueden concurrir con mutaciones en *TET2*, y la asociación de ambas mutaciones podría estar relacionada con una enfermedad más progresiva⁽²²⁾ (Tabla 2).

COHESINAS

El complejo multiproteico de las cohesinas está codificado por los genes *STAG2*, *SMC1A*, *SMC3*, *RAD21*, y *STAG1*. Está involucrado en la cohesión de las cromátidas hermanas, en la reparación del DNA post-replicativo y en la regulación transcripcional.

Las mutaciones en cualquiera de ellos pueden encontrarse en pacientes con SMD, tanto en los de bajo riesgo (con una incidencia del 11%) como en los de alto riesgo (17%), así como en la LMMC (10%)⁽³³⁾. Las mutaciones suelen generar codones de terminación prematura de la traducción, bien por mutación de tipo *nonsense*, o por mutación de tipo *frameshift*. Las mutaciones en estos genes tienden a ser mutuamente excluyentes y se considera que causan una pérdida de función. Sin embargo, actualmente solo se incluye en los paneles de diagnóstico el gen *STAG2* que es el único que tiene una relación causal demostrada en estas enfermedades.

STAG2 (Xq25)

El gen *STAG2* codifica para una subunidad del complejo de las cohesinas, el cual está involucrado en muchos procesos celulares, como por ejemplo la reparación de la rotura de la doble cadena de DNA, y la segregación de las cromátidas durante la mitosis⁽³³⁾. Se han descrito mutaciones en *STAG2* en SMD (5-10%), más frecuentemente en los subtipos CRDM y AREB, así como en LMA (1%) y en otros cánceres⁽¹⁾. El papel de las mutaciones de *STAG2* en los SMD no es muy conocido, aunque se cree que podría estar involucrado en la leucemogénesis⁽³³⁾. Las mutaciones en *STAG2* se han asociado con una peor supervivencia en los SMD pero no en LMA⁽⁷⁾. Hasta el momento, no se han validado asociaciones pronósticas con mutaciones en otros genes del complejo de las cohesinas.

FACTORES DE TRANSCRIPCIÓN

Los factores de transcripción son unas proteínas que participan en la regulación de la transcripción, sin formar parte de la RNA polimerasa. Pueden actuar reconociendo y uniéndose a secuencias concretas del DNA, uniéndose a otros factores, o uniéndose directamente a la RNA polimerasa.

Aproximadamente el 20% de los pacientes con SMD y LMMC presentan alteraciones en algún factor de transcripción, como *RUNX1*, *GATA2* y *ETV6*, ocasionando la pérdida de función de estos factores⁽¹⁾.

RUNX1 (21q22)

RUNX1 (*Runt-related transcription factor 1*, también conocido como *AML1*) es un gen que codifica para la subunidad alfa de la proteína del factor de unión central que participa en la regulación de la expresión de múltiples genes implicados en la hematopoyesis normal.

Las mutaciones de *RUNX1* pueden ser de tipo *missense*, *nonsense*, *frameshift* o de *splicing*, y dan lugar a una proteína truncada o aberrante con un efecto dominante-negativo o de haploinsuficiencia^(7,34). En SMD y LMMC se han descrito mutaciones de

RUNX1 en el 10-15% de los pacientes⁽³⁴⁾. Este porcentaje aumenta hasta más del 20% en SMD de alto riesgo, SMD-t y en LMA post SMD con bajo porcentaje de blastos⁽³⁴⁾, lo que sugiere su asociación con un fenotipo de alto riesgo. Frecuentemente se asocia con -7/del(7q) y trombocitopenia severa. Las mutaciones en *RUNX1* son un biomarcador pronóstico asociado a supervivencias globales más cortas confirmado en análisis multivariantes de varias series^(7,30).

En el caso de las mutaciones de *RUNX1* de origen germinal, existe una alta probabilidad de que produzcan insuficiencia de la médula ósea que con el tiempo provoque la evolución a distintas enfermedades hematopoyéticas de la línea mieloide^(4,5,35).

En LMMC, las mutaciones de *RUNX1* junto con (*ASXL1*, *NRAS* y *SETBP1*), se incluyen dentro del índice pronóstico molecular CPSS-Mol que identifica 4 grupos de riesgo con diferente supervivencia global e incidencia acumulada de progresión leucémica⁽³⁰⁾.

SETBP1 (18q12.3)

El gen *SETBP1* codifica para una proteína nuclear que interactúa con el gen supresor de tumores *PP2A*, desempeñando un papel crítico en la proliferación celular⁽³⁶⁾. Las mutaciones de *SETBP1* se encuentran en el 2-5% de los pacientes con SMD⁽¹⁾ y en el 5-10% de pacientes con LMMC^(1,30). Estudios recientes describen mutaciones de *SETBP1* en hasta un 9% de todos los pacientes con SMD/NMP, con frecuencia especialmente alta (24-32%) en la LMC atípica⁽³⁷⁾.

La presencia de la mutación *SETBP1* a menudo se asocia con leucocitosis, -7/del(7q), isocromosoma 17p, y mutaciones de *ASXL1*, *EZH2* y *SRSF2*^(36,37). Se ha descrito que las

mutaciones en *SETBP1* son un indicador de pronóstico adverso en la supervivencia de los SMD⁽³⁶⁾. Estas mutaciones se producen durante la evolución, lo que sugiere que podrían estar involucradas en la transformación leucémica de los SMD⁽³⁶⁾ (Tabla 2).

De manera similar, se ha demostrado el valor pronóstico independiente de *SETBP1* en LMMC, ya que mutaciones en este gen se asocian a niveles altos de leucocitos, enfermedad extramedular frecuente, supervivencia libre de LMA y progresión a LMA⁽³⁰⁾. Por ello, estas mutaciones han sido incorporadas a los distintos índices pronósticos específicos para LMMC^(30,38) (Tabla 2).

GATA2 (3q21.3)

El gen *GATA2* es un factor de transcripción que es crucial para la proliferación y la supervivencia de las células hematopoyéticas. Se han descrito mutaciones de este gen en diferentes tipos de enfermedades. Las mutaciones *GATA2* hereditarias están asociadas con sordera, linfedema, citopenias mononucleares, susceptibilidad a infecciones, mielodisplasia familiar y LMA⁽⁷⁾.

La identificación de mutaciones *GATA2* en pacientes con neutropenia crónica leve que evoluciona posteriormente a SMD/LMA sugiere que puede ser un hallazgo temprano del inicio de la enfermedad^(7,10) (Tabla 2).

ETV6 (12p13)

ETV6 (*ETS variant gene 6*, también conocido como *TEL*) es un factor de transcripción que tiene un papel crítico en la embriogénesis y en la hematopoyesis. Las alteraciones más frecuentes que afectan a este gen son translocaciones (se han descrito más de 30 *partners*, aunque en SMD el más frecuente *EV11*), aunque también se han descrito deleciones y mutaciones en el 2-5% de las SMD. Las deleciones de este gen suelen asociarse con monosomías del cromosoma 7 y a un pronóstico adverso^(34,39) (Tabla 2).

BCOR y BCORL1 (Xp11.4)

BCOR y *BCORL1* codifican para proteínas que funcionan como el co-represor de *BCL6*⁽⁹⁷⁾. El complejo co-represor *BCL6/BCOR* juega un papel importante en el desarrollo embrionario temprano. *BCORL1*, también ubicado en el cromosoma X, tiene una actividad funcional similar⁽⁴⁰⁾.

Mutaciones en *BCOR* y *BCORL1* se detectan en un 5% y 1% de pacientes con SMD, respectivamente, con mayor frecuencia en pacientes con LMMC (7%) y LMA (9%, aunque si nos centramos en las de cariotipo normal alcanzan el 17%)⁽⁴¹⁾. Curiosamente, las mutaciones de *BCOR* y *BCORL1* parecen concurrir en pacientes que tienen mutaciones en *RUNX1* y *DNMT3A*, y se asocian con una menor supervivencia global y un mayor riesgo de transformación leucémica⁽⁴¹⁾. El análisis multivariante incluyendo edad, IPSS, dependencia transfusional y estado mutacional, mostró que la presencia

de una mutación en *BCOR* era un factor pronóstico desfavorable independiente para supervivencia global⁽⁴¹⁾ (Tabla 2).

TRANSDUCCIÓN DE SEÑALES (VÍAS DE SEÑALIZACIÓN)

Es frecuente encontrar pacientes con SMD y LMMC portadores de mutaciones en los genes *CBL*, *JAK2*, *NRAS*, *KRAS*, *MPL*, *NF1*, *PTPN11*, *KIT* y *FLT3*. Las alteraciones heterocigóticas descritas en estos genes se localizan, mayoritariamente, en regiones *hotspot* siendo más frecuentes las mutaciones *missense* (como en los genes *NRAS*, *KRAS* y *MPL*) y las *frameshift* (como *FLT3*). Todas ellas causan la activación constitutiva de las proteínas codificantes^(30,34).

VÍA RAS (*NRAS*, *KRAS*, *PTPN11*, *CBL*, *NF1* y *KIT*)

Se encuentran mutaciones en genes implicados en esta vía en el 5-10% de los casos con SMD.

Las mutaciones más frecuentes son en *NRAS* (5-10%) en SMD, mientras que en la LMMC son frecuentes tanto las mutaciones en *NRAS* como en *KRAS* (10-20%, respectivamente). Las mutaciones en *NRAS* se asocian a trombocitopenia severa⁽³⁴⁾.

A estas mutaciones las siguen en frecuencia las de *CBL*. El gen *CBL* (*Casitas B-lineage lymphoma*), ubicado en 11q23.3, codifica para una proteína con actividad E3 *ubiquitina ligasa*, que se cree que es un regulador negativo de la señal del receptor tirosina quinasa que impulsa la oncogénesis⁽⁴²⁾. Se han detectado mutaciones de *CBL* en el 8-18% de los pacientes con LMMC⁽⁴³⁾ y en un 10% de los pacientes con otros SMD/NMP⁽⁴⁴⁾. Curiosamente, la tasa de mutación de *CBL* es significativamente mayor en SMD secundarios en comparación con SMD de *novo* (12% frente a 1%, respectivamente)⁽⁴²⁾ (Tabla 2).

En LMMC, las mutaciones de *NRAS* junto con *ASXL1*, *RUNX1* y *SETBP1*, se incluyen dentro del índice pronóstico molecular CPSS-Mol que identifica 4 grupos de riesgo con diferente supervivencia global e incidencia acumulada de progresión leucémica⁽⁹⁰⁾ (Tabla 2).

VÍA TIROSINA QUINASA (*FLT3* y *JAK2*)

Las alteraciones en estas vías son poco frecuentes en los SMD. Las alteraciones en *FLT3-ITD* son comunes en LMA y confieren mal pronóstico, pero en SMD su incidencia es menor de un 5%⁽¹⁾.

Por otro lado, encontramos mutaciones en *JAK2* (principalmente la *V617F*) en el 50% de las SMD con sideroblastos en anillo (SA) y con trombocitosis (ARSA-T), en el 2-10% de las LMMC y en el 6% de pacientes con del(5q) aislada⁽⁴⁶⁾. Las evidencias sobre la importancia clínica de la mutación *JAK2* en los SMD son limitadas. Un trabajo reciente en pacientes con SMD, demostró que la mutación *JAK2 V617F* aparentemente estaba asociada con una menor progresión a LMA y una tendencia a una mayor supervivencia global⁽⁴⁷⁾ (Tabla 2). Además, los inhibidores de *JAK2* actualmente disponibles, y ampliamente utilizados en NMP, podrían mejorar el curso evolutivo de la enfermedad.

REPARACIÓN DEL DAÑO CELULAR

TP53 (17p13)

TP53 (17p13) es un gen que codifica para una proteína supresora tumoral que también funciona como factor transcripcional, y que en situaciones de estrés celular regula la expresión de genes involucrados en arresto del ciclo celular, apoptosis, senescencia, reparación del DNA, y cambios en el metabolismo de la célula. *TP53* es el gen más frecuentemente mutado en cáncer (presenta mutaciones en la mitad de los casos de cáncer). Las mutaciones van a alterar la función normal del gen, provocando un daño en el DNA y afectando a la supervivencia celular y la proliferación celular. Las mutaciones de *TP53* se dan en aproximadamente el 5-10% de las SMD, más frecuentemente en SMD de alto riesgo, y se encuentran mutaciones a lo largo de todo el gen. Además, ocurren frecuentemente junto con la pérdida del segundo alelo de *TP53* mediante la delección de uno de los brazos cortos del cromosoma 17⁽³⁴⁾.

Las mutaciones de *TP53* en los SMD se han asociado a manifestaciones clínicas desfavorables, como por ejemplo enfermedad avanzada, cariotipos complejos, en ocasiones con afectación del cromosoma 5 (72%), y con los cariotipos monosómicos, exceso de blastos en la médula ósea y trombocitopenia severa presentando un valor

pronóstico independiente^(34,46). Además, ocurren frecuentemente junto con la pérdida del segundo alelo de *TP53* mediante la delección de uno de los brazos cortos del cromosoma 17⁽³⁴⁾.

En general, la presencia de mutaciones de *TP53* tiene un impacto negativo independiente en el pronóstico de todos los subtipos de SMD⁽⁴⁶⁾. La supervivencia global de los pacientes con *TP53* es inferior a la de los no mutados (6 vs. 66 meses)⁽⁴⁸⁾. Los pacientes con esta mutación también tienen mayor riesgo de transformación a LMA aunque el riesgo se ajuste a otras variables pronósticas⁽⁶⁾. Las mutaciones de *TP53* se han asociado también a un peor pronóstico en pacientes que han recibido trasplante de progenitores hematopoyéticos⁽²⁴⁾ y con pacientes que previamente han sido sometidos a tratamiento con quimioterapia⁽⁴⁹⁾.

En relación con los pacientes que presentan un cariotipo complejo, aproximadamente el 50% de estos pacientes no tienen una alteración de *TP53* detectable y presentan una supervivencia global comparable a la de los pacientes con cariotipos no complejos. Por lo tanto, el estado mutacional de *TP53* puede ser útil para refinar el pronóstico de estos pacientes típicamente considerados como de mayor riesgo⁽⁴⁶⁾.

Por otro lado, las mutaciones de *TP53* se han asociado con la resistencia al tratamiento con lenalidomida en pacientes con SMD y del(5q), ya sea como anomalía aislada o como parte de un cariotipo complejo⁽⁵⁰⁾. Un 18% de pacientes con SMD y del(5q) presentan mutaciones de *TP53* y, en estos casos, el tratamiento con lenalidomida produce respuestas subóptimas, recaídas o resistencia al tratamiento⁽⁷⁾. En estos casos, las mutaciones *TP53* pueden ser eventos secundarios y están a menudo presentes en subclones pequeños que pueden expandirse durante el tratamiento. Por este motivo, la OMS recomienda evaluar las mutaciones de *TP53* en pacientes con SMD y del(5q) aislado, para ayudar a identificar un subgrupo de pronóstico adverso en esta entidad generalmente de pronóstico favorable⁽⁴⁶⁾ (Tabla 2).

A nivel de tratamiento, se han investigado fármacos *antisense* complementarios a *TP53* que inhibiría la expresión de la versión mutada de p53, y restauraría la eritropoyesis en pacientes con SMD del(5q)⁽¹⁶⁾. Recientemente, varias publicaciones asocian mutaciones en *TP53* a una mayor tasa de respuesta a decitabina en SMD y LMA⁽⁵¹⁾, aunque hay discrepancias en los dos estudios en cuanto a la asociación de esta respuesta a una mejor supervivencia global (Tabla 2).

Tabla 2. Impacto pronóstico de los genes comúnmente mutados en SMD y LMMC.

GEN	SMD		LMMC	
	INCIDENCIA	SIGNIFICADO CLÍNICO	INCIDENCIA	SIGNIFICADO CLÍNICO
<i>ASXL1</i>	5-25%	Desfavorable	40-50%	Desfavorable Progresión a LMA
<i>CSNK1A1</i>	5-10%	Desconocido Asociado a del(5q) y SF3B1, 42%	<1%	Desconocido
<i>DNMT3A</i>	12-18%	Desfavorable Asociado a CN	2-10%	Desfavorable Progresión a LMA
<i>EZH2</i>	5-10%	Desfavorable	5-12%	Desfavorable Asociado a LMMC-2
<i>IDH1</i>	<5%	Desfavorable	<1%	Desfavorable
<i>IDH2</i>	<5%	Desfavorable	5-10%	Desfavorable
<i>JAK2</i>	<5%	SMD con del(5q), 5-7%; ARSA-T, 50%	2-10%	Asociado a LMMC-MP
<i>KRAS</i>	5-10%	Desfavorable	10-20%	Asociado a LMMC-MP
<i>NRAS</i>	5-10%	Desfavorable	10-20%	Desfavorable Asociado a LMMC-MP
<i>RUNX1</i>	10-15%	Desfavorable Pueden ser congénitas	10-30%	Desfavorable Progresión a LMA
<i>SETBP1</i>	<5%	Desfavorable	5-10%	Desfavorable Progresión a LMA
<i>SF3B1</i>	20-30% 80% SA	Asociado a SA Favorable	5-10%	Incierto
<i>SRSF2</i>	10-15%	Desfavorable	30-50%	Desfavorable
<i>STAG2</i>	5-10%	Desfavorable	5-10%	Desfavorable
<i>TET2</i>	20-25%	Asociado a cariotipo normal Mayor tasa de respuesta a hipometilantes	45-60%	Incierto
<i>TP53</i>	8-12%	Desfavorable Asociado a CC (50%) y del(5q) (15-20%) Predictor fallo de Lenalidomida	<5%	Desfavorable
<i>U2AF1</i>	8-12%	Desfavorable	5-10%	Desfavorable
<i>ZRSR2</i>	5-10%	Desconocido	5-10%	Desconocido

ARSA-T: anemia refractaria con sideroblastos en anillo y trombocitosis; CC: cariotipo complejo; CN: cariotipo normal;
LMMC: leucemia mielomonocítica crónica; LMMC-MP: LMMC mieloproliferativa; SA: sideroblastos en anillo;
SMD: síndromes mielodisplásicos.

VARIANTES GERMINALES DE PREDISPOSICIÓN Y HEMATOPOYESIS CLONAL EN EDAD AVANZADA

Durante los últimos años se han identificado síndromes con predisposición a desarrollar algún tipo de cáncer, incluyendo leucemias y linfomas. Este nuevo conocimiento hace que la interpretación de las variantes deba tener en cuenta la posibilidad de que el desarrollo de neoplasias mieloides puede deberse a mutaciones que pueden ser tanto somáticas como germinales. Hoy en día son múltiples los genes que se han encontrado asociados a predisposición de neoplasias hematológicas (Tabla 3), lo que indica que identificar variantes de origen germinal en un test genómico sea altamente probable.

Tabla 3. Genes asociados a predisposición de neoplasias hematológicas (adaptada a Frutani and Shimamura *et al.*, 2017)⁽⁵²⁾.

SÍNDROME	GEN	TIPO DE HERENCIA	NEOPLASIA HEMATOLÓGICA ASOCIADA
Noonan-like	<i>CBL</i>	AD	LMMJ
Leucemia Mielode Aguda	<i>CEBPA</i>	AD	LMA
Familiar Trombocitopenia 5	<i>ETV6</i>	AD	SMD, LMA, LMMC, LLA
Trastornos relacionados con GATA2	<i>GATA2</i>	AD	SMD, LMA, LMMC
Trombocitopenia amegacariocítica congénita	<i>MPL</i>		SMD, LMA
Neurofibromatosis	<i>NF1</i>	AD	LMMJ, LMA
Síndrome de Noonan	<i>PTPN11</i>	AD	LMMJ, LLA
Trastorno plaquetario familiar con predisposición a neoplasias mieloides	<i>RUNX1</i>	AD	SMD, LMA
Li Fraumeni	<i>TP53</i>	AD	SMD, LMA, LLA

AD: autosómica dominante; LMMJ: leucemia mielomonocítica juvenil; LMA: leucemia mielode aguda; LLA: leucemia linfoblástica aguda; SMD: síndrome mielodisplásico.

De hecho, la clasificación de la OMS 2016 recalca la importancia de tener en cuenta las neoplasias mieloides con predisposición germinal (Tabla 4).

Tabla 4. Clasificación de neoplasias mieloides con predisposición germinal (adaptada a de Arber *et al.*, 2016)⁽⁹⁾.

CLASIFICACIÓN DE LAS NEOPLASIAS MIELOIDES
Neoplasias mieloides con predisposición germinal sin patología previa ni disfunción orgánica <ul style="list-style-type: none"> • LMA con mutación germinal de <i>CEBPA</i> • Neoplasias mieloides con mutación germinal de <i>DDX41</i>*
Neoplasias mieloides con predisposición germinal y patología plaquetaria previa <ul style="list-style-type: none"> • Neoplasias mieloides con mutación germinal de <i>RUNX1</i>* • Neoplasias mieloides con mutación germinal de <i>ANKRD26</i>* • Neoplasias mieloides con mutación germinal de <i>ETV6</i>*
Neoplasias mieloides con predisposición germinal y otra disfunción orgánica <ul style="list-style-type: none"> • Neoplasias mieloides con mutación germinal de <i>GATA2</i> • Neoplasias mieloides asociadas a síndrome de fallo medular • Neoplasias mieloides asociadas a desórdenes teloméricos • LMMJ asociada a neurofibromatosis, o a Síndromes de Noonan o Noonan-like • Neoplasias mieloides asociadas a Síndrome de Down

*También descritas en neoplasias linfoides. Los genes *DDX41*, *ANKRD26* y los genes asociados a los telómeros no están incluidos en el listado de mínimos genes recomendados.

El hecho de que al estudiar neoplasias hematológicas de origen mielode se realice el análisis de la muestra tumoral en médula ósea o en sangre periférica hace posible que se detecte por accidente variantes que podrían ser germinales.

La detección de variantes con frecuencias alélicas (VAF, *variant allele frequency*) entre 40-60% puede hacer sospechar de la presencia de una variante de origen germinal. Sin embargo, hay que tener precaución al informar una variante como germinal. Hay que tener en cuenta que presentar una VAF del 40-60% no es un dato concluyente, ya que, por lo general, se desconoce la infiltración de células tumorales en la muestra, dato necesario para estimar la cigosidad de una variante en base a su VAF.

Desafortunadamente, en el análisis genómico que se realiza en el contexto de la rutina diagnóstica no se suelen utilizar muestras control (tampoco está perfectamente definida y consensuada qué muestra es la más adecuada como control), que permitan descartar o confirmar la presencia de una variante de la línea germinal, y por tanto, de una posible enfermedad hereditaria. Un modo alternativo de descartar si una variante es o no germinal es analizar si su VAF varía en las muestras que se analicen durante el seguimiento de la enfermedad: las variaciones reflejarían fluctuaciones en la carga tumoral de la muestra, probando así su naturaleza somática. Otro dato que ayuda a

identificar las variantes que afectan a la línea germinal es que dichas variantes suelen ir acompañadas de un cuadro clínico/historia familiar característico, por lo que en el caso que no se pueda descartar su origen germinal con el análisis de muestras control o al seguimiento, se tendría que recomendar su interpretación dentro del contexto clínico del paciente.

Finalmente, también se recomienda precaución al interpretar las variantes encontradas en personas de edad avanzada, ya que se han descrito mutaciones recurrentes en determinados genes en personas mayores que no presentan evidencia de enfermedad hematológica. Curiosamente, los genes que se han encontrado mutados en esta población son los mismos que en SMD, aunque con frecuencias diferentes, de mayor a menor frecuencia serían: *TET2*, *DNMT3A*, *ASXL1*, *SRSF2*, *TP53*, *SF3B1*, *JAK2* y *CBL*⁽⁵⁴⁾ (Tabla 5). La mayoría de estos casos no desarrollarán nunca una neoplasia hematológica, por lo que el hallazgo de estas mutaciones, que indican clonalidad, se ha denominado hematopoyesis clonal de potencial indeterminado (CHIP, por sus siglas en inglés). En los casos con CHIP, la detección de una citopenia (que puede deberse a déficit vitamínico, toxicidad alcohólica, infección vírica o a alguna otra causa no neoplásica) junto con la identificación de una mutación, no bastan para establecer un diagnóstico de SMD.

No obstante, estos casos no han de confundirse con los casos de citopenias idiopáticas de significado incierto (ICUS) que cursan con citopenias (por causas varias) pero sin terminar de cumplir criterios de SMD. En estos casos, la presencia de mutaciones en los mismos genes es mucho mayor, así como el tamaño del clon portador de la mutación. El 40% de los pacientes con ICUS tienen evidencia de hematopoyesis clonal y se podrían clasificar como citopenias clonales de significado incierto (CCUS). En estos casos, si el paciente es portador de alguna mutación tienen mayor probabilidad de desarrollar una neoplasia hematológica⁽⁵⁴⁾ (Tabla 5).

Tabla 5. Comparación de las características genéticas entre CHIP, CCUS y SMD⁽⁵⁴⁾.

	CHIP POBLACIÓN NO SELECCIONADA	CCUS AL DIAGNÓSTICO	CCUS PREVIO A PROGRESIÓN A SMD/LMA	SMD, TODOS LOS GRUPOS DE RIESGO
Genes mutados	<i>TET2</i> , <i>DNMT3A</i> , <i>ASXL1</i> , <i>PPM1D</i> , <i>TP53</i> , <i>JAK2</i>	<i>TET2</i> , <i>DNMT3A</i> , <i>ASXL1</i> , <i>SRSF2</i> , <i>TP53</i>	<i>TET2</i> , <i>DNMT3A</i> , <i>ASXL1</i> , <i>SRSF2</i> , <i>U2AF1</i>	<i>TET2</i> , <i>DNMT3A</i> , <i>ASXL1</i> , <i>SRSF2</i> , <i>SF3B1</i>
Media de mutaciones	~1	~1,6	~2	~2,6
VAF aproximado	9-12%	30-40%	~40%	30-50%
Incidencia	~10-15% de los >70 años	~35% de ICUS	~90% de los ICUS que progresan a SMD/LMA	<50% de los pacientes con citopenias

CHIP: hematopoyesis clonal de potencial indeterminado; CCUS: citopenias clonales de significado incierto; ICUS: citopenias idiopáticas de significado incierto; LMA: leucemia mieloide aguda; SMD: síndrome mielodisplásico; VAF: frecuencia del alelo variante.

La mayoría de veces los laboratorios que realizan los estudios genómicos en SMD empleando paneles dirigidos no disponen de esta información clínica, por ello es importante el solicitante de la prueba lo tenga cuenta antes de tomar una decisión diagnóstica/terapéutica.

REFERENCIAS

1. Kennedy JA, Ebert BL. Clinical Implications of Genetic Mutations in Myelodysplastic Syndrome. *J Clin Oncol*. 2017 Mar 20;35(9):968-974.
2. Greenberg PL, Stone RM, Al-Kali A, et al. Myelodysplastic Syndromes, Version 2.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw*. 2017 Jan;15(1):60-87.
3. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 2011;478:64–9.
4. Abrahamsson AE, Geron I, Gotlib J, et al. Glycogen synthase kinase 3beta missplicing contributes to leukemia stem cell generation. *Proc Natl Acad Sci U S A*. 2009 Mar 10;106(10):3925-9.
5. Papaemmanuil E, Gerstung M, Malcovati L, et al. Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013 Nov 21;122(22):3616-27.
6. Malcovati L, Papaemmanuil E, Bowen DT, et al. Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium and of the Associazione Italiana per la Ricerca sul Cancro Gruppo Italiano Malattie Mieloproliferative. Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. *Blood*. 2011 Dec 8;118(24):6239-46.
7. Cazzola M, Della Porta MG, Malcovati L. The genetic basis of myelodysplasia and its clinical relevance. *Blood*. 2013 Dec 12;122(25):4021-34.
8. Martín I, Such E, Navarro B, et al. Negative impact on clinical outcome of the mutational co-occurrence of SF3B1 and DNMT3A in refractory anemia with ring sideroblasts (RARS). *Leuk Lymphoma*. 2017 Jul;58(7):1686-1693.
9. Kim E, Ilagan JO, Liang Y, et al. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell*. 2015 May 11;27(5):617-30.
10. Thol F, Kade S, Schlarmann C, et al. Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. *Blood*. 2012 Apr 12;119(15):3578-84.
11. Meggendorfer M, Roller A, Haferlach T, et al. SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood*. 2012 Oct 11;120(15):3080-8.
12. Makishima H, Yoshizato T, Yoshida K, et al. Dynamics of clonal evolution in myelodysplastic syndromes. *Nat Genet*. 2017 Feb;49(2):204-212.

13. Bejar R, Stevenson KE, Caughey BA, et al. Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. *J Clin Oncol* 30:3376-3382, 2012.
14. Przychodzen B, Jerez A, Guinta K, et al. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood*. 2013 Aug 8;122(6):999-1006.
15. Graubert TA, Shen D, Ding L, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *NatGenet* 2011;44:53–7.
16. Caceres G, McGraw K, Yip BH, et al. TP53 suppression promotes erythropoiesis in del(5q) MDS, suggesting a targeted therapeutic strategy in lenalidomide-resistant patients. *Proc Natl Acad Sci U S A*. 2013 Oct 1;110(40):16127-32.
17. Wu SJ, Tang JL, Lin CT, et al. Clinical implications of U2AF1 mutation in patients with myelodysplastic syndrome and its stability during disease progression. *Am J Hematol* 2013;88:E277–82.
18. Shih AH, Abdel-Wahab O, Patel JP, et al. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat Rev Cancer*. 2012 Sep;12(9):599-612.
19. Yamazaki J, Jelinek J, Lu Y, et al. TET2 Mutations Affect Non-CpG Island DNA Methylation at Enhancers and Transcription Factor-Binding Sites in Chronic Myelomonocytic Leukemia. *Cancer Res*. 2015 Jul 15;75(14):2833-43.
20. Gelsi-Boyer V, Brecqueville M, Devillier R, et al. Mutations in ASXL1 are associated with poor prognosis across the spectrum of malignant myeloid diseases. *J Hematol Oncol*. 2012 Mar 21;5:12.
21. Solary E, Bernard OA, Tefferi A, et al. The Ten-Eleven Translocation-2 (TET2) gene in hematopoiesis and hematopoietic diseases. *Leukemia*. 2014 Mar;28(3):485-96.
22. Muto T, Sashida G, Oshima M, et al. Concurrent loss of Ezh2 and Tet2 cooperates in the pathogenesis of myelodysplastic disorders. *J Exp Med* 2013;210:2627–39.
23. Itzykson R, Kosmider O, Cluzeau T, et al. Impact of TET2 mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia* 2011;25:1147–52.
24. Bejar R, Stevenson KE, Caughey B, et al. Somatic mutations predict poor outcome in patients with myelodysplastic syndrome after hematopoietic stem-cell transplantation. *J Clin Oncol* 32:2691-8, 2014.
25. Walter MJ, Ding L, Shen D, et al. Recurrent DNMT3A mutations in patients with myelodysplastic syndromes. *Leukemia*. 2011 Jul;25(7):1153-8.

26. Traina F, Visconte V, Elson P, *et al.* Impact of molecular mutations on treatment response to DNMT inhibitors in myelodysplasia and related neoplasms. *Leukemia* 2014;28:78–87.
27. Lin CC, Hou HA, Chou WC, *et al.* IDH mutations are closely associated with mutations of DNMT3A, ASXL1 and SRSF2 in patients with myelodysplastic syndromes and are stable during disease evolution. *Am J Hematol* 2014;89:137–44.
28. Krell D, Mulholland P, Frampton AE, *et al.* IDH mutations in tumorigenesis and their potential role as novel therapeutic targets. *Future Oncol* 2013;9:1923–35.
29. Boulwood J, Perry J, Pellagatti A, *et al.* Frequent mutation of the polycomb-associated gene ASXL1 in the myelodysplastic syndromes and in acute myeloid leukemia. *Leukemia*. 2010 May;24(5):1062–5.
30. Elena C, Galli A, Such E, *et al.* Integrating clinical features and genetic lesions in the risk assessment of patients with chronic myelomonocytic leukemia. *Blood* 128:1408–17, 2016.
31. Jerez A, Sugimoto Y, Makishima H, *et al.* Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis. *Blood*. 2012 Jun 21;119(25):6109–17.
32. Grossmann V, Kohlmann A, Eder C, *et al.* Molecular profiling of chronic myelomonocytic leukemia reveals diverse mutations in >80% of patients with TET2 and EZH2 being of high prognostic relevance. *Leukemia* 2011;25:877–9.
33. Kon A, Shih LY, Minamino M, *et al.* Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet*. 2013 Oct;45(10):1232–7.
34. Bejar R, Stevenson K, Abdel-Wahab O, *et al.* Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*. 2011;364(26):2496–2506.
35. Chen CY, Lin LI, Tang JL, *et al.* RUNX1 gene mutation in primary myelodysplastic syndrome—the mutation can be detected early at diagnosis or acquired during disease progression and is associated with poor outcome. *Br J Haematol*. 2007 Nov;139(3):405–14.
36. Hou HA, Kuo YY, Tang JL, *et al.* Clinical implications of the SETBP1 mutation in patients with primary myelodysplastic syndrome and its stability during disease progression. *Am J Hematol* 2014;89:181–6.
37. Makishima H, Yoshida K, Nguyen N, *et al.* Somatic SETBP1 mutations in myeloid malignancies. *Nat Genet*. 2013 Aug;45(8):942–6.

38. Itzykson R, Kosmider O, Renneville A, *et al.* Prognostic score including gene mutations in chronic myelomonocytic leukemia. *J Clin Oncol* 2013;31:2428–36.
39. Wall M, Rayeroux KC, MacKinnon RN, *et al.* ETV6 deletion is a common additional abnormality in patients with myelodysplastic syndromes or acute myeloid leukemia and monosomy 7. *Haematologica*. 2012 Dec;97(12):1933–6.
40. Li M, Collins R, Jiao Y, *et al.* Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. *Blood*. 2011 Nov 24;118(22):5914–7.
41. Damm F, Itzykson R, Kosmider O, *et al.* SETBP1 mutations in 658 patients with myelodysplastic syndromes, chronic myelomonocytic leukemia and secondary acute myeloid leukemias. *Leukemia* 2013;27:1401–3.
42. Nadeau S, An W, Palermo N, *et al.* Oncogenic signaling by leukemia-associated mutant Cbl proteins. *Biochem Anal Biochem Suppl* 2012;6:7921.
43. Kohlmann A, Grossmann V, Klein HU, *et al.* Next-generation sequencing technology reveals a characteristic pattern of molecular mutations in 72.8% of chronic myelomonocytic leukemia by detecting frequent alterations in TET2, CBL, RAS, and RUNX1. *J Clin Oncol*. 2010 Aug 20;28(24):3858–65.
44. Schwaab J, Ernst T, Erben P, *et al.* Activating CBL mutations are associated with a distinct MDS/MPN phenotype. *Ann Hematol* 2012;91:1713–20.
45. Kao HW, Sanada M, Liang DC, *et al.* A high occurrence of acquisition and/or expansion of C-CBL mutant clones in the progression of high-risk myelodysplastic syndrome to acute myeloid leukemia. *Neoplasia* 2011;13:1035–42.
46. Meggendorfer M, Haferlach C, Kern W, *et al.* Molecular analysis of myelodysplastic syndrome with isolated del(5q) reveals a specific spectrum of molecular mutations with prognostic impact: a study on 123 patients and 27 genes. *Haematologica*. 2017 Jun 22.
47. De Renzis B, Mansat-De Mas V, Wattel E, *et al.* Prognostic impact of JAK2V617F mutation in myelodysplastic syndromes: a matched case–control study. *Leuk Res Rep* 2013;2:64–6.
48. Kulasekararaj AG, Smith AE, Mian SA, *et al.* TP53 mutations in myelodysplastic syndrome are strongly correlated with aberrations of chromosome 5, and correlate with adverse prognosis. *Br J Haematol*. 2013 Mar;160(5):660–72.
49. Ok CY, Patel KP, Garcia-Manero G, *et al.* Mutational profiling of therapy-related myelodysplastic syndromes and acute myeloid leukemia by next generation sequencing, a comparison with de novo diseases. *Leuk Res*. 2015 Mar;39(3):348–54.

50. Jädersten M, Saft L, Smith A, et al. TP53 mutations in low-risk myelodysplastic syndromes with del(5q) predict disease progression. *J Clin Oncol* 2011;29:1971–9.

51. Chang CK, Zhao YS, Xu F, et al. TP53 mutations predict decitabine-induced complete responses in patients with myelodysplastic syndromes. *Br J Haematol*. 2017 Feb;176(4):600-608.

52. Frutani E, Shimamura A. Germline Genetic Predisposition to Hematologic Malignancy. *J Clin Oncol*. 2017;35(9):1018-1028.

53. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 127:2391-405, 2016.

54. Bejar R. Implications of molecular genetic diversity in myelodysplastic syndromes. *Curr Opin Hematol*. 2017 Mar;24(2):73-78.



PLATAFORMAS Y CRITERIOS MÍNIMOS

GUÍA DE APLICACIÓN CLÍNICA DE
LA SECUENCIACIÓN MASIVA EN
SÍNDROMES MIELODISPLÁSICOS Y
LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:

Laura Palomo, Mariam Ibáñez, Bárbara Tazón-Vega, Iria Vázquez, Rocío Benito, Inmaculada Rapado, Pamela Acha, Marta Fernandez-Mercado, Jesús María Hernández-Sánchez, Jesús María Hernández-Rivas, Francesc Solé y Esperanza Such.

PLATAFORMAS Y CRITERIOS MÍNIMOS

PLATAFORMAS

En general, podemos clasificar las plataformas de secuenciación en base a la longitud media de lectura (larga o corta); al modo de secuenciación, que incluye la secuenciación de extremo único (los fragmentos de DNA se secuencian en un solo extremo, *single-end sequencing*) y la secuenciación de ambos extremos del fragmento con lecturas pareadas (*paired-end sequencing*); y a la tecnología de secuenciación empleada (secuenciación por ligación, hibridación o síntesis, principalmente). Debido a que la tecnología de secuenciación depende de la plataforma y del sistema elegido, el diseño de la prueba debe basarse en el tipo de alteración específica que se desea detectar. Además, a la hora de decidir en qué secuenciador realizar la caracterización molecular de los pacientes, el laboratorio debe considerar cuidadosamente el tamaño de la región secuenciada, la profundidad de cobertura requerida, el volumen de muestras previsto, los tiempos de proceso requeridos y los costes totales por muestra.

A continuación se detallan las plataformas de secuenciación masiva de lectura corta *gold-standard* en los laboratorios de diagnóstico clínico.

PLATAFORMA ILLUMINA

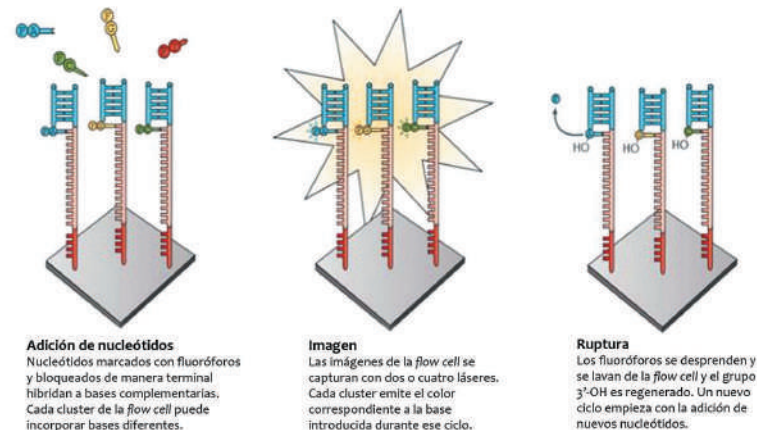
Las plataformas de Illumina dominan la industria de secuenciación masiva de lectura corta debido, en parte, a su madurez como tecnología, a la amplia gama de equipos disponibles, así como al elevado nivel de compatibilidad entre ellos. La gama de instrumentos disponibles varía desde el MiniSeq de bajo rendimiento hasta el HiSeq X de ultra rendimiento (Tabla 6 y Tabla 7).

Las plataformas de Illumina se basan en la secuenciación por síntesis mediante terminación reversible cíclica (*Cyclic Reversible Termination, CRT*), es decir, emplean nucleótidos terminadores de cadena marcados con moléculas fluorescentes, al igual que en la secuenciación de *Sanger*, aunque difieren de ésta en que tras la obtención de la imagen se elimina el nucleótido fluorescente. De este modo, al desbloquear el carbono 3', la cadena puede aceptar una nueva base y continuar la reacción de

secuenciación. En esta tecnología, el primer paso consiste en la fragmentación del DNA y la ligación de adaptadores e índices; los adaptadores permiten la hibridación de las moléculas que se desean secuenciar a la superficie de la *flow cell*, mientras que los índices permiten la identificación de cada muestra dentro de un grupo de muestras que se han secuenciado a la vez. Tras la ligación, se da un paso de amplificación sobre la superficie sólida de la *flow cell*, donde se realiza a continuación la reacción de secuenciación (Figura 2). Estos equipos captan la fluorescencia emitida durante la síntesis del DNA en forma de imágenes a través de canales ópticos situados en cámaras LED, lo que abarata los costes de los equipos, aunque ralentiza el análisis de los datos. En general, la plataforma Illumina tiene una tasa de precisión global mayor al 99,5%⁽¹⁾. Esto implica una baja tasa de errores, todos ellos correlacionados con la posición en la lectura; dichos errores se acumulan con mayor frecuencia al final de los amplicones, y son recurrentes inter- e intra- experimentos. En general, la plataforma muestra cierta sobrerrepresentación de errores de sustitución en regiones ricas en AT y ricas en GC⁽²⁻⁵⁾. Sin embargo, esta tecnología es mucho menos susceptible a los errores en las regiones de homopolímeros observados en las plataformas de secuenciación por la adición de un solo nucleótido, y en el caso de presentarse, aparecen en un porcentaje muy bajo de las lecturas.

En conjunto, las plataformas de Illumina permiten llevar a cabo una amplia gama de aplicaciones: secuenciación de genes mediante paneles de amplicones y de captura, secuenciación del genoma a través de secuenciación *Whole Genome/ Exome Sequencing* (WGS, WES); análisis epigenómicos, tales como *ChIP-Seq* (inmunoprecipitación de la cromatina seguida de secuenciación), *ATAC-Seq* (ensayo de cromatina accesible a la transposasa mediante secuenciación) o secuenciación de metilación del DNA (*Methyl-Seq*), así como análisis del transcriptoma a través de la secuenciación del RNA (*RNA-Seq*), entre otras.

a. Illumina



b. Ion Torrent (Thermo Fisher)

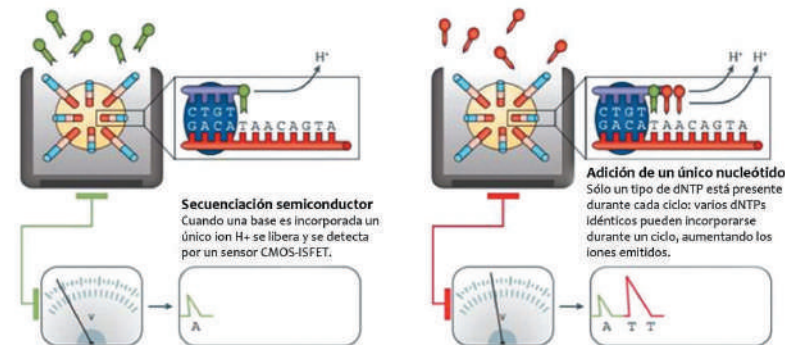


Figura 2. Secuenciación por síntesis (Modificada de Goodwin et al., 2016)⁽⁶⁾. a. Illumina. Secuenciación mediante terminación reversible cíclica (*Cyclic Reversible Termination, CRT*). Tras el enriquecimiento del DNA molde en fase sólida, la mezcla de cebadores específicos junto con la DNA polimerasa y los nucleótidos modificados se añaden a la *flow cell*. Cada nucleótido está bloqueado por un grupo 3'-O-azidometil y está marcado con un fluoróforo (F) específico de base. Durante cada ciclo, los fragmentos incorporarán solo un nucleótido ya que el grupo 3' bloqueado impide incorporaciones adicionales. Después de la incorporación de la base el dispositivo se visualiza mediante microscopía de fluorescencia de reflexión interna total (TIRF) usando dos o cuatro canales láser. El color identifica qué base se incorporó. El fluorocromo se escinde y el 3'-OH se regenera con el agente reductor tris (2 carboxietil) fosfina (TCEP). Tras esto, el ciclo de adición, elongación y escisión de nucleótidos comienza de nuevo. b. Ion Torrent. Secuenciación mediante adición de un solo tipo de nucleótido

(*Single-Nucleotide Addition, SNA*). Después del enriquecimiento del DNA molde basado en bolas (*beads*), las bolas se colocan cuidadosamente en una placa de microtitulación donde una bola ocupa un solo pocillo de reacción. Los nucleótidos se añaden a los pocillos todos a la vez y se lleva a cabo una reacción de elongación estándar. A medida que se incorpora cada base, se genera un solo ion H+ como subproducto. La liberación de H+ da lugar a un cambio de 0,02 unidades en el pH, detectado por un semiconductor complementario de óxido metálico (CMOS) y un transistor de efecto de campo sensible a iones (ISFET). Después de la adición de cada nucleótido, las bases no incorporadas se eliminan y se añade la siguiente.

PLATAFORMA ION TORRENT (THERMO FISHER)

La plataforma Ion Torrent ofrece longitudes de lectura superiores a las de otros secuenciadores de lectura corta, con un tamaño de lectura promedio de hasta 400pb, proporcionando algunas ventajas para aplicaciones que se centran en DNA repetitivo o complejo.

Tabla 6. Comparación de las plataformas de secuenciación masiva por síntesis.

PLATAFORMA	SISTEMA DE SECUENCIACIÓN	ERRORES FRECUENTES	APLICACIONES
Illumina MiniSeq (Mid/High output); MiSeq v2 y v3, NextSeq 500/550 (Mid/High output); HiSeq2500 v2, v3 y v4; HiSeq3000/4000; HiSeq X, NovaSeq 5000/6000	Secuenciación por síntesis mediante terminación reversible cíclica (<i>Cyclic Reversible Termination, CRT</i>).	Errores de sustitución en regiones ricas en AT o GC. Se acumulan al final del amplicón. Recurrentes inter e intra-experimentos.	Secuenciación del genoma y exoma, amplicones de genes específicos, análisis epigenómicos (<i>ChIP-seq, ATAC-seq o Methyl-Seq</i>) y análisis del transcriptoma (<i>RNA-Seq</i>).
Ion Torrent Ion PGM 314, 316 y 318; Ion Proton; Ion S5 520, 530 y 540	Secuenciación por síntesis mediante adición de un solo nucleótido (<i>Single-Nucleotide Addition, SNA</i>).	Errores tipo indels, frecuentes en regiones de homopolímeros. Errores en sustituciones en el 0,04%–0,1% del total de bases leídas asociadas con un <i>strand-bias</i> .	Secuenciación del exoma, amplicones de genes específicos, secuenciación dirigida del RNA, análisis de CNV y aneuploidias, perfiles de transcriptoma e identificación de sitios de <i>splicing</i> .

PLATAFORMA	LONGITUD MEDIA DE LECTURA	RENDIMIENTO	LECTURAS	TIEMPO POR CARRERA	TASA DE ERROR ESTIMADO
Illumina MiniSeq (Mid/High output); MiSeq v2 y v3, NextSeq 500/550 (Mid/High output); HiSeq2500 v2, v3 y v4; HiSeq3000/4000; HiSeq X, NovaSeq 5000/6000	36-300 (SE) 50-300 (PE)	540Mb- 900Gb	12M-2.500M (SE) 24M-4.000M (PE)	4h-3.5d	<1%, sustituciones
Ion Torrent Ion PGM 314, 316 y 318; Ion Proton; Ion S5 520, 530 y 540	200-400 (SE)	30Mb-15Gb	400.000 pb-80M	2-23h	<3%, indel

d: días; Gb: gigabases; h: horas; Kb: kilobases; M: millón; Mb: megabases;
pb: pares de bases; PE: paired-end sequencing; SE: single-end sequencing.

La plataforma Ion Torrent se basa en la secuenciación por síntesis mediante adición de un solo tipo de nucleótido (*Single-Nucleotide Addition, SNA*) y registra los cambios en la concentración de protones producidos durante la incorporación de dNTPs en la síntesis del DNA; esta señal de pH se transforma en digital mediante un chip semiconductor. El cambio de pH es detectado por un sensor micropHmetro; la intensidad de dicha señal es proporcional al número de nucleótidos incorporados, aunque tiene una precisión limitada en la medición de longitudes de homopolímeros (Figura 2). La tasa de error global es similar a la de otras plataformas de NGS en regiones no homopolimerizadas; sin embargo, al ser una plataforma basada en SNA, se han detectado más errores de secuenciación en la lectura de inserciones y deleciones (*indels*), siendo más frecuentes para las inserciones. Las regiones de homopolímeros son muy problemáticas para estas plataformas, que carecen de precisión de una sola base en la medición de regiones repetitivas de dNTPs mayores de 6-8 pb⁽⁶⁾. El error en la precisión de las *flow-calls* provoca la detección de *indels* artefactuales con una tasa de hasta el 3%, (*over-called short-homopolymers* y *under-called long-homopolymers*). Afortunadamente, se ha observado que esta tasa de artefactos en la detección de *indels* decrece al aumentar el número de ciclos de secuenciación. Otro error mucho menos frecuente es la inducción de *indels* con una frecuencia alélica muy alta en regiones específicas del genoma, que se reproducen con cierta constancia en todos los experimentos. Este artefacto se observa en menos del 0,1% del total de las bases leídas y generalmente se asocia a *strand-bias*, es decir, tan solo están presentes en una de las direcciones de lectura (directa o reversa). Por lo general, este tipo de artefactos se manifiestan como inserciones de A/T o deleciones de C/G. Por último, se ha descrito que la tasa de errores en sustituciones es del 0,04%–0,1% del total de bases leídas por cada carrera, y se detectan fácilmente ya que se asocian, por lo general, a *strand-bias*.

La plataforma Ion Torrent ofrece diferentes tipos de chips e instrumentos para ajustar el rendimiento del secuenciador a las necesidades del investigador (Tabla 6 y Tabla 7). El rendimiento de estos chips oscila entre ~ 50Mb y 15Gb, con tiempos de

de ejecución de la secuenciación de entre 2 y 7 horas, lo que la convierte en la más rápida de las plataformas actuales. Esto hace que el dispositivo sea adecuado para la secuenciación de grupos de genes (secuenciación dirigida o *targeted sequencing*), incluyendo el perfil de transcriptoma y la identificación de sitios de *splicing* (aunque no al nivel de secuenciadores de lectura larga)⁽⁷⁾. Ion Torrent cuenta con instrumentos de dos capacidades: el Ion *Personal Genome Machine* (PGM) y el Ion S5. Cuando se combinan con la preparación de las librerías Ion Chef y el dispositivo de carga de chips, se transforman en plataformas sencillas de operar.

Tabla 7. Comparación de las plataformas e instrumentos de secuenciación masiva por síntesis. (Modificada de Goodwin, et al., 2016)⁽⁸⁾.

PLATAFORMA	LONGITUD DE LECTURA (pb)	RENDIMIENTO	LECTURAS	TIEMPO POR CARRERA
Sistema de secuenciación SNA				
Ion PGM 314	200 (SE)	30-50Mb	400.000-	23h
	400 (SE)	60-100Mb*	550.000bp*	3,7h*
Ion PGM 316	200 (SE)	300-500Mb	2-3M*	3h
	400 (SE)*	600 Mb-1Gb*		4,9h
Ion PGM 318	200 (SE)	600 Mb-1Gb*	4.5-5M*	4h
	400 (SE)*	1-2Gb*		7,3h
Ion Proton	Up to 200 (SE)	Hasta 10Gb*	60-80M*	2-4h
Ion S5 520	200 (SE)	600 Mb-1Gb	3-5M*	2,5h
	400 (SE)*	1,2-2Gb*		4h*
Ion S5 530	200 (SE)	3-4Gb	15-20M*	2,5h
	400 (SE)*	6-8Gb*		4h*
Ion S5 540	200 (SE)	10-15Gb*	60-80M*	2,5h

Continúa...

...Tabla 7_Continuación

PLATAFORMA	LONGITUD DE LECTURA (pb)	RENDIMIENTO	LECTURAS	TIEMPO POR CARRERA
Sistema de secuenciación CRT				
Illumina MiniSeq <i>Mid output</i>	150 (SE)*	2,1-2,4Gb*	14-16M*	17h*
Illumina MiniSeq <i>High output</i>	75 (SE)*	1,6-1,8Gb	22-25M (SE)*	7h
	75 (PE)*	3,3-3,7Gb	44-50M (PE)*	13h
	150 (PE)*	6,6-7,5Gb*		24h*
Illumina MiSeq v2	36 (SE)	540-610Mb	12-15M (SE)	4h
	25 (PE)	750-850Mb	24-30M (PE)*	5.5h
	150 (PE)	4,5-5,1Gb		24h
	250 (PE)*	7,5-8,5Gb		39h*
Illumina MiSeq v3	75 (PE)	3,3-3,8Gb	44-50M (PE)*	21-56h*
	300 (PE)*	13,2-15Gb*	22-25M (PE)*	
Illumina NextSeq 500/550 <i>Mid output</i>	75 (PE)	16-20Gb	Hasta 260M (PE)*	15h
	150 (PE)*	32-40Gb*		26h*
Illumina NextSeq 500/550 <i>High output</i>	75 (SE)	25-30Gb	400M (SE)*	11h
	75 (PE)	50-60Gb	800M (PE)*	18h
	150 (PE)*	100-120Gb*		29h*
Illumina HiSeq2500 v2 <i>Rapid run</i>	36 (SE)	9-11Gb	300M (SE)*	7h
	50 (PE)	25-30Gb	600M (PE)*	16h
	100 (PE)	50-60Gb		27h
	150 (PE)	75-90Gb		40h
	250 (PE)*	125-150Gb		60h*
Illumina HiSeq2500 v3	36 (SE)	47-52Gb	1.500 (SE)	2d
	50 (PE)	135-150Gb	3.000M (PE)*	5,5d
	100 (PE)	270-300Gb		11d*
Illumina HiSeq2500 v4	36 (SE)	64-72Gb	2.000M (SE)	29h
	50 (PE)	180-200Gb	4.000M (PE)*	2,5d
	100 (PE)	360-400Gb		5d
	125 (PE)*	450-500Gb*		6d*
Illumina HiSeq3000/4000	50 (SE)	105-125Gb	2.500M (SE)*	1-3,5d*
	75 (PE)	325-375Gb		
	150 (PE)*	650-750Gb*		
Illumina HiSeq X	150 (PE)*	900-900Gb por flowcell*	2,6-3.000M (PE)*	<3d*
NovaSeq 5000/6000	50 (PE)*	flowcell s1, s2	>10B*	<19h*
	100 (PE)*	flowcell s1, s2		<29h*
	150 (PE)*	flowcell s1,s2,s3,s4		<40h*

CRT: Cyclic Reversible Termination; d: días; Gb: gigabases; h: horas; Kb: kilobases; M: millón; Mb: megabases; pb: pares de bases; PE: paired-end sequencing; SE: single-end sequencing; SNA: Single-Nucleotide Addition;

*Información según el proveedor.

DISEÑO DEL PANEL DE NGS

ESTRATEGIA DEL DISEÑO

La secuenciación masiva dirigida permite estudiar un grupo de genes o regiones de interés relevantes para la patología en estudio. Existen numerosos métodos de preparación de librerías de paneles de genes para secuenciación dirigida, y su elección dependerá de diversos factores, como la plataforma de secuenciación, el tamaño del panel de genes, el tiempo de respuesta diagnóstica y el coste de secuenciación. Existen dos estrategias principales para la preparación de librerías en secuenciación dirigida: amplicones y captura por hibridación (Figura 3 y Tabla 8).

Los **paneles de amplicones** se basan en la amplificación por PCR de las regiones de interés mediante la unión de *primers* específicos. El protocolo para la obtención de librerías consta de dos pasos principales. En primer lugar, se generarán los amplicones de interés mediante amplificación por PCR. Después se ligará, en los extremos de cada amplicón, un índice (también llamado *index* o *barcode*) específico de cada muestra, y un adaptador compatible con la plataforma de secuenciación que vaya a emplearse. Algunos paneles de amplicones permiten, además, marcar con un índice único (índice molecular, *molecular barcode* -MBC- o *unique molecular barcode* -UMI-) cada fragmento original de DNA. La presencia de UMIs, permite identificar y eliminar durante el análisis los duplicados de PCR generados en la preparación de las librerías.

Los **paneles de captura** por hibridación se basan en el enriquecimiento del DNA mediante captura con sondas específicas. El protocolo incluye tres etapas. Primero, fragmentación del DNA, bien mecánica o enzimática. Segundo, generación y amplificación de las librerías mediante la ligación de los índices y adaptadores. Tercero, enriquecimiento de la librería en las regiones de interés mediante hibridación y **captura** con sondas complementarias.

Tabla 8. Características principales de los paneles de amplicones y de captura.

	AMPLICIONES	CAPTURA
Fundamento	Amplificación de las regiones de interés mediante la unión de <i>primers</i> específicos.	El DNA se enriquece en las regiones de interés por captura con sondas complementarias.
Tamaño del panel	Recomendado para librerías pequeñas ($\leq 200\text{Kb}$).	Compatible para librerías de hasta 200Mb (el rango de tamaños compatibles vendrá determinado por la química y la plataforma empleada).
Tipo de cobertura	Escalonada: útil para la detección de variantes puntuales y pequeñas <i>indels</i> .	Cobertura en forma de campana: permite la detección de variantes puntuales, pequeñas <i>indels</i> , duplicaciones exónicas y, en función del tamaño, cambios en el número de copias.
Material de partida*	La mayoría a partir de 50ng (algunos protocolos permiten bajar hasta 10-20ng).	La fragmentación enzimática permite trabajar a partir de 50ng de DNA, mientras que la fragmentación por sonicación requiere una cantidad superior (>500ng en la mayoría de casos, pudiendo llegar hasta 1.000ng).
Coste	Depende del tamaño de la librería y de la química empleada. De forma general y en condiciones similares: paneles más económicos.	Depende del tamaño de la librería y de la química empleada. De forma general y en condiciones similares: paneles más caros.
Ventajas	No se secuencian zonas <i>off-target</i> . Protocolos más cortos (2-3 días).	Mayor uniformidad de cobertura entre regiones. Más probabilidad de capturar regiones problemáticas por arrastre debido a la fragmentación aleatoria.
Desventajas	Menor uniformidad de cobertura entre regiones. Poca o ninguna cobertura en regiones difíciles de amplificar.	Captura de regiones <i>off-target</i> , que consumirán capacidad de secuenciación. Protocolos largos (3-4 días).

*La cantidad necesaria de DNA de partida dependerá de la química empleada. Inicialmente, deberían seguirse las recomendaciones de cada protocolo, aunque éstas pueden optimizarse: es posible que haya protocolos que puedan funcionar con cantidades menores de DNA de partida, y otros que incluso mejoren al partir de una mayor cantidad de DNA. En caso de necesitarse la modificación de este parámetro, antes se debe validar mediante secuenciación de una misma muestra partiendo de distintas cantidades de DNA, y comparando los resultados.

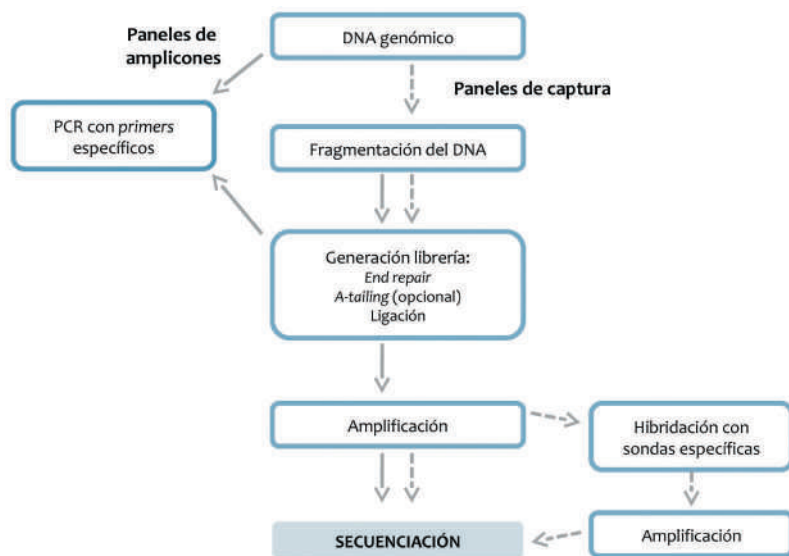


Figura 3. Principales pasos del protocolo de generación de librerías en paneles de amplicones y de captura (Modificada de Rehm et al., 2013)⁽⁹⁾.

DISEÑOS COMERCIALES Y PERSONALIZADOS

A la hora de secuenciar un panel, se puede elegir entre utilizar un panel comercial, con un diseño cerrado, o realizar un diseño personalizado (*panel custom*) (Tabla 9). La ventaja de los paneles comerciales es que no requieren una etapa de diseño previa y están optimizados, aunque requieren un proceso de comprobación o validación en cada laboratorio. En cambio, los paneles personalizados requieren, inicialmente, un trabajo previo de selección de genes y/o regiones y un diseño de los *primers*/sondas correspondientes. Aunque existen herramientas elaboradas con este propósito que son de gran ayuda, este tipo de paneles requiere de un proceso largo (generalmente de varios meses) de optimización y validación del diseño. Por otro lado, la ventaja de estos últimos es la flexibilidad que ofrecen, ya que se pueden incluir las regiones que uno desee e ir modificando este diseño con el tiempo (añadir o quitar regiones) en función de las necesidades de cada patología.

Tabla 9. Características de los paneles comerciales mieloides existentes y de las químicas más comúnmente utilizadas para los paneles personalizados.

PANEL	QUÍMICA	ESTRATEGIA	PLATAFORMA	DETALLES DEL DISEÑO
Paneles comerciales				
<i>Archer FusionPlex Myeloid Kit</i>	Archer	Amplicones (RNA-Seq)	Illumina Ion Torrent	Regiones seleccionadas de 84 genes para el estudio de fusiones, mutaciones puntuales y niveles de expresión. Librerías con índice molecular único (UMIs).
<i>Clear Seq AML</i>	Agilent	Amplicones	Illumina Ion Torrent	Regiones seleccionadas de 20 genes.
<i>Ion AmpliSeq AML Panel</i>	Thermo Fisher	Amplicones	Ion Torrent	Regiones seleccionadas de 19 genes (264 amplicones, 22Kb).
<i>Myeloid Solution by Sophia</i>	Sophia Genetics	Captura	Illumina	Regiones seleccionadas de 30 genes. Posibilidad de añadir nuevas regiones.
<i>QiaSeq Targeted DNA Human Myeloid Neoplasms Panel</i>	Qiagen	Amplicones	Illumina Ion Torrent	Regiones seleccionadas de 50 genes (2536 amplicones, 236Kb).
<i>QiaSeq Targeted RNAscan Human Leukemia Panel</i>	Qiagen	Amplicones (RNA-Seq)	Illumina Ion Torrent	Regiones seleccionadas de 24 genes para el estudio de fusiones. Librerías con índice molecular único (UMIs).
<i>SureSeq Myeloid Panel</i>	Oxford Gene Technology	Amplicones	Illumina	Regiones seleccionadas de 25 genes.
<i>TruSight Myeloid Panel</i>	Illumina	Amplicones	Illumina	Regiones seleccionadas de 54 genes (568 amplicones, 141Kb).
Tecnologías disponibles para el diseño de paneles personalizados				
<i>AmpliSeq</i>	Thermo Fisher	Amplicones	Ion Torrent	Hasta 24.000 amplicones, desde 1Kb hasta 5Mb.
<i>Haloplex</i>	Agilent	Amplicones	Illumina Ion Torrent	Hasta 200.000 amplicones o 5Mb.
<i>HaloPlex HS</i>	Agilent	Amplicones	Illumina Ion Torrent	Hasta 5Mb. Librerías con índice molecular único (UMIs).
<i>HeatSeq</i>	Roche (Nimblegen)	Amplicones	Illumina	Hasta 15Mb. Librerías con índice molecular único (UMIs) basado en sondas MIP (sondas moleculares de inversión).
<i>Nextera Rapid Capture Custom Enrichment</i>	Illumina	Captura	Illumina	Desde 0.5Kb hasta 15Mb.
<i>QiaSeq Targeted DNA</i>	Qiagen	Amplicones	Illumina Ion Torrent	Hasta 12.000 amplicones. Librerías con índice molecular único (UMIs).
<i>SeqCap EZ Choice</i>	Roche (Nimblegen)	Captura	Illumina	Hasta 7Mb la versión normal y hasta 200Mb la versión XL.
<i>SureSelect</i>	Agilent	Captura	Illumina Ion Torrent	Desde 1Kb hasta 24Mb.
<i>TruSeq Custom Amplicon</i>	Illumina	Amplicones	Illumina	Hasta 1536 amplicones, desde 2Kb hasta 650Kb.

RECURSOS PARA EL DISEÑO DE PANELES CUSTOM

Programas para hacer el diseño

La mayoría de compañías ofrecen programas gratuitos *online* para el diseño de paneles personalizados. Algunos ejemplos son: *Design Studio* (Illumina), *SureDesign* (Agilent), *NimbleDesign* (Roche), *Ion AmpliSeq Designer* (Thermo Fisher). La mayoría de estos programas funcionan siguiendo pasos similares (Figura 4).

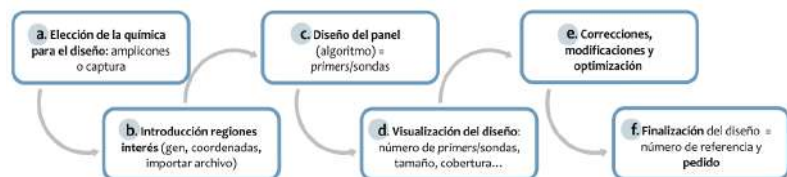


Figura 4. Esquema del proceso de diseño de un panel personalizado (custom). a. Elección de la química para el diseño del panel (amplicones/captura). b. Introducción de las regiones de interés: las regiones se pueden introducir, una a una, por genes o por posición genómica (cromosoma y coordenada). Otra opción es importar un archivo .bed o .csv que contenga todas las posiciones (el formato de este archivo vendrá determinado por el programa que se esté utilizando). c. Diseño del panel: una vez introducidas las regiones de interés, el programa diseña mediante un algoritmo propio un conjunto de *primers/sondas* adecuados para amplificar o capturar esas regiones. d. Visualización del diseño: una vez hecho el diseño, el usuario puede comprobar el número de *primers/sondas* que se van a necesitar, el tamaño final del panel, la cobertura del diseño (ver qué regiones están cubiertas y cuáles no) y otros parámetros. Este tipo de programas suelen tener alertas que avisan de las regiones problemáticas o difíciles de capturar, como por ejemplo las zonas repetitivas o ricas en GC. e. Correcciones y optimización: una vez terminado el diseño, uno puede hacer las modificaciones que desee para intentar corregir los posibles problemas. Por ejemplo, aumentar la densidad de sondas en zonas difíciles de capturar, o mover las coordenadas de regiones difíciles de amplificar para que haya nuevas opciones para colocar los *primers/sondas* que van dirigidos a esa región. Las mismas compañías, muchas veces, ofrecen un soporte para la optimización del diseño de los paneles. Una vez hechas las correcciones necesarias, el programa recalcula y los resultados se pueden volver a visualizar y/o modificar. f. Realización del pedido: una vez terminado este proceso, se finaliza el diseño, que queda asociado a un número de referencia. Con este número, y muchas veces a través de la propia aplicación, se puede realizar el pedido de los reactivos.

Programas de visualización del panel diseñado

Existen otros recursos que pueden ser útiles a la hora de visualizar el panel una vez ya diseñado. Uno de ellos es el *genome browser* (navegador genómico) de la Universidad de California en Santa Cruz (UCSC), que es abierto y se utiliza *online*. El *UCSC Genome Browser* es una herramienta para la visualización y consulta de numerosos tipos de datos genómicos públicos, y que además permite la carga y visualización de datos propios. Otra opción popular es el programa gratuito *Integrative Genome Viewer* (IGV), desarrollado por el *Broad Institute*, en el cual se pueden importar y visualizar archivos con información de localización genómica, como los archivos .bed, entre otros.

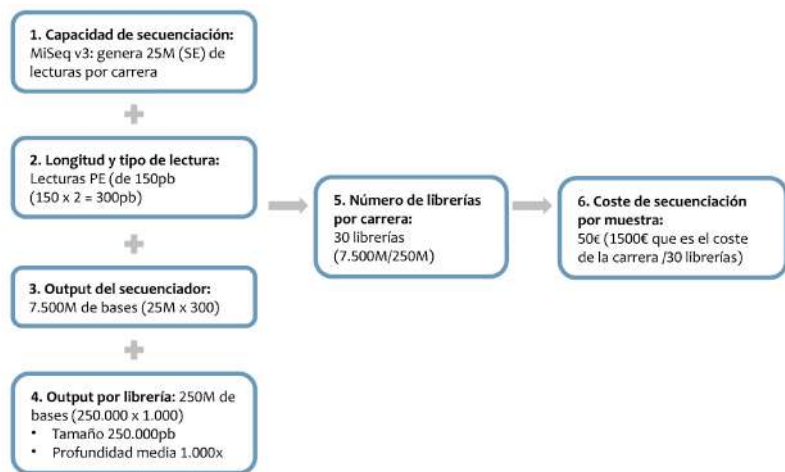
Parámetros del diseño que influirán en la secuenciación

A la hora de realizar el diseño de un panel personalizado hay que tener en cuenta una serie de parámetros que pueden influir en la secuenciación y en el coste final por muestra:

1. **Capacidad de secuenciación:** cantidad de lecturas que genera el secuenciador por carrera teniendo en cuenta solamente un sentido de la secuenciación, *single-end* sequencing (SE).
2. **Longitud de lectura y tipo de lectura:** *single-end* sequencing (SE) o *paired-end* sequencing (PE).
3. **Output del secuenciador:** cantidad de datos que genera el secuenciador (cantidad de lecturas multiplicado por el tamaño de lectura).
4. **Output necesario por librería** (o muestra), depende de:
 - a. Tamaño del panel (incluyendo regiones *on-* y *off-target*)
 - b. Profundidad media.
5. **Número de librerías por carrera** número de librerías que se podrán cargar juntas para ser secuenciadas en una misma carrera (multiplexar): se calcula como *output* total del secuenciador dividido por *output* necesario por librería.
6. **Coste de secuenciación por librería:** se calcula como el coste por carrera dividido por número de librerías por carrera.

De forma resumida, en primer lugar se debe tener en cuenta el secuenciador y tipo de lectura, que vendrá determinada por la química elegida para el diseño. En segundo lugar, se deben considerar las características del panel (tamaño y profundidad

deseada). Definiendo estos parámetros se podrá determinar el número de librerías que se podrán incluir en cada carrera, lo cual nos dará una idea del coste de secuenciación por cada muestra. La Figura 5 recoge un ejemplo que ilustra este tipo de cálculos.



SE: single-end sequencing; PE: paired-end sequencing

Figura 5. Parámetros del diseño que influyen en el coste de secuenciación. Estos parámetros se deben tener en cuenta a la hora de calcular el número de muestras que se pueden multiplexar en una carrera, lo que influye sustancialmente en el coste final de la secuenciación para cada muestra.

PREPARACIÓN DE LIBRERÍAS DE NGS

En estas guías se comentan únicamente los puntos críticos y recomendaciones generales comunes a todos los protocolos. A continuación se especifican los métodos utilizados en algunos centros participantes en el GESMD. En todo momento deben seguirse las recomendaciones y controles de calidad específicos del protocolo realizado descritos por el fabricante de los reactivos.

Tabla 10. Protocolos utilizados por los miembros del GESMD

PROTOCOLO

Agilent SureSelect QXT/QT
IonAmpliseq AML Panel
IonAmpliseq DNA Custom Panel
Nextera Rapid Capture Custom Enrichment
Myeloid Solution by Sophia Genetics (QiaSeq FX kit)
Nimblegen HeatSeq
Nimblegen SeqCap EZ Choice
QiaSeq Targeted DNA panel
TruSeq Custom AmpliconTru
Sight Myeloid Sequencing Panel

MUESTRA DE PARTIDA

La muestra de partida para la preparación de librerías dependerá del objetivo del panel. Estas guías se centran en el uso de paneles para la detección de mutaciones somáticas, donde la muestra de partida es DNA tumoral. Para detectar reordenamientos o genes de fusión, así como para cuantificar los niveles de expresión de determinadas dianas, la muestra de partida será cDNA. Algunos paneles permiten también detectar reordenamientos a partir de DNA genómico.

Muestra biológica

Para el estudio de hemopatías de origen mielode normalmente se extrae DNA de muestras de aspirado de médula ósea preferentemente, aunque también podría emplearse sangre periférica.

Métodos de extracción

Según nuestra experiencia, el método de extracción de DNA influye mínimamente en la calidad de las librerías. En el GESMD se ha utilizado DNA extraído con los siguientes métodos con buenos resultados: *QIAmp DNA Micro/Mini Kit* (Qiagen), *Maxwell 16 Blood/Cell DNA Purification Kit* (Promega), *Qiasymphony DNA Midi Kit* (Qiagen).

Calidad del DNA

Aunque la muestra idónea es el DNA extraído de muestra fresca, se puede trabajar con otro tipo de muestras, siempre y cuando el DNA extraído cumpla los requisitos mínimos de calidad (Tabla 11). Según nuestra experiencia, la pureza del DNA no es crítica para la obtención de buenas librerías. Las muestras congeladas, así como las procedentes de células fijadas en Carnoy (preservadas a -80°), en general funcionan bien, sin embargo, es aconsejable realizar siempre un estudio de integridad para evaluar si es adecuado para secuenciación masiva. La evaluación de la integridad del DNA se realiza generalmente mediante ensayos de *Tape Station* (Agilent) o PCR cuantitativa. Esto es especialmente importante si las muestras proceden de material parafinado o se van a utilizar paneles de captura. El DNA parafinado normalmente solo es compatible con paneles de amplicones, aunque conviene comprobar en cada caso las recomendaciones del fabricante de los reactivos para la preparación de librerías, que indicarán el método más adecuado.

Tabla 11. Muestras especiales para la preparación de librerías

MUESTRA DE PARTIDA	TIPO DE PANEL	COMENTARIOS
Células de sangre periférica o médula ósea congeladas a -80°C o criopreservadas	Amplicones y captura	Buenos resultados.
Células plasmáticas preservadas en suero-DMSO	Amplicones y captura	Buenos resultados.
Material fijado en Carnoy (preservado a -20°C)	Amplicones y captura	Buenos resultados.
Muestras parafinadas	Amplicones	Buenos resultados; atención a las falsas variantes introducidas por la desparafinización ⁽¹⁰⁾ .
DNA amplificado mediante procesos de <i>whole genome amplification</i> (WGA)	Amplicones y captura	Resultados variables según las regiones a secuenciar, atención a las falsas variantes introducidas por el proceso de WGA, especialmente a bajas frecuencias alélicas ⁽¹¹⁾ .

Los instrumentos más aceptados para evaluar la pureza del DNA son:

- *Nanodrop* (Thermo Fisher Scientific): ratios 260/280 y pureza ~1.8, y ratio 260/230 ~2.0-2.2
- *Tape Station* (Agilent): integridad del DNA, valor superior a 8.

Cuantificación del DNA

La cantidad de DNA de partida es crítica para el buen funcionamiento de la preparación de librerías. Es imprescindible utilizar un método fluorimétrico de cuantificación de DNA de doble cadena. Los más aceptados son *Qubit fluorometer* (Thermo Fisher Scientific) y el ensayo *Quant-iT PicoGreen* (Thermo Fisher Scientific). Se desaconseja la utilización de métodos espectrofotométricos como *Nanodrop* (Thermo Fisher Scientific).

La cantidad mínima de DNA de partida para la preparación de librerías viene determinada por el fabricante de los reactivos; en general se utiliza, como mínimo, 50ng de DNA. Para captura se requiere una cantidad bastante superior, como mínimo 500ng.

PREPARACIÓN DE LIBRERÍAS

Durante la preparación de librerías se deberá seguir los controles de calidad en distintos puntos recomendados por el fabricante de los reactivos. Una vez realizadas las librerías, se evaluará su calidad y cantidad, y en los métodos de captura se evaluará además el rendimiento de captura.

Calidad de las librerías

La calidad de las librerías está determinada por su tamaño y la ausencia de adaptadores y oligos de PCR libres. Es recomendable evaluarla mediante electroforesis capilar de alta resolución. Los instrumentos más aceptados son el 2100 *Bioanalyzer Instrument* (Agilent), la 2200 o 4200 *Tape Station* (Agilent) y el *QIAxcel Advanced System* (Qiagen). Estos instrumentos generan un electroferograma en el que se comprobará que:

- El tamaño medio de las librerías es el especificado por el fabricante.
- No existe ningún pico de tamaño correspondiente al tamaño de los adaptadores o de los oligos de PCR.

Cuantificación de las librerías

Se utilizan los mismos métodos que para la cuantificación de DNA, es decir, métodos fluorimétricos de cuantificación de DNA de doble cadena. Otro método recomendado en muchos protocolos para la cuantificación de las librerías, debido a su precisión, es el *KAPA Library Quantification Kit* (*KAPA Biosystems*, Roche). Este método requiere más tiempo y es más caro que los mencionados anteriormente, pero puede resultar útil en la fase de optimización del panel. La cantidad mínima de librería para secuenciar viene determinada por la plataforma de secuenciación.

Rendimiento de captura

Como se ha mencionado anteriormente, conviene evaluar el rendimiento de la captura en los paneles que usan este sistema de enriquecimiento. El método específico de evaluación de este rendimiento es recomendado por los fabricantes de los reactivos.

Optimización del protocolo

Para algunos tipos de química, es posible optimizar la cantidad/volumen de los reactivos empleados en el protocolo de preparación de librerías. Esta optimización suele ser útil para que el kit sea más rentable. Sin embargo, el criterio económico no es suficiente para modificar las condiciones de uso de un kit, sino que las optimizaciones del protocolo deben validarse, mediante secuenciación de una misma muestra con las distintas condiciones, y posterior comparación de los resultados.

PARÁMETROS DE SECUENCIACIÓN

Entre los parámetros principales asociados a la secuenciación masiva dirigida, independientemente de los tipos de librería y plataforma empleados, se encuentran la profundidad de cobertura, el porcentaje de lecturas mapeadas en dichas regiones y la uniformidad conseguida en la distribución de las mismas. Todos estos parámetros dependen de la optimización del proceso de secuenciación. Además, estos parámetros están determinados, en ocasiones de forma más importante, por otras variables como son la calidad del DNA/RNA de partida, la preparación adecuada de la librería, el diseño correcto de los amplicones o sondas de captura e incluso por variables relativas a estimaciones iniciales erróneas del número de lecturas necesarias (a veces se hacen

estimaciones demasiado restrictivas a fin de minimizar costes). Por último, la calidad de las lecturas es la variable que evalúa de manera más directa la fase de secuenciación.

1.PROFUNDIDAD DE COBERTURA

Para cada muestra se debe establecer antes de su secuenciación la profundidad de cobertura media óptima, es decir, el número de lecturas que se desea realizar en la secuencia de referencia diana. Como regla general, cuanto mayor sea la profundidad, mayor será la sensibilidad y la confianza en la identificación de una variante. La representación de este parámetro se realiza mediante histogramas, que también pueden emplearse para calcular la uniformidad de la secuenciación.

No se debe confundir la profundidad de cobertura con la amplitud de cobertura, que se refiere a cuántas regiones del panel están cubiertas por el diseño a una profundidad adecuada.

Una regla habitual para calcular el número de lecturas necesario para un panel de genes nuevo, es hacer una estimación en la que se prevé una cobertura cuatro veces menor que la buscada para los fragmentos con menos lecturas. Este dato servirá para ajustar la cobertura media a fin de que la mínima sea superior a un valor establecido. El cálculo de la profundidad de cobertura se realiza teniendo en cuenta la longitud de la secuencia diana, la celularidad de la muestra (si no se ha realizado un aislamiento del tipo celular concreto), y la eventual presencia de clones minoritarios de interés. En ambos casos se requerirá una mayor sensibilidad.

En cualquier caso, frecuentemente será necesario reevaluar la profundidad de cobertura tras los resultados de las primeras muestras secuenciadas con un determinado panel de genes. Esto se debe a que tanto la inespecificidad como la deficiencia de amplificación o captura de algunas regiones precisan de comprobación experimental. La corrección de estas deficiencias requiere aumentar el número de lecturas a fin de alcanzar la cobertura mínima establecida para todos los fragmentos. Los valores de profundidad de cobertura recomendados son:

- **Profundidad de cobertura mínima de las regiones de interés:** se recomienda que $\geq 95\%$ de las bases detectadas estén cubiertas por un mínimo de 100 lecturas ($>100X$). Este sería el punto de corte para dar una muestra como válida para poder proceder con el análisis.

- **Profundidad de cobertura media:** es el valor mínimo de lecturas promedio de las regiones secuenciadas para cada muestra. Se recomienda llegar a una profundidad de cobertura media $\geq 1000X$. Por ejemplo, la detección de variantes somáticas con una profundidad de 1000 lecturas por muestra (1000X) alcanzaría una sensibilidad media del 2,5%, si se requiere la presencia de la variante en al menos 25 lecturas.

A continuación se dan algunas expresiones útiles para calcular la profundidad de cobertura para una muestra y para el cálculo del número de muestras que es posible secuenciar en una carrera:

- Estimación del número de lecturas en función de la cobertura media,

$$n^{\circ} \text{ lecturas muestra} = n^{\circ} \text{ amplicones} \times \text{cobertura media}$$

- Cobertura muestra con panel de captura,

$$n^{\circ} \text{ lecturas muestra} = \frac{\text{bases secuencia diana} \times \text{cobertura media}}{\text{tamaño medio lecturas}}$$

- Número de muestras por carrera,

$$n^{\circ} \text{ muestras} = \frac{\text{lecturas totales carrera}}{n^{\circ} \text{ lecturas muestra}}$$

$$n^{\circ} \text{ muestras} = \frac{\text{bases totales output carrera}}{n^{\circ} \text{ lecturas muestra} \times \text{tamaño medio lectura}}$$

$$n^{\circ} \text{ muestras} = \frac{\text{bases totales output carrera}}{\text{bases secuencia diana} \times \text{cobertura media}}$$

El número de amplicones y las bases de la secuencia diana son datos proporcionados en el diseño del panel. Las lecturas totales de la carrera y las bases totales en el **output** de la carrera son datos proporcionados por el fabricante, y asociados a la plataforma, los kits de secuenciación (reactivos y chip o *flow cell*) y el número de ciclos de la carrera. Las lecturas

totales equivalen al número de pocillos enriquecidos clonales (Ion Torrent) o número de *clusters* (Illumina). Sin embargo, estas variables facilitadas conviene sustituirlas por las obtenidas habitualmente en el laboratorio cuando se disponga de la información.

Una profundidad de cobertura más baja de lo esperada para una librería está relacionada frecuentemente con una cuantificación defectuosa, bien por exceso o por defecto, previa a la amplificación clonal. Esto se verá reflejado en el valor de policlonalidad obtenido, en una carga de pocillos/*clusters* escasa, o en los equipos Ion Torrent en forma de un valor bajo de enriquecimiento (<95%, debido a la presencia de beads sin fragmentos unidos en la fase de emulsión PCR).

2. PORCENTAJE DE LECTURAS MAPEADAS CORRECTAMENTE

Este parámetro depende principalmente del diseño del panel y suele ser más o menos constante entre muestras y carreras. Aunque la preparación de la librería, la amplificación clonal de la misma y su secuenciación sean correctas, la amplificación de fragmentos inespecíficos en PCRs altamente multiplexadas es inevitable. También en los paneles de captura se hibridan numerosos fragmentos que mapean en regiones del genoma fuera de las regiones diana. El porcentaje mapeado correctamente en paneles de amplicones suele ser superior al 90%, mientras que en los paneles de captura este valor puede disminuir sensiblemente. En estos casos, se recomienda que el porcentaje mapeado *on-target* sea superior al 70%. Los paneles comerciales de ambos tipos, tras mejoras sucesivas, consiguen un aumento del porcentaje mapeado en las regiones diana. Por supuesto, la repetitividad en el genoma de las secuencias de las regiones diana es un factor importante para obtener resultados pobres para el mismo.

Sin embargo, una disminución manifiesta del valor habitual del parámetro puede indicar un error en la preparación de la librería, o bien un problema en las fases de amplificación o secuenciación que disminuya la calidad de las lecturas provocando un alineamiento incorrecto. El cálculo del porcentaje de lecturas mapeadas a la región de interés para una muestra se puede calcular con la siguiente ecuación:

$$\% \text{ lecturas correctamente mapeadas} = \frac{\text{lecturas mapeadas diana}}{\text{lecturas totales}}$$

3. UNIFORMIDAD DE LAS LECTURAS

En este parámetro también pueden influir, al igual que en el porcentaje de lecturas mapeadas, problemas en la realización de cualquiera de las tres fases de secuenciación masiva. Fallos en la uniformidad suelen estar relacionados con el contenido de G+C de la secuencia y la eficiencia de amplificación de cada amplicón o fragmento. Pero también pueden estar asociados a una profundidad de cobertura media escasa, lo que se hace más evidente en los fragmentos menos representados. La baja calidad del material genético de partida, como es el caso de las muestras parafinadas, conduce asimismo a histogramas de lecturas poco uniformes.

La uniformidad se evaluará en función de la plataforma usando como criterios mínimos $\geq 80\%$ para los equipos de Ion Torrent y una captura uniforme para los equipos de Illumina (Tabla 12).

Tabla 12. Parámetros de calidad mínimos

COBERTURA MÍNIMA	COBERTURA MEDIA	LECTURAS MAPEADAS CORRECTAMENTE	UNIFORMIDAD
>95% bases >100X	>1000X (variante >25 lecturas)	Amplicones > 90% Captura > 70%	Ion Torrent $\geq 80\%$; Illumina, basarse en los parámetros de calidad anteriores.

Por último, cabe señalar que en ocasiones la falta relativa de uniformidad de algunas regiones no se debe a un error, sino que puede producirse en algunas muestras por variaciones importantes en el número de copias de esa región.

4. ESCALA DE CALIDAD PHRED

Las escalas de calidad de secuenciación miden la calidad de *base calling* (asignación de bases), definida como la probabilidad de que una base sea llamada/asignada incorrectamente. La escala más frecuentemente usada se basa en un algoritmo *Phred* similar al desarrollado en secuenciación Sanger calculado con la siguiente fórmula, siendo P la probabilidad de error (Tabla 12).

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

Tabla 13. Escala de calidad *Phred*

ESCALA PHRED	PROBABILIDAD DE BASE INCORRECTA	EXACTITUD DE BASE
Q10	1 de 10	90%
Q20	1 de 100	99%
Q30	1 de 1000	99,9%

A pesar de que la evaluación de los errores de secuenciación varía entre fabricantes, de forma general se requiere un valor mmo de $Q=30$ (Tabla 13).

A continuación (Tabla 14) se resume la capacidad en número de muestras para un panel de 100Kb y con una cobertura media de 1000x en las diferentes plataformas.

Tabla 14. Ejemplos de capacidad mínima de número de muestras por carrera para un panel de 100Kb con una cobertura media de 1000X.

PLATAFORMA	LONGITUD DE LECTURA	RENDIMIENTO	LECTURAS	Nº MUESTRAS (PANEL 100KB, 1000X)
ILLUMINA MiniSeq Mid output	150 (SE)	2,1-2,4Gb	14-16M	21
ILLUMINA MiniSeq High output	75 (SE)	1,6-1,8Gb	22-25M (SE)	16
	75 (PE)	3,3-3,7Gb	44-50M (PE)	
	150 (PE)	6,6-7,5Gb		33
ILLUMINA MiSeq v2	36 (SE)	540-610Mb	12-15M (SE)	4
	25 (PE)	750-850Mb	24-30M (PE)	3
	150 (PE)	4,5-5,1Gb		18
	250 (PE)	7,5-8,5Gb		30
ILLUMINA MiSeq v3	75 (PE)	3,3-3,8Gb	44-50M (PE)	16
	300 (PE)	13,2-15Gb		66
ILLUMINA NextSeq 500/550 Mid output	75 (PE)	16-20Gb	Up to 260M (PE)	80
	150 (PE)	32-40Gb		160
ILLUMINA NextSeq 500/550 High output	75 (SE)	25-30Gb	400M (SE)	250
	75 (PE)	50-60Gb	800M (PE)	250
	150 (PE)	100-120Gb		500
ION PGM 314	200 (SE)	30-50Mb	400.000-550.000bp	0,3
	400 (SE)	60-100Mb		0,6
ION PGM 316	200 (SE)	300-500Mb	2-3M	3
	400 (SE)	600 Mb-1Gb		6
ION PGM 318	200 (SE)	600Mb-1Gb	4.5-5M	6
	400 (SE)	1-2Gb		10
ION Proton	200 (SE)	10-15Gb	80M	100

VALIDACIÓN DE PANELES DE NGS

Para el desarrollo e implantación clínica de un nuevo test de diagnóstico molecular mediante NGS para SMD y LMMC, es necesario estimar la sensibilidad (verdaderos positivos) y especificidad (falsos negativos) realizando un plan de validación de la prueba diagnóstica. Es necesario estimar la sensibilidad y especificidad (asegurar que lo que se detecta son verdades positivas y que no se pierdan falsos negativos) realizando un plan de validación de la prueba diagnóstica. La validación consta de dos fases: la validación técnica y la validación clínica.

VALIDACIÓN TÉCNICA

El objetivo de esta fase es conocer cuáles y cómo están cubiertas las regiones del panel diseñado. Se trata de evaluar la calidad del diseño del panel y detectar en esta primera fase las regiones que quedan sin cubrir o que quedan mal cubiertas. Dependiendo de estos resultados será conveniente rediseñar el panel, es decir, añadir sondas o amplicones, para asegurar que las regiones de interés se secuencian bien. Se deberán tener en cuenta las siguientes consideraciones:

- Para esta validación se seleccionarán muestras control (con cariotipo normal o, al menos, que sean diploides en las regiones de interés).
- El DNA debe ser de la máxima calidad posible, cumpliendo los requisitos de cantidad y calidad recomendados por la plataforma y tecnología que se vaya a utilizar.
- Si el panel incluyera cambios en el número de copias (CNVs) es importante incluir tanto mujeres como hombres.
- La generación de las librerías y secuenciación de las mismas se llevará a cabo siguiendo el protocolo del fabricante.
- En el análisis informático de las secuencias obtenidas se determinará:
 - El porcentaje de las bases cubiertas a distintas profundidades (500X, 400X, 300X...). El nivel de profundidad es clave y necesario para la identificación de mosaicismos. Por ello, puede que la validación técnica indique que alguna muestra necesita más cobertura y habrá que reajustar el número de muestras por carrera.

- Se detectará también la amplitud de cobertura, los genes o exones bien cubiertos; hay regiones que por características intrínsecas de la secuencia no se cubren bien (por ejemplo regiones ricas en G+C). En este caso se pueden tomar dos decisiones: eliminarlas del diseño del panel o, en caso de ser regiones muy interesantes, rediseñarlas.

- Las nuevas variantes identificadas se validarán mediante otras técnicas: secuenciación Sanger, ARMS-PCR, ASO-PCR, PCR digital, otras plataformas de NGS, etc., teniendo en cuenta para la interpretación de los resultados de concordancia los límites de detección de estas técnicas.

Mediante la realización de esta segunda fase se podrá valorar la capacidad de detección.

A continuación se describen comprobaciones que es recomendable realizar:

- Réplicas técnicas:
 - Secuenciar dos veces la misma librería en el mismo instrumento.
 - Secuenciar dos librerías de la misma muestra hechas por dos personas diferentes.
 - Secuenciar una librería en dos instrumentos distintos (mismo modelo, misma química).
- Réplicas biológicas:
 - Secuenciar dos librerías hechas a partir del mismo DNA.
 - Secuenciar dos librerías de dos extracciones de DNA de la misma muestra.

VALIDACIÓN CLÍNICA

El objetivo de esta fase es determinar si el panel diseñado es capaz de detectar alteraciones en las regiones de diseño.

- Para esta fase de la validación se seleccionarán muestras de pacientes con SMD o LMMC que ya hayan sido bien caracterizados molecularmente y que presenten alguna alteración patogénica, cambio de un solo nucleótido (*single nucleotide variant*, SNV) o *indel*, en las regiones de interés. Las muestras se analizarán con el panel desarrollado para ver la correlación con lo detectado anteriormente por otras técnicas. Mediante el panel de NGS deberían detectarse, como mínimo, todas aquellas alteraciones previamente detectadas mediante técnicas moleculares convencionales. Probablemente se detectarán, también, nuevas alteraciones que pasaron inadvertidas por los métodos convencionales debido a la menor sensibilidad de detección de éstos.

REFERENCIAS

1. Bentley DR, Balasubramanian S, Swerdlow HP, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59.
2. Dohm JC, Lottaz C, Borodina T, *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 2008;36(16):e105.
3. Harismendy O, Ng PC, Strausberg RL, *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*. 2009;10(3):R32.
4. Nakamura K, Oshima T, Morimoto T, *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*. 2011;39(13):e90.
5. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol*. 2011;12(11):R112.
6. Loman NJ, Misra R V, Dallman TJ, *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(5):434-439.
7. Li S, Tighe SW, Nicolet CM, *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol*. 2014;32(9):915-925.
8. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351.
9. Rehm HL, Bale SJ, Bayrak-Toydemir P, *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733-747.
10. Oh E, Choi Y-L, Kwon MJ, *et al.* Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PLoS One*. 2015;10(12):e0144162.
11. Hou L, Sun N, Mane S, *et al.* Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genet Epidemiol*. 2017 Feb;41(2):152-162.



ANÁLISIS Y FILTRADO DE VARIANTES

GUÍA DE APLICACIÓN CLÍNICA DE LA SECUENCIACIÓN MASIVA EN SÍNDROMES MIELODISPLÁSICOS Y LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:

Francisco Fuster-Tormo, Paolo Maietta, Jesús María Hernández-Sánchez, Marta Fernandez-Mercado, Iria Vázquez, Laura Palomo, Bárbara Tazón-Vega, Mariam Ibáñez, Lurdes Zamora, María Abáigar, Joaquín Martínez, Francesc Solé y Esperanza Such.

ANÁLISIS Y FILTRADO DE VARIANTES

SECUENCIACIÓN

Las tecnologías de NGS comprenden una compleja interacción entre química, *hardware* y sensores ópticos o semiconductores. A esta interacción hay que agregarle el uso de programas (*software*) que analizan los datos de los sensores para predecir las bases que contiene cada secuencia a estudio, paso conocido como *base calling*⁽¹⁾.

PLATAFORMA ILLUMINA

Como se explica en el capítulo anterior, el *base calling* de las plataformas de Illumina consiste en el procesamiento de las imágenes que se captan en cada ciclo de la fluorescencia emitida durante la síntesis de DNA.

El *software* del secuenciador controla la generación de imágenes en las distintas áreas o mosaicos de la *flow cell*. De esta forma, se toma una imagen de cada base (en todos los *clusters*), agrupando cuatro imágenes de mosaico por cada ciclo de secuenciación. Las imágenes se van recogiendo, junto con un fichero de configuración que contiene información de la carrera, y se procesan con el *software* interno para extraer las intensidades y las localizaciones de los *clusters*. Cada imagen refleja la fluorescencia en uno de los cuatro canales correspondientes a los cuatro posibles nucleótidos: G, A, T o C. El programa de detección de *clusters* busca el mayor área de intensidad del *cluster* y mide el umbral de ruido de fondo dentro de la *flow cell*. Una vez que todos los *clusters* han sido detectados, se ordenan y almacenan las intensidades. Éstas son luego procesadas por el *software* de *base calling*, junto con las localizaciones previamente guardadas. Es aquí donde se establece la secuencia nucleotídica (*base call*) y se determina la calidad de cada una de las bases. Al final de este proceso, se genera un fichero fastq (Tabla 17) por cada cadena de DNA.

COMPROBACIÓN DE LA CALIDAD DE LA CARRERA EN LA PLATAFORMA ILLUMINA

Junto con las muestras secuenciadas, Illumina proporciona al usuario una serie de datos y gráficos que permiten evaluar la calidad de la secuenciación. Estos datos están disponibles desde el inicio de la secuenciación y quedan almacenados en la plataforma de soporte (*BaseSpace*) junto con el resto de información de la carrera. Algunos parámetros a los que conviene prestar atención son:

- Rendimiento (*yield*): representa la cantidad de información que se generará durante la secuenciación. Está relacionado tanto con el secuenciador como con el kit usado (Figura 6).

	CYLES	YIELD	PROJECTED YIELD	ALIGNER (%)	ERROR RATE (%)	IDENTITY CYCLE 1	Yield
Read 1	301	4.02 Mbp	4.02 Mbp	0.98	1.02	301	95.46
Read 2 (0)	0	140.94 Mbp	140.94 Mbp	0.99	0.99	1,277	95.61
Read 3 (0)	0	140.94 Mbp	140.94 Mbp	0.99	0.99	106	97.43
Read 4	301	4.02 Mbp	4.02 Mbp	0.96	1.04	277	94.76
Non-Index Reads Total	602	8.05 Mbp	8.05 Mbp	0.97	1.04	255	95.76
Totals	618	8.05 Mbp	8.05 Mbp	0.97	1.04	453	95.76

LANE	READ	INDEX	CONVERT	CLUSTERS PF	PASS/FILTER (%)	READS	READS PF	YIELD	YIELD	CYLES	ALIGNER	ERROR RATE	ERROR RATE	ERROR RATE	ERROR RATE	IDENTITY	COMMANDS	STATUS
			(K/MP)	(%)	(%)						(%)	(%)	(%)	(%)	(%)			
1	1	30	672 ±13	94.49 ±0.01	8.158 ±0.071	31,297,042	35,179,830	96.86	4.02 Mbp	200	0.98 ±0.01	1.02 ±0.01	0.54 ±0.01	0.21 ±0.01	0.28 ±0.01	201 ±07		OK Passed
2 (0)	30	672 ±13	94.49 ±0.01	8.669 ±0.066	31,297,042	35,179,830	88.41	140.94 Mbp	0	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	1,877 ±113			
3 (0)	30	672 ±13	94.49 ±0.01	8.669 ±0.066	31,297,042	35,179,830	97.43	140.94 Mbp	0	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	106 ±19			
4	30	672 ±13	94.49 ±0.01	8.127 ±0.080	31,297,042	35,179,830	94.75	4.02 Mbp	200	0.96 ±0.01	1.04 ±0.01	0.79 ±0.01	0.28 ±0.01	0.24 ±0.01	277 ±08			

Figura 6. Rendimiento (Yield). En el apartado Lane Metrics encontramos información sobre la cantidad de información generada. El rendimiento (*Yield*), número de *clusters* (*Density K/MM²*), el porcentaje de *clusters* que pasan los filtros (*Cluster PF%*), número de lecturas, y porcentaje del Q30.

- Gráfico *data by lane*: muestra, entre otras opciones, la densidad de *cluster* de las muestras. Es importante tener una densidad correcta, ya que se pueden perder muchas lecturas, debido al descarte hecho por el *software*, en caso de exceso de DNA (*saturación de clusters*) o se puede tener una secuenciación de muy baja cobertura por no haber introducido suficiente cantidad de DNA (*underclustering*). Se recomienda visitar las especificaciones de cada plataforma de Illumina para conocer la densidad de *cluster* idónea, puesto que este parámetro es dependiente de la plataforma y del kit usado (Figura 7).

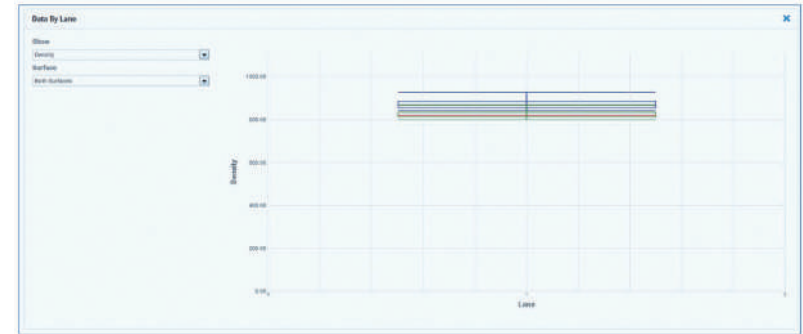


Figura 7. Muestra la densidad de los *clusters* para la carrera.

- *Q score distribution*: gráfico de barras que permite ver el número de bases en cada nivel de calidad (*Q score/Phred score*). El *Q score* indica la probabilidad de que una base se haya llamado erróneamente.

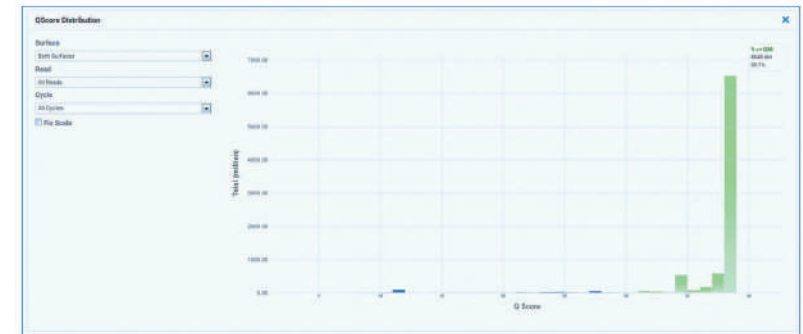


Figura 8. En la gráfica de *Q score distribution* visualizamos la distribución de las bases según su nivel de calidad (Q30).

- *Q score heatmap*: gráfico que muestra de forma global todos los *Q scores* (Figura 9).
- *Indexing QC*: histograma que permite comprobar el número de lecturas que se han generado por cada muestra multiplexada. Es útil para ver si hubo problemas a la hora de secuenciar alguna muestra (Figura 10).

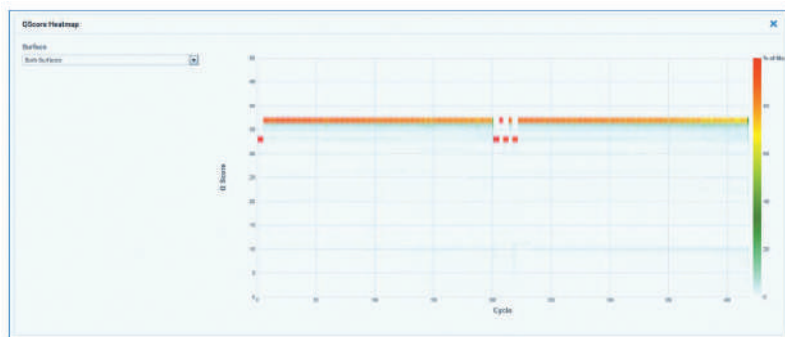


Figura 9. El *Q score heatmap* muestra la distribución de forma global del *Q score* para cada base.



Figura 10. El *Indexing QC* permite visualizar la distribución de las lecturas para cada una de las muestras cargadas en la carrera. Para una carrera con muestras de un solo tipo de panel la distribución debe ser homogénea; si se multiplexan muestras pertenecientes a distintos paneles la distribución puede ser heterogénea dependiendo de los tamaños de los paneles multiplexados.

PLATAFORMA ION TORRENT (THERMO FISHER)

Esta plataforma secuenciará el DNA usando un chip semiconductor. Durante el proceso de secuenciación, cuando un nucleótido se incorpora a la cadena de DNA, la reacción de polimerización del DNA libera un ión de hidrógeno, lo que provoca un cambio en el pH dentro del pocillo. Este cambio es captado por los sensores de pH del pocillo

(micropHmetros), que transforman ese cambio de pH en un voltaje. Este voltaje es detectado por el sistema, que registra la incorporación de la base correspondiente a la cadena resultante. Este proceso se repite secuencialmente durante un número de ciclos establecido.

En caso de no producirse una incorporación de nucleótido a la cadena de DNA, no habrá cambios en el pH y, por tanto, no se registrará ninguna base en esa posición. Si hay más de una base igual consecutiva, el cambio de pH será más pronunciado, con lo que el cambio de voltaje será mayor y el sensor convertirá esa señal en el número de bases proporcional a ese aumento. Este proceso se produce simultáneamente en todos los pocillos del chip.

COMPROBACIÓN DE LA CALIDAD DE LA CARRERA EN LA PLATAFORMA ION TORRENT

Los parámetros que se recomienda revisar, en cada muestra secuenciada son:

- La precisión (*accuracy*) de las lecturas: no debe ser inferior al 98% a lo largo de toda la longitud de las lecturas.
- La longitud de lectura: se debe confirmar que es la esperada. Si se obtienen secuencias cortas podría deberse a algún error durante la preparación de librerías o la secuenciación.

CALIDAD DE LAS MUESTRAS (PHRED SCORE)

Ambas plataformas reportan un control de calidad por cada base llamado *Phred score*. En la sección Escala de calidad *Phred* del capítulo Plataformas y criterios mínimos se explica con más detalle en qué consiste el *Phred score*.

Una lectura se considerará de buena calidad si tiene un *Phred score* mayor o igual a 20 si las muestras se han secuenciado usando la plataforma Ion Torrent, o si se ha secuenciado un genoma o un exoma. Sin embargo, para paneles secuenciados por Illumina, se recomienda que el *Phred score* alcance o supere el valor de 30.

ANÁLISIS DE LAS MUESTRAS

Una vez completado el proceso de secuenciación, se deberían tener dos ficheros fastq, generalmente comprimidos en gzip, con la secuencia de cada cadena de DNA (*forward* y *reverse*). Estos ficheros se compararán con un genoma de referencia para buscar variantes. Para ello, se pueden usar *software* comerciales que reporten directamente unas variantes candidatas, o crear un *pipeline* propio para realizar el análisis. Para poder realizar un *pipeline* propio es recomendable la participación de una persona con experiencia en el análisis bioinformático de muestras de NGS, ya que en algunos pasos se requieren conocimientos especializados para evitar falsos positivos y la pérdida de información relevante.

ANÁLISIS DE MUESTRAS SECUENCIADAS EN PLATAFORMAS ILLUMINA

El análisis de las muestras suele comprender los siguientes pasos, algunos opcionales:

1. Control de calidad previo
2. Pre-procesamiento (opcional)
3. Alineamiento contra el genoma de referencia
4. Post-alineamiento (opcional)
5. Detección (llamada) de variantes (*variant calling*)
6. Anotación de variantes

1.CONTROL DE CALIDAD PREVIO

El control de calidad previo se utiliza para obtener una visión global de la calidad de la secuenciación de la muestra. Este paso es independiente de los criterios de calidad que se muestran en el secuenciador, ya que puede haber *carreras* cuyos parámetros de calidad sean muy buenos, pero la muestra no se haya secuenciado con la calidad esperada.

Como se nombra en la sección anterior, cada base viene asociada con una calidad que se corresponde con la probabilidad de que se haya producido un error. Los programas

usados en este paso suelen mostrar distintos gráficos donde se muestra el número de lecturas que tiene cada muestra, la calidad general de las lecturas, la calidad de las bases y el número de bases no especificadas (contenido de 'N' de nuestras lecturas), entre otros. El programa recomendado para este paso es FastQC^(A), aunque hay otros programas disponibles con una eficiencia similar. Este programa acepta como entrada, ficheros fastq, en formato descomprimido o comprimido en gzip, o ficheros sam/bam (Tabla 17). El programa genera un conjunto de gráficos y test estadísticos que permitirán evaluar la calidad de la muestra.

2.PRE-PROCESAMIENTO (OPCIONAL)

En base a estos test, se determinará si se debe ejecutar el pre-procesamiento. Este paso es opcional y consiste en recortar las lecturas (a este proceso se le denomina *trimming*) en caso de que se vea que hay alguna zona en la que la calidad disminuye drásticamente (normalmente en el inicio o el final de la lectura), y/o descartar aquellas lecturas que no tengan una media de calidad mínima de sus bases. Para este proceso se recomienda usar el programa Trimmomatic⁽²⁾, aunque existen otros programas disponibles, como Trim Galore!^(B), QcReads⁽³⁾ o SeqtrimNEXT⁽⁴⁾, por ejemplo.

Una vez realizado el *trimming*, se recomienda volver a analizar con FastQC para comprobar que la calidad ha aumentado hasta parámetros deseables. En caso contrario, se recomienda repetir el paso del pre-procesamiento usando criterios más estrictos. Idealmente, se recomienda una calidad mínima (*Phred score*) de 30 pero en algunos casos no será posible ya que la cobertura de la muestra bajará drásticamente. En estos casos, se recomienda tener una calidad mínima entre 20 y 25.

3.ALINEAMIENTO CONTRA EL GENOMA DE REFERENCIA

Con el conjunto de lecturas con buena calidad, se procederá al alineamiento de las lecturas contra el genoma de referencia. Este paso consiste en buscar la posición dentro del genoma de referencia que más encaja con cada una de las lecturas guardadas en el archivo fastq. Se usan distintos algoritmos estadísticos para encontrar la posición más correcta, ya que, el genoma humano tiene muchas zonas repetitivas y,

en algunos casos, una lectura puede encajar perfectamente en distintas regiones del genoma. Por ello, para la validación de un panel es recomendable usar distintos alineadores para captar los posibles errores que pueda cometer el alineador.

Para realizar este paso, se recomienda usar el programa BWA⁽⁵⁾. Este programa permite un alineado con una sensibilidad superior al 90-95%⁽⁶⁾. En la Tabla 15 se muestran otros alineadores usados en la literatura.

Tabla 15. Alineadores más comunes (en negrita se muestra el recomendado por el GESMD).

PROGRAMA	ESTRATEGIA DE ALINEAMIENTO
BWA	Algoritmo de Burrows-Wheeler ⁽⁵⁾
Bowtie2	Algoritmo de Burrows-Wheeler ⁽⁷⁾
Stampy	Tablas hash ⁽⁸⁾
NovoAlign	Algoritmo de búsqueda iterativa ^(c)
SHRiMP2	Algoritmo de Smith-Waterman ⁽⁹⁾

Una vez se han alineado las lecturas se deben ordenar por coordenada genómica. Este paso se puede hacer tanto con el grupo de herramientas *samtools*⁽¹⁰⁾, usando la herramienta *sort*, como con el grupo de herramientas *Picard*^(b), usando la herramienta *SortSam*.

El alineamiento finaliza con la creación de un fichero índice para el fichero bam (suele tener extensión .bai). Este índice permite a los programas que se ejecutarán en los siguientes pasos acceder más rápidamente a la información contenida dentro de los ficheros sam/bam. Para crear el índice se pueden usar los mismos grupos de herramientas (*samtools*, herramienta *index*, o *Picard*, herramienta *BuildBamIndex*).

4. POST-ALINEAMIENTO (OPCIONAL)

Finalizado el alineamiento, la ordenación de las lecturas y la creación del índice del fichero bam, el siguiente paso es el post-alineamiento. Este paso se desglosa en marcado de

duplicados de PCR, realineamiento alrededor de las inserciones/delecciones (*indels*) y recalibrado de las bases.

El marcado de duplicados de PCR es un paso controvertido, hay grupos que lo desaconsejan para análisis de paneles de genes debido a la alta cobertura que suelen tener estas muestras. Al tener una cobertura tan alta, existe una gran probabilidad de que el programa considere falsos duplicados de lecturas. De esta manera, pueden perderse parte de las lecturas, con la consecuente pérdida de cobertura del análisis. Si además se ha usado un panel de amplicones para procesar la muestra, la pérdida de información será mucho mayor. Por este motivo, el GESMD considera totalmente contraindicado marcar duplicados en paneles de amplicones. En caso de paneles de captura, si se considera conveniente eliminar los duplicados, se recomienda usar la herramienta *MarkDuplicates*, dentro del grupo de herramientas de *Picard*.

El realineamiento de *indels* permite corregir errores creados por los alineadores. De esta forma, las lecturas serán más consistentes en las regiones donde pueda haber inserciones o delecciones. Para poder ejecutar este paso se necesita el grupo de herramientas GATK⁽¹¹⁾, en concreto se utilizarán *RealignerTargetCreator* e *IndelRealigner*.

El recalibrado de bases corrige errores de secuenciación y otros artefactos experimentales. Para ejecutar este paso se deben usar las herramientas *BaseRecalibrator* y *PrintReads*, dentro del grupo de herramientas de GATK.

Existen discrepancias respecto de la incorporación de estas dos herramientas en la *pipeline* de análisis⁽¹²⁾. Algunas pruebas realizadas por miembros del GESMD no muestran una mejoría en los resultados del análisis, ya que se ha llegado a los mismos resultados sin realizar el post-alineamiento. Es por esto que se considera un paso opcional dentro del análisis de muestras analizadas por NGS.

Para evitar la controversia que hay alrededor de la eliminación de duplicados se ha desarrollado un método alternativo en el proceso de fabricación de librerías basado en el marcado de cada molécula original de DNA con un índice único (UMI, *Unique Molecular Identifier*). La secuencia de estos UMIs sirve para identificar qué lecturas proceden de moléculas duplicadas por PCR (los que comparten la secuencia y el UMI), de las lecturas que se han originado de moléculas similares presentes en la muestra de partida (misma secuencia, pero distinto UMI). La explicación del análisis

de datos de secuenciación generados de librerías que incorporan UMIs no se aborda en la presente edición de estas guías, ya que esta opción no está aún extendida en el ámbito de la NGS con fines diagnósticos⁽¹³⁾. Una ventaja añadida del empleo de UMIs en la fabricación de librerías es que facilita la detección de variantes de muy baja frecuencia⁽¹⁴⁾.

5.LLAMADA DE VARIANTES (VARIANT CALLING)

El siguiente paso es la llamada de variantes o *variant calling*. En este paso, los programas (*variant callers*) comparan las lecturas alineadas con el genoma de referencia, buscando aquellas posiciones donde exista una discrepancia entre la lectura y el genoma de referencia (*mismatch*). En caso de encontrar suficientes lecturas con el mismo *mismatch* para descartar un error de secuenciación o de alineamiento, este *mismatch* se reporta como una variante. Normalmente, esta variante vendrá acompañada de una frecuencia alélica (VAF, *Variant Allele Frequency*), que se corresponde con el porcentaje en el que aparece la variante respecto de la cobertura total de esa base.

Existen distintos tipos de variantes, dependiendo del *mismatch*:

- SNV (*Single Nucleotide Variant*): son las más comunes, consisten en un cambio de un nucleótido por otro, lo que puede producir un cambio de aminoácido (*missense*), mantener el mismo aminoácido (*synonymous*), o ganancias o pérdidas del codón de terminación (*nonsense* o *stop*).
- *Indels*: inserciones o deleciones, que pueden cambiar o no la pauta de lectura (*frameshift* o *in-frame*, respectivamente), o pueden provocar la aparición un codón *stop*.
- *Splicing*: variantes SNV o indels localizadas en las regiones para el reconocimiento del inicio y final de intrones (regiones de *splicing*) y que pueden alterar el sitio de *splicing* de las regiones intrónicas, lo que con frecuencia tiene repercusiones en la secuencia del RNA mensajero maduro.

Existe una gran cantidad de *variant callers*, cada uno con sus ventajas y limitaciones.

En la Tabla 16 se muestran algunos de los más usados y el tipo de análisis en el que está recomendado su uso. El GESMD recomienda el uso combinado de la herramienta *mpileup*, dentro del grupo de herramientas de *samtools*, y VarScan2⁽¹⁵⁾ por su flexibilidad en el análisis.

Tabla 16. Variant callers más populares (en negrita se muestra el recomendado por el GESMD).

NOMBRE	USO RECOMENDADO PARA	REFERENCIA
GATK HaplotypeCaller	Variantes germinales	McKenna <i>et al.</i> , 2010 ⁽¹¹⁾
Samtools + VarScan 2	Variantes somáticas y germinales	Koboldt <i>et al.</i>, 2012⁽¹⁵⁾
Mutect	Variantes somáticas	Cibulskis <i>et al.</i> , 2013 ⁽¹⁶⁾
Strelka	Variantes somáticas	Saunders <i>et al.</i> , 2012 ⁽¹⁷⁾
Seurat	Variantes somáticas	Christoforides <i>et al.</i> , 2013 ⁽¹⁸⁾

El fichero de salida de estos *variant callers* suele ser un fichero vcf (Tabla 17). Este fichero contiene todas las discrepancias entre las lecturas de la muestra y el genoma de referencia. Sin embargo, carece de la información de qué genes se encuentran en la región donde se ha encontrado la variante.

Tabla 17. Tipos de fichero de salida.

TIPO DE FICHERO	DESCRIPCIÓN
bcl	Formato binario que contiene los nucleótidos y su calidad en cada mosaico (<i>tile</i>) y ciclo. El fichero puede ser analizado durante la secuenciación.
FASTA	Formato texto, contiene una única línea de cabecera que comienza con el símbolo ">" seguido del identificador de secuencia. Es un formato válido tanto para secuencias de DNA/RNA como para secuencias de proteínas. El resto de líneas que siguen a este encabezado son parte de la secuencia de DNA, RNA o proteína y suele contener 80 caracteres de largo por cada línea. Estos caracteres se representan siguiendo las normas de la IUB/IUPAC.
FASTQ	Fichero en formato texto que contiene tanto la secuencia de DNA como información sobre la calidad de cada base. Cada secuencia está compuesta por 4 líneas: <ul style="list-style-type: none"> o Encabezado único que permite distinguir todas las lecturas generadas en cada muestra; comienza con "@". Si el demultiplexado es realizado por CASAVA (Illumina), contiene, además, la siguiente información: Nombre, número de secuenciación, ID de la <i>flow cell</i>, línea, mosaico, coordenadas X e Y en la <i>flow cell</i>, filtro de calidad de la lectura y secuencia del index. o Una línea con los nucleótidos secuenciados (combinación de las letras G, A, T y C). o Una línea, generalmente, sólo con el símbolo "+". o Una línea en formato ASCII (33+score) igual a la longitud y número de bases de la secuencia. Codifica la calidad en formato Phred score para cada base secuenciada.
sam/bam/bai	<ul style="list-style-type: none"> o SAM (<i>Sequence Alignment/Map</i>): fichero en formato texto, contiene información sobre cada lectura respecto a la posición del genoma de referencia, las variantes en la lectura, la calidad del alineamiento y la calidad de la lectura en formato ASCII, representando el Phred score. o BAM: secuencia binaria y comprimida del fichero SAM. o BAI: índice del BAM.
ubam	bam no alineado. El formato de este fichero es similar al sam/bam, pero sin la información de alineamiento. Se puede considerar un fichero fastq ampliado, ya que incluye información típica de la lectura, junto con información de la carrera que han generado las lecturas. Esta información adicional viene incluida dentro de la cabecera del fichero.
vcf/bcf	vcf (<i>Variant Calling Format</i>): formato texto, contiene información sobre variantes. Suele estar compuesto por una cabecera extensa, de longitud variable, cuyas líneas empiezan por "##", donde se explica el tipo de fichero y la información guardada dentro de la columna "información adicional". Después de la cabecera, viene la información de las variantes, separadas en columnas, siguiendo el siguiente formato. Cada fila se compone de 8 columnas: <ol style="list-style-type: none"> 1. Cromosoma 2. Posición en el genoma 3. Identificador 4. Base/s de referencia 5. Base/s que difiere del genoma 6. Calidad de la base alterada 7. Filtro 8. Información adicional. Dependiendo del programa que se use para hacer el <i>variant calling</i>, el contenido de esta columna variará bcf: formato binario (comprimido) del vcf
BED	Formato de texto que contiene una lista de posiciones en un genoma. Cada fila se compone de: identificador del cromosoma, posición inicial y final. Se puede añadir información adicional en las columnas siguientes, como un identificador.
GFF (General Feature Format)	Formato de texto, contiene una línea por cada característica (<i>feature</i>) (gen, exón u otras regiones genómicas). Contiene 9 columnas: <ol style="list-style-type: none"> 1. Cromosoma 2. Fuente de anotación 3. Tipo de característica 4. Posición inicial 5. Posición final 6. Puntuación 7. Cadena, + o - 8. Marco, (0, 1 ó 2) indicando si la primera base de esta característica es la segunda o tercera base en la proteína. 9. Atributo: separado por ";", contiene la información adicional sobre la característica.

6. ANOTACIÓN DE VARIANTES

Para poder conocer toda la información referente a la zona genómica donde se ha detectado una variante, se debe usar un programa de anotación de variantes, o si se prefiere, se puede buscar la información de la posición de cada variante directamente en un *genome browser*, como UCSC^(E), RefSeq Genome Data Viewer^(F), o ENSEMBL^(G). Generalmente se obtendrá una elevada cantidad de variantes en cada muestra analizada, aunque el número suele variar dependiendo de la cantidad de regiones que se analicen dentro del panel. Por tanto, se recomienda usar programas de anotación de variantes. El más utilizado y recomendado por el GESMD es ANNOVAR⁽¹⁹⁾. ANNOVAR permite combinar toda esta información y guardarla en un único fichero de texto separado en columnas que se puede abrir fácilmente usando Microsoft® Excel® o *softwares* similares. La anotación es importante para el posterior filtrado de variantes, paso que se explicará más detalladamente al final de este capítulo.

ANÁLISIS DE MUESTRAS SECUENCIADAS EN PLATAFORMAS ION TORRENT

1.CONTROL DE CALIDAD

En el caso de las muestras secuenciadas en las plataformas Ion Torrent, el programa de gestión (Torrent Suite™) genera automáticamente un informe donde se detallan las estadísticas globales de la carrera, junto con algunos datos específicos de cada muestra, como son el porcentaje de bases con Q score mayor que 20, el número de lecturas totales, la cobertura media de la secuenciación, la uniformidad de la cobertura en las regiones de interés, el *strand-bias*,... Dependiendo de la tecnología de captura elegida y de la cobertura deseada, se establecerán umbrales específicos para considerar la muestra como apta para el análisis.

2. PRE-PROCESAMIENTO, ALINEAMIENTO Y VARIANT CALLING

El Torrent Suite™ extraerá las lecturas de cada muestra por separado, recortando los adaptadores de cada una de ellas y excluyendo aquellas que tengan una baja calidad. Además, este programa alinea las lecturas contra el genoma de referencia usando el alineador TMAP.

La salida de TMAP es un fichero bam, el cual se puede usar como bam de entrada para ejecutar Torrent Variant Caller y hacer el *variant calling*, obteniendo así un fichero vcf. Este fichero vcf podrá ser anotado de forma análoga a como se anotan los vcf resultantes del análisis de las plataformas Illumina.

Todas las herramientas mencionadas en este apartado son gratuitas y se pueden descargar desde su repositorio en Github. A pesar de ello, el GESMD recomienda usar el *software* comercial de Ion Torrent, ya que garantiza un uso de los parámetros de configuración del análisis optimizado.

ANÁLISIS MEDIANTE EL USO DE SOFTWARE COMERCIAL

En el GESMD somos conscientes de que no todos los equipos de trabajo cuentan con personal con la suficiente formación en biología computacional como para procesar los datos de secuenciación de modo manual. Afortunadamente, diversas casas comerciales cuentan con *softwares* para facilitar el análisis de los datos de secuenciación. Estos programas varían en cuanto al grado de apertura del *software* para ser adaptado a las diferentes necesidades, y en cuanto a los tipos de análisis y anotaciones que permiten realizar. La Tabla 18 resume las ventajas e inconvenientes que el Grupo GESMD ha detectado en cada uno de los *softwares* comerciales que se utilizan en el Grupo. Una ventaja común a todos ellos es que permiten trazar el análisis realizado, mediante un sistema de versiones que quedan registradas en el propio análisis. Esto es de gran utilidad, particularmente en el contexto de trabajo en grupo, ya que aunque cada centro realice su propio análisis, siempre es posible comparar los resultados entre los distintos centros, trazando exactamente qué parámetros se han empleado en cada análisis.

En la Tabla 18 se muestran características de los programas comerciales usados por el GESMD: 1) proveedor del *software*, 2) los ficheros de entrada que necesita para poder ejecutarse, 3) los ficheros de salida (resultados) que genera y 4) la *pipeline* interna (si esta información está disponible, en caso contrario se muestra N/A). Además, se incluyen ventajas e inconvenientes de cada *software* y, si se dispone, cómo se detectan los errores de secuenciación.

Tabla 18. Características de los software comerciales usados por el GESMD.

SOFTWARE, PROVEEDOR	FICHEROS ENTRADA	FICHEROS SALIDA	ALINEAMIENTO	DETECCIÓN VARIANTES	ANOTACIÓN VARIANTES	CONTROL ERRORES SECUENCIACIÓN
MiSeq Reporter, Illumina	FASTQ	BAM Bam-bai VCF	Smith-Waterman algorithm	MiSeq Reporter	N/A	N/A
Variant Studio, Illumina	VCF	TXT	NA	NA	VariantStudio Custom	Permite analizar y comparar un conjunto de muestras. Permite anotar variantes conocidas como errores de secuenciación.
DNAnexus, DNAnexus	FASTQ	BAM Bam.bai VCF VCF.gz	BWA	DNAnexus	DNAnexus	N/A
Sophia DDM® Sophia Genetics	FASTQ	BAM Bam-bai VCF TXT	PEPPER™, MUSKAT™ y MOKA™ (software propio de Sophia DDM®)			Señal de alarma en las zonas de baja cobertura. Señala las muestras analizadas en las que se encuentra la misma variante.
Ion Reporter®, ThermoFisher	BAM	VCF, TSV	TMAP	TVC	IonReporter	Permite anotar variantes conocidas como errores de secuenciación.
SOFTWARE	VENTAJAS		INCONVENIENTES			
MiSeq Reporter, Illumina	Genera el VCF en formato para VariantStudio.		Sensibilidad del 1%.			
Variant Studio, Illumina	Permite trabajar con anotaciones <i>custom</i> Filtrado muy ágil.		Permite analizar VCF de otros software pero para ello se debe adecuar el formato del vcf al del VariantStudio. Variantes generadas por MiSeqReporter con sensibilidad del 1%.			
DNAnexus, DNAnexus	Permite analizar con diferente sensibilidad.		No anota datos poblacionales de ExAC o EVS. No permite generar un fichero que incluya varias muestras. Filtrado manual en Excel.			
Sophia DDM® Sophia Genetics	Especificidad y sensibilidad del 99.9%. Fácil manejo. Interpretación y categorización de variantes. Clasificación somática basada en asociación variante-fármaco-enfermedad. Acceso a la "Clinical Genomics Community". Actualización constante del software.		Sistema cerrado. Imposibilidad de hacer variaciones en el pipeline.			
Ion Reporter®, ThermoFisher	Sistema de análisis sencillo integrado en los secuenciadores Ion Torrent. Incorpora diferentes workflow de análisis para diferentes tipos de muestras. Permite comparar muestras entre sí.		Sistema cerrado sólo para Ion Torrent. No permite actualización ni modificación.			

FILTRADO DE VARIANTES

Como se ha mencionado anteriormente, la lista de variantes que proporciona el *pipeline* de análisis, comercial o propio, será bastante extensa debido a la detección de polimorfismos (*Single Nucleotide Polymorphism*, SNP), errores de secuenciación, variantes que no están situadas dentro de la zona codificante del gen, etc. Por ello, antes de comenzar a interpretar las variantes candidatas, se deben filtrar aquellas variantes que no se vayan a informar. Cabe destacar que los criterios de filtrado de variantes están en evaluación continua, ya que, dependiendo de los avances que se realicen en el campo de la secuenciación masiva, puede ser conveniente modificar la forma de evaluar las variantes.

PRE-FILTRADO DE VARIANTES

En este proceso se eliminarán aquellas variantes fuera de las regiones codificantes, así como los errores de secuenciación. Esta sección se puede dividir en dos pasos: el primero es prácticamente automático y consiste en eliminar aquellas variantes fuera de las regiones codificantes; el segundo es más manual, y consiste en visualizar las lecturas para eliminar posibles errores de secuenciación que no se han detectado durante el análisis previo.

PRE-FILTRADO AUTOMÁTICO DE VARIANTES

En este paso se deben preservar las variantes que estén en zonas codificantes (exónicas, 5'UTR y 3'UTR) y de *splicing* (mínimo de -5 a +5). Se eliminarán las variantes que estén en zonas intergénicas, *downstream*, *upstream*, en RNA no codificantes y en zonas intrónicas alejadas de las regiones de *splicing*. Para agilizar el pre-filtrado, se pueden usar programas de hojas de cálculo, como Microsoft® Excel® o similares.

PRE-FILTRADO MANUAL DE VARIANTES

El pre-filtrado manual de variantes permite descartar errores de secuenciación que no hayan sido detectados por el *software* de análisis usado previamente.

Para poder realizar este filtrado se debe tener instalado en el ordenador un visor genómico que pueda abrir ficheros de alineamiento (bam o sam). El más popular es *Integrative Genomics Viewer* (IGV)^(20,H), que se encuentra disponible para descargar de forma gratuita. Para optimizar el tiempo de análisis, se recomienda visualizar en el IGV solo las variantes candidatas, una vez ya se hayan filtrado las variantes que no son de interés, como se comenta más abajo.

Para cada variante que se esté visualizando, se recomienda descartar los siguientes errores de secuenciación:

- Cobertura no uniforme: la región donde se encuentra la variante debe tener una captura uniforme en un rango de +/- 15 bases alrededor de la variante estudiada.
- Detección en un único fragmento: la variante debe estar secuenciada por más de una lectura de coordenadas no idénticas.
- *Strand bias*: se debe comprobar que la proporción de lecturas en las cadenas *forward* y *reverse* se mantiene equilibrada entre las lecturas mutadas y las de la referencia, descartando así el *strand bias*. Si no se pudiera descartar, se recomienda hacer, si es posible, una validación por secuenciación con el método de Sanger de esta región.
- Zonas repetitivas: se deben descartar las variantes en zonas repetitivas y en homopolímeros.
- *Edge effect*: se deben descartar las variantes que se encuentren en los extremos de las lecturas.

FILTRADO DE POLIMORFISMOS

La MAF (*minor allele frequency*) hace referencia a la frecuencia en que una variante está descrita en la población. Las variantes con una MAF total (o en una subpoblación étnica) mayor del 1% son consideradas actualmente polimorfismos (SNPs). Estas variantes no serán categorizadas debido a que, en base al conocimiento actual, no presentan una utilidad clínica, pudiendo dificultar la interpretación. Se recomienda describir en el informe la actitud ante estas variantes (por ejemplo: “Las variantes descritas en las bases de datos poblacionales clasificadas como polimorfismos o SNPs -*Single Nucleotide Polymorphisms*- no se incluyen en este informe”) o bien

facilitar esta información en una tabla anexa. A continuación, se listan algunas bases de datos de poblaciones que resultan útiles para excluir polimorfismos:

BASE DE DATOS	PÁGINA WEB
1000 Genomes Project	http://browser.1000genomes.org
Exome Variant Server	http://evs.gs.washington.edu/EVS
dbSNP	http://www.ncbi.nlm.nih.gov/snp
dbVar	http://www.ncbi.nlm.nih.gov/dbvar
ExAC	http://exac.broadinstitute.org

FILTRADO DE VARIANTES SINÓNIMAS

Se recomienda filtrar las variantes sinónimas, que son aquellas que no generan un cambio en la proteína codificante, ya que, en base al conocimiento actual, aunque podrían tener utilidad clínica, siguen presentando dificultad en la interpretación.

FILTRADO DE VARIANTES EN REGIONES UTR

Se recomienda filtrar las variantes encontradas en las regiones 3'UTR y 5'UTR, ya que en base al conocimiento actual, aunque podrían tener utilidad clínica, siguen presentando dificultad en la interpretación.

REFERENCIAS

1. Ledergerber C, Dessimoz C. Base-calling for next generation sequencing platforms. *Brief Bioinform.* 2011;12(5):489-497.
2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120.
3. Ma Y, Xie H, Han X, et al. QcReads: an adapter and quality trimming tool for next-generation sequencing reads. *J Genet Genomics.* 2013;40(12):639-642.
4. Falgueras J, Lara AJ, Fernandez-Pozo N, et al. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics.* 2010;11:38.
5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589-595.
6. Hatem A, Bozdog D, Toland AE, et al. Benchmarking short sequence mapping tools. *BMC Bioinformatics.* 2013;14:184.
7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359.
8. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 2011;21(6):936-939.
9. David M, Dzamba M, Lister D, et al. SHRIMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics.* 2011;27(7):1011-1012.
10. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993.
11. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.
12. Tian S, Yan H, Kalmbach M, et al. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics.* 2016;17(1):403.
13. Kou R, Lam H, Duan H, et al. Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLoS One.* 2016;11(1):e0146638.

14. Kinde I, Wu J, Papadopoulos N, et al. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A.* 2011;108(23):9530-9535.
15. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568-576.
16. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;31(3):213-219.
17. Saunders CT, Wong WSW, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics.* 2012;28(14):1811-1817.
18. Christoforides A, Carpten JD, Weiss GJ, et al. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics.* 2013;14:302.
19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.
20. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26.

REFERENCIAS WEB

- A. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- B. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore
- C. <http://www.novocraft.com/products/novoalign/>
- D. <http://broadinstitute.github.io/picard>
- E. <http://genome.ucsc.edu/>
- F. <https://www.ncbi.nlm.nih.gov/genome/gdv/>
- G. <http://www.ensembl.org/index.html>
- H. <http://software.broadinstitute.org/software/igv/download>

INTERPRETACIÓN DE VARIANTES

GUÍA DE APLICACIÓN CLÍNICA DE LA SECUENCIACIÓN MASIVA EN SÍNDROMES MIELODISPLÁSICOS Y LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:

Sara Álvarez, Iria Vázquez, Bárbara Tazón-Vega, Marta Cabezón, María José Larráyo, Laura Palomo, Mariam Ibáñez, Lurdes Zamora, Marta Fernandez-Mercado, José Cervera, Francesc Solé y Esperanza Such.

INTERPRETACIÓN DE VARIANTES

La categorización y clasificación de las variantes resultantes de un estudio de secuenciación masiva tiene como objetivo principal facilitar la interpretación de los resultados en el contexto clínico de los SMD y armonizar los datos entre laboratorios.

La categorización de una variante en el momento del diagnóstico estará basada en los registros de las bases de datos consultadas y en la literatura disponible en el momento de la realización del estudio. En revisiones posteriores, la identificación recurrente de una variante, variaciones en la frecuencia poblacional o nuevos datos sobre la funcionalidad de un gen serán determinantes para su reclasificación.

Se interpretarán y categorizarán las variantes que superen los criterios de calidad (reflejados en el apartado de filtrado de variantes) y que cumplan las siguientes condiciones:

- 1) MAF < 1% (MAF, de las siglas en inglés, frecuencia del alelo menor)
- 2) VAF > 10% (VAF, de las siglas en inglés, frecuencia del alelo variante)
- 3) Profundidad mínima de 25 lecturas del alelo alternativo

La categorización se realizará en base a 3 características, la presencia en bases de datos clínicas, la accionabilidad del gen alterado y la predicción de patogenicidad mediante algoritmos *in silico*.

BASES DE DATOS CLÍNICAS

BASES DE DATOS DE VARIANTES

Para la interpretación de variantes es necesario conocer si las variantes identificadas están previamente recogidas en las bases de datos y el tejido y/o tumor en que han sido reportadas. Las direcciones web disponibles para su consulta se detallan en la Tabla 19.

Tabla 19. Bases de datos de variantes somáticas.

BASE DE DATOS	PÁGINA WEB
ClinVar*	http://www.ncbi.nlm.nih.gov/clinvar
Catalog of Somatic Mutations in Cancer (COSMIC)*	http://cancer.sanger.ac.uk/cosmic
IARC (WHO) TP53 mutation database	http://p53.iarc.fr
My Cancer Genome	http://www.mycancergenome.org
Personalized cancer therapy, MDACC	https://pct.mdanderson.org
Intogen	https://www.intogen.org/search
Pediatric Cancer Genome Project	http://explorepcgp.org
International Cancer Genome Consortium (ICGC)	https://dcc.icgc.org
The Cancer Genome Atlas (TCGA)	https://portal.gdc.cancer.gov
cBioPortal for Cancer Genomics	http://www.cbioportal.org
Human Gene Mutation Database (HGMD)	http://www.hgmd.org
Leiden Open Variation Database (LOVD)	http://www.lovd.nl

*Bases de datos frecuentemente utilizadas por los miembros del GESMD.

La interpretación de los cambios genéticos identificados en la muestra a estudio se realizará utilizando las bases de datos de variantes somáticas conocidas que permitirá determinar su recurrencia en la patología a estudio, teniendo en cuenta también el tejido/tumor en el que han sido identificadas.

En base a los recursos actualmente disponibles, se recomiendan como bases de datos de obligada consulta ClinVar, COSMIC e IARC TP53. En ausencia de información relevante, la consulta de otras bases de datos (Tabla 19) y de bibliografía actualizada es esencial para la adecuada categorización de las variantes. Con este fin, el GESMD ha creado una base de datos de variantes frecuentes en SMD y LMMC (GESMDmol) que podrá ser una herramienta adicional para la interpretación de los resultados.

Para la interpretación de las variantes en genes para los que no sea evidente su accionabilidad se recomienda la consulta de bases de datos de ensayos clínicos (Tabla 20).

BASES DE DATOS DE ACCIONABILIDAD

Para determinar la accionabilidad del gen alterado y su repercusión terapéutica pueden consultarse si existen ensayos clínicos específicos para ese gen. Las principales bases de datos de ensayos clínicos y sus direcciones web se detallan en la Tabla 20.

Tabla 20. Bases de datos de ensayos clínicos.

BASE DE DATOS	PÁGINA WEB
ClinicalTrials.gov	https://clinicaltrials.gov/
EU Clinical Trial Register	https://www.clinicaltrialsregister.eu/ctr-search/search

ALGORITMOS DE PREDICCIÓN

Los algoritmos de predicción estiman la probabilidad de que una variante, en un determinado gen, tenga consecuencias deletéreas para la proteína codificada. Los criterios que determinan el efecto del cambio genético identificado, son dependientes del algoritmo utilizado. Estos algoritmos son necesarios para evaluar variantes de cambio de sentido y variantes de *splicing*, previamente no descritas en las bases de datos o en la literatura (Tabla 21).

Se han desarrollado múltiples algoritmos. Actualmente, los más frecuentemente utilizados por los miembros del GESMD para las variantes de cambio de sentido, son PolyPhen2, SIFT, MutationTaster y PROVEAN.

Tabla 21. Algoritmos de Predicción.

UTILIDAD/FUNCIÓN	ALGORITMO	DIRECCIÓN WEB
Variantes <i>missense</i>	PolyPhen2*	http://genetics.bwh.harvard.edu/pph2
	SIFT*	http://sift.jcvi.org
	MutationAssessor	http://mutationassessor.org
	MutationTaster*	http://www.mutationtaster.org
	PROVEAN*	http://provean.jcvi.org
	Condel	http://bg.upf.edu/fannsd/
	CoVEC	https://sourceforge.net/projects/covec/files
	CADD	http://cadd.gs.washington.edu
Regiones de <i>splicing</i>	GERP++	http://mendel.stanford.edu/sidowlab/downloads/gerp
	Finder42	http://www.umd.be/HSF3/
	MaxEntScan43	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
	NetGene244	http://www.cbs.dtu.dk/services/NetGene2
	NNSplice50	http://www.fruitfly.org/seq_tools/splice.html
GeneSplicer51	http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml	

*Los algoritmos más frecuentemente utilizados por los miembros del GESMD.

CRITERIOS PARA LA CATEGORIZACIÓN DE VARIANTES

Hasta el momento los sistemas para variantes somáticas no están claramente establecidos. Se propone establecer una clasificación de variantes basada en su accionabilidad. Esto incluye variantes identificadas como dianas terapéuticas, con valor pronóstico/predictivo y/o con utilidad en el diagnóstico o en la clasificación de la entidad clínica.

Las variantes serán clasificadas teniendo en cuenta los siguientes criterios:

I.- Presencia recurrente en bases de datos o en la literatura, que indiquen la patogenicidad o la recurrencia de la variante identificada.

Se considera una variante como recurrentemente asociada a la patología a estudio cuando ha sido descrita al menos por 2 autores distintos.

II.- Tejido y/o histología del tumor en la que se ha descrito la variante.

III.- Relación causal o accionabilidad del gen afectado por la variante identificada en la patología a estudio.

Se define como un **gen accionable** aquel que presenta un **alto nivel de evidencia** en al menos una de las siguientes categorías en la entidad clínica a estudio:

- 1) gen considerado como diana terapéutica,
- 2) gen con valor predictivo/pronóstico,
- 3) gen asociado al diagnóstico/clasificación.

La accionabilidad de un gen se determinará en función de las siguientes evidencias:

- Alto nivel de evidencia si:
 - Biomarcador aprobado por la FDA (*Food and Drug Administration*), ESMO (*European Society for Medical Oncology*) o EMA (*European Medicine Agency*).
 - Biomarcador incluido en guías de tratamiento/diagnóstico.
 - Evidencias de accionabilidad en meta-análisis.
 - Resultados positivos en ensayos clínicos en fase II y fase III de carácter multicéntrico o con análisis retrospectivo y prospectivo del biomarcador.

- Bajo nivel de evidencia si:

- Modelos *in vivo*.
- Modelos *in vitro*.
- Predictores *in silico*.
- *Abstracts* o resultados preliminares.

El criterio III es especialmente relevante, cuando se utilizan paneles panhematológicos u oncológicos.

IV.- Algoritmos de predicción y estudios funcionales.

Las variantes no descritas (variantes 3 y 4) serán subcategorizadas en posiblemente patogénicas, de significado incierto o posiblemente benigno, en función de la disponibilidad de estudios funcionales y de los resultados de los algoritmos de predicción funcional.

La combinación de los resultados entre los diferentes predictores será utilizado para establecer la categorización de aquellas variantes previamente no descritas en las bases de datos o en la literatura consultada. Se recomienda la utilización de al menos cuatro predictores. El peso de cada uno de los sistemas de predicción y la combinación de los resultados obtenidos deberá ser establecido por cada laboratorio.

Para el análisis de las regiones de *splicing* queda a criterio del laboratorio la selección de los algoritmos a utilizar. Se recomienda incluir los 5pb flanqueantes de cada exón en todos los análisis. Los algoritmos más utilizados se listan en la Tabla 21.

Las variantes que dan lugar a un cambio en el marco de lectura o generan un codón de parada, en aquellos genes cuya patogenicidad se asocie a una pérdida de función, se interpretarán como variantes posiblemente patogénicas, siempre que no afecten al último exón. Las variantes localizadas en el último exón que impliquen una pérdida de función serán evaluadas en función de otros elementos como los dominios proteicos alterados o el efecto de otras variantes en posiciones cercanas, pudiéndose categorizar como variantes probablemente patogénicas o de significado incierto.

Aquellos cambios que no alteren el marco de lectura o que conlleven una pérdida de función y que afecten a genes que, por el contrario, su patogenicidad se asocie a una ganancia de función (por ejemplo *JAK2*), se interpretarán como variantes de significado incierto.

CLASIFICACIÓN DE LAS VARIANTES

Siguiendo los criterios de clasificación, se establecerán 5 clases funcionales (Tabla 22).

VARIANTE CLASE 1

Descritas en la literatura como relevantes en el MANEJO CLÍNICO de los SMDs o en otras hemopatías mieloides. Esta variante ha sido establecida como una variante patogénica en hemopatías mieloides, y altera un gen accionable.

VARIANTE CLASE 2

Descritas en la literatura pero sin valor establecido en el manejo clínico de los SMD o en otras hemopatías mieloides.

Esta variante ha sido establecida como una variante patogénica en tumores sólidos o hemopatías NO mieloides y altera un gen accionable.

VARIANTE CLASE 3

Variante no descrita previamente, en un gen establecido como accionable en SMDs o en otras hemopatías mieloides.

En función de los algoritmos de predicción (Tabla 21) y la disponibilidad de estudios funcionales serán clasificadas como:

- 3A: posiblemente patogénica - La mayoría de los algoritmos de predicción utilizados o los estudios funcionales disponibles indican que la variante altera la función de la proteína.
- 3B: de significado incierto - Los algoritmos de predicción utilizados o los estudios funcionales disponibles no pueden establecer o presentan resultados contradictorios con respecto al efecto sobre la proteína.
- 3C: posiblemente benigna - La mayoría de los algoritmos de predicción utilizados o los estudios funcionales disponibles indican que la variante no tiene un efecto que modifique sustancialmente la función de la proteína.

VARIANTE CLASE 4

Variante no descrita previamente, en un gen que no ha sido establecido como accionable en hemopatías mieloides, pero sí en otros tumores sólidos o hematológicos no mieloides.

En función de los algoritmos de predicción (Tabla 21) y la disponibilidad de estudios funcionales serán clasificadas como:

- 4A: posiblemente patogénica - La mayoría de los algoritmos de predicción utilizados o los estudios funcionales disponibles indican que la variante altera la función de la proteína.
- 4B: de significado incierto - Los algoritmos de predicción utilizados o los estudios funcionales disponibles no pueden establecer o presentan resultados contradictorios con respecto al efecto sobre la proteína.
- 4C: posiblemente benigna - La mayoría de los algoritmos de predicción utilizados o los estudios funcionales disponibles indican que la variante no tiene un efecto que modifique sustancialmente la función de la proteína.

VARIANTE CLASE 5

Variantes en este gen NO han sido establecidas como accionables en ningún tumor sólido o hematológico.

RECLASIFICACIÓN DE VARIANTES:

La categorización de las variantes se modificará en base a las evidencias actualizadas en las bases de datos y en la literatura.

Es responsabilidad del laboratorio determinar la periodicidad con la que se considera necesaria la re-categorización de variantes, debiéndose realizar al menos una vez al año.

Tabla 22. Clasificación de las variantes.

CATEGORÍA	CRITERIO I VARIANTE RECURRENTE EN BASE DE DATOS ⁽¹⁾	CRITERIO II TEJIDO DONDE SE HA DESCRITO LA VARIANTE ⁽²⁾	CRITERIO III ACCIONABILIDAD DEL GEN ALTERADO	CRITERIO IV PREDICTORES Y ESTUDIOS FUNCIONALES
CLASE 1	Variante descrita como PATOGENICA	En la enfermedad o tejido a estudio	Gen Accionable en la enfermedad a estudio	NP
CLASE 2	Variante descrita como PATOGENICA	En OTRA enfermedad o tejido a estudio	Gen Accionable en OTROS tumores sólidos o hematológicos	NP
CLASE 3	Variante NO descrita	NP	Gen Accionable en la enfermedad a estudio	3A. Posiblemente patogénica 3B. De significado incierto 3C. Posiblemente benigna
CLASE 4	Variante NO descrita	NP	Gen accionable en OTROS tumores sólidos o hematológicos	4A. Posiblemente patogénica 4B. De significado incierto 4C. Posiblemente benigna
CLASE 5	Variante NO descrita	NP	Accionabilidad desconocida	

NP: No procede. ⁽¹⁾ Variante identificada asociada a la patología a estudio por al menos 2 autores.

⁽²⁾ Implicación clínica de la identificación de la variante. Se considera un gen accionable aquel que tiene en una determinada entidad clínica un valor como diana terapéutica, como elemento predictivo/ pronóstico y/o presenta implicaciones en el diagnóstico / clasificación.

REFERENCIAS

1. Li MM, Datto M, Duncavage EJ, *et al.* Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn.* 2017;19(1):4-23.
2. Richards S, Aziz N, Bale S, *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):405-424.
3. Sukhai MA, Craddock KJ, Thomas M, *et al.* A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genet Med.* 2016;18(2):128-136.
4. Van Allen EM, Wagle N, Stojanov P, *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med.* 2014;20(6):682-688.



ELABORACIÓN DE INFORMES

GUÍA DE APLICACIÓN CLÍNICA DE
LA SECUENCIACIÓN MASIVA EN
SÍNDROMES MIELODISPLÁSICOS Y
LEUCEMIA MIELOMONOCÍTICA CRÓNICA

AUTORES:

María José Calasanz, Sara Álvarez, Iria Vázquez, Marta Fernandez-Mercado, María Diez Campelo, David Valcárcel, Lurdes Zamora, Pamela Acha, Guillermo Sanz, Francesc Solé y Esperanza Such.

ELABORACIÓN DE INFORMES

En el grupo se propone la realización de un informe consensado en cuanto a la información que debe contener.

INFORMACIÓN RECOGIDA EN EL INFORME DE NGS

La información que debe recoger el informe está dividida en los siguientes puntos:

IDENTIFICACIÓN COMPLETA DEL PACIENTE

Datos demográficos, diagnóstico confirmado, momento, fecha, hospital y médico solicitante. Declaración de haber firmado el consentimiento informado.

METODOLOGÍA

Material empleado y de qué tipo de muestra proviene, plataforma y genes analizados (haciendo referencia a si son regiones *hotspots* o completas) y si excluye algún gen relevante (por ejemplo *FLT3*-ITD) y la referencia del genoma utilizado.

RESULTADOS

Comenzar con una frase donde se resuman los principales hallazgos encontrados de manera comprensible para el clínico, categorizando la variante en función del significado de la mutación (categorización de variantes propuesta por el GESMD).

INFORMACIÓN COMPLEMENTARIA

A continuación incluir la siguiente información:

- Detallar que se exponen únicamente los genes que presentan variantes que pueden ser relevantes para el diagnóstico, pronóstico o tratamiento.

- Detallar que se informan las variantes con un VAF >10% y la profundidad mínima de lectura (100x).
- Detallar los resultados en formato tabla donde se incluya la siguiente información (Tabla 23).
- Además, las variantes con un VAF <10% (siempre que haya un número de 25 lecturas para el alelo alternativo) se describirán en una tabla anexa al final del informe para valorar el posible seguimiento. Si la variante tuviera relevancia clínica, podría incluirse en la tabla de resultados resaltando que la frecuencia alélica es inferior al 5%.
- Los SNP no relevantes clínicamente no se informarán pero podrán ser solicitados por el médico responsable y/o paciente.
- Se debe adjuntar un comentario donde se incluyan las bases de datos empleadas para la interpretación.
- Incluir algún detalle de las regiones no cubiertas de interés diagnóstico y cómo se han resuelto (informe complementario de Sanger, ASO-PCR, fragmentos, etc).
- Además se puede incluir, si procede, una tabla con las variantes encontradas en secuenciaciones previas del mismo paciente para conocer la evolución de los resultados.

Tabla 23. Modelo de tabla de resultados en el informe de NGS.

GEN	EXÓN	CAMBIO DE NUCLEÓTIDO	CAMBIO DE PROTEÍNA	TIPO DE VARIANTE	TRANSCRITO RefSeq ID	BASE DE DATOS	VAF (%)	PROFUNDIDAD	CATEGORÍA

Nota: las posiciones genómicas están descritas respecto a la secuencia del genoma de referencia GRCh37 (hg19).

OBSERVACIONES DEL INFORME

Se recomienda añadir las siguientes observaciones en el informe:

Limitaciones del estudio

Existen algunos tipos de mutaciones que NO pueden ser identificados mediante esta prueba:

- Hay regiones del genoma cuyas características hacen que no sea posible determinar con exactitud los cambios en la secuencia.
- Hay regiones codificantes de los genes analizados que pueden no haberse cubierto al 100%. Incluir una tabla con las regiones no cubiertas o señalar que estarán disponibles bajo petición al laboratorio.
- Esta tecnología no permite la identificación de translocaciones ni reordenamientos equilibrados ni variaciones en el número de copias (CNV).
- Las regiones homopoliméricas (>8nt) no pueden ser secuenciadas con precisión.
- La presencia de variantes poco frecuentes en la secuencia de DNA podrían impedir la obtención de un número suficiente de secuencias amplificadas, dificultando la obtención de un resultado fiable para esa región del genoma.
- No se incluyen variantes en el DNA no codificante, pseudogenes, regiones repetitivas, o alteraciones epigenéticas.
- Ninguna variante se considera por debajo de 25 lecturas, por tanto, pueden existir variantes reales en la muestra que no se hayan comunicado por estar bajo el umbral de sensibilidad que se ha establecido.

Consideraciones

El estudio se ha realizado con DNA procedente de muestra tumoral, de manera que no se han realizado estudios específicos para descartar o confirmar alteraciones germinales potencialmente hereditarias. Si se sospecha una enfermedad hereditaria, su médico debe referirle a una consulta de consejo genético.

Cláusula de exención de responsabilidad diagnóstica

Se debería incluir una cláusula de exención de responsabilidad diagnóstica.

MODELO DE INFORME DE NGS CONSENSUADO POR EL GESMD



Logo Hospital



Nombre Laboratorio

INFORME DEL ESTUDIO DE MUTACIONES MEDIANTE PANEL SMD (NGS)

Nombre paciente: Código muestra:
 Tipo de estudio: Tipo de muestra:
 Diagnóstico: Fecha de entrada:
 Servicio / Unidad: Centro:
 Facultativo: Fecha de análisis:

CONSENTIMIENTO INFORMADO

Se asume la posesión del consentimiento informado por parte del solicitante.

CONSENTIMIENTO INFORMADO

El estudio genético llevado a cabo en el DNA extraído de la muestra recibida consiste en el análisis de la región codificante completa o de las regiones *hotspot* y regiones intrónicas flanqueantes (+/- 5pb) de los genes incluidos en el panel de secuenciación masiva de nueva generación (NGS). Se ha utilizado el panel de genes XXXXX que identifica de manera simultánea mutaciones en 54 genes detallados a continuación:

- región codificante completa: BCOR, BCORL1, CDKN2A, CEBPA, CUX1, DNMT3A, ETV6/TEL, EZH2, IKZF1, KDM6A, PHF6, RAD21, RUNX1, STAG2, ZRSR2.
- regiones de hotspot: ABL, ASXL1, ATRX, BRAF, CALR, CBL, CBLB, CBL, CSF3R, FBXW7, FLT3, GATA1, GATA2, GNAS, HRAS, IDH1, IDH2, JAK2, JAK3, KIT, KRAS, MLL, MPL, MYD88, NOTCH1, NPM1, NRAS, PDGFRA, PTEN, PTPN11, SETBP1, SF3B1, SMC1A, SMC3, SRSF2, TET2, TP53, U2AF1, WT1.

*el análisis de FLT3-ITD no está incluido en el panel.

Para la secuenciación de esta muestra de DNA se ha empleado la tecnología Illumina MiSeq con una cobertura media de 1000X y una uniformidad >95%.

Para el análisis de las variantes se ha empleado el software XXXX. Se han aplicado filtros para eliminar del análisis variantes intrónicas (se excluyen de los filtros sitios de splicing flanqueantes al exón), variantes sin efecto en la proteína y variantes con elevada frecuencia en la población (MAF>1%), ya que se consideran cambios polimórficos (SNPs) sin relevancia clínica. Para establecer la categoría de las variantes encontradas se ha utilizado la Guía de secuenciación masiva en SMD y LMMC propuestas por el Grupo Español de Síndromes Mielodisplásicos -GESMD- (Tabla 1).

1



Logo Hospital



Nombre Laboratorio

INFORME DEL ESTUDIO DE MUTACIONES MEDIANTE PANEL SMD (NGS)

Nombre paciente: Código muestra:
 Tipo de estudio: Tipo de muestra:
 Diagnóstico: Fecha de entrada:
 Servicio / Unidad: Centro:
 Facultativo: Fecha de análisis:

Tabla 1

CATEGORÍA	CRITERIO I VARIANTE RECURRENTE EN BASE DE DATOS ¹⁰¹	CRITERIO II TIPO DONDE SE HA DESCRITO LA VARIANTE ¹⁰²	CRITERIO III ACCIONABILIDAD DEL GEN ALTERADO	CRITERIO IV PREDICTORES Y ESTUDIOS FUNCIONALES
CLASE 1	Variante descrita como PATOGENICA	En la enfermedad o tejido a estudio	Gen Accionable en la enfermedad a estudio	NP
CLASE 2	Variante descrita como PATOGENICA	En OTRA enfermedad o tejido a estudio	Gen Accionable en OTROS tumores sólidos o hematológicos	NP
CLASE 3	Variante NO descrita	NP	Gen Accionable en la enfermedad a estudio	3A. Posiblemente patogénica 3B. De significado incierto 3C. Posiblemente benigna
CLASE 4	Variante NO descrita	NP	Gen accionable en OTROS tumores sólidos o hematológicos	4A. Posiblemente patogénica 4B. De significado incierto 4C. Posiblemente benigna
CLASE 5	Variante NO descrita	NP	Accionabilidad desconocida	

NP: No procede. ¹⁰¹ Variante identificada asociada a la patología a estudio por al menos 2 autores.
¹⁰² Implicación clínica de la identificación de la variante. Se considera un gen accionable aquel que tiene en una determinada entidad clínica un valor como diana terapéutica, como elemento predictivo/ pronóstico y/o presenta implicaciones en el diagnóstico / clasificación.

RESULTADOS

NO se han detectado variantes en el análisis realizado / Se han identificado variantes en los genes _____(1), _____(2), etc. (Tabla 2).

Tabla 2. Detalle de los resultados obtenidos.

GEN	EXÓN	CAMBIO DE NUCLEÓTIDO	CAMBIO DE PROTEÍNA	TIPO DE VARIANTE	TRANSCRITO RefSeq ID	BASE DE DATOS	VAF (%)	PROFUNDIDAD	CATEGORÍA

Abreviaturas: VAF; frecuencia de la variante alélica; NM del transcrito de referencia utilizado; RefSeq ID; Categoría según las Guías de NGS en SMD y LMMC del GESMD. *Mutaciones validadas mediante las técnicas de RFLP, ASO-PCR, ARMS-PCR y de Sanger.

2



Logo Hospital



Nombre Laboratorio

INFORME DEL ESTUDIO DE MUTACIONES MEDIANTE PANEL SMD (NGS)

Nombre paciente:	Código muestra:
Tipo de estudio:	Tipo de muestra:
Diagnóstico:	Fecha de entrada:
Servicio / Unidad:	Centro:
Facultativo:	Fecha de análisis:

INTERPRETACIÓN CLÍNICA / CONCLUSIONES

Interpretación de los resultados y/o conclusiones, para que los resultados sean comprensibles para el médico solicitante y/o el paciente. Da mucho valor al informe.

RECOMENDACIONES (SI APLICA)

Recomendaciones para la repetición de pruebas o la realización de pruebas adicionales (cuando se considere necesario); o recomendaciones sobre si hay indicaciones para el tratamiento, etc. (ver cómo lo definimos/utilizamos...).

Fecha de emisión del informe: XX/XX/XXXX

Informe realizado por:

Informe revisado por:

Firma: Dra. _____
Unidad/Servicio _____

Firma: Dra. _____
Unidad/Servicio _____

REGIONES NO CUBIERTAS

A continuación se detallan los genes y exones que no alcanzan a la profundidad mínima establecida en el 100% de las bases.

Tabla 3. Regiones que no alcanzan la profundidad media establecida.

GEN	Cr.	EXÓN	TRANSCRITO RefSeq ID	cDNA	POSICIÓN DE INICIO	TAMAÑO	MEDIA DE COBERTURA
-----	-----	------	----------------------	------	--------------------	--------	--------------------

Abreviaturas: Cr. Cromosoma



Logo Hospital



Nombre Laboratorio

INFORME DEL ESTUDIO DE MUTACIONES MEDIANTE PANEL SMD (NGS)

Nombre paciente:	Código muestra:
Tipo de estudio:	Tipo de muestra:
Diagnóstico:	Fecha de entrada:
Servicio / Unidad:	Centro:
Facultativo:	Fecha de análisis:

DETALLES ADICIONALES

Descripción detallada de la metodología incluyendo el tipo de secuenciación (captura/amplicón) y la plataforma secuenciación utilizada.

Descripción controles de calidad empleados, cobertura media alcanzada y la profundidad mínima.

Descripción del porcentaje de bases cubiertas a la profundidad mínima establecida (ej. 98% a 500x).

Listado bases de datos utilizadas: poblacionales, somáticas, propias, etc.

Descripción de la categorización según el GESMD.

Tabla 3. Regiones concretas cubiertas con el panel empleado.

GEN	TRANSCRITO	EXONES



Logo Hospital



Nombre Laboratorio

INFORME DEL ESTUDIO DE MUTACIONES MEDIANTE PANEL SMD (NGS)

Nombre paciente:	Código muestra:
Tipo de estudio:	Tipo de muestra:
Diagnóstico:	Fecha de entrada:
Servicio / Unidad:	Centro:
Facultativo:	Fecha de análisis:

LIMITACIONES

No pueden ser identificadas mediante esta prueba:

- Esta tecnología no permite la identificación de translocaciones ni de reordenamientos equilibrados, ni variaciones en el número de copias (CNV).
- No se incluyen variantes en el DNA no codificante, pseudogenes, expansiones repetitivas de trinucleótidos, o alteraciones epigenéticas.
- Hay regiones del genoma cuyas características hacen que no sea posible determinar con exactitud los cambios en la secuencia (ej. regiones homopoliméricas >8 nucleótidos).
- La presencia de variantes poco frecuentes en la secuencia de DNA podría impedir la obtención de un número suficiente de secuencias amplificadas, dificultando la obtención de un resultado fiable para esa región del genoma.
- Ninguna variante se considera por debajo de 500 lecturas, por tanto, pueden existir variantes reales en la muestra que no se hayan comunicado por estar bajo el umbral de sensibilidad que se ha establecido.
- El estudio se ha realizado con DNA procedente de tumor, de manera que no se han realizado estudios específicos para descartar o confirmar alteraciones germinales, potencialmente hereditarias. Si se sospecha una enfermedad hereditaria, su médico debe referirle a una consulta de consejo genético.

CLÁUSULA DE EXENCIÓN DE RESPONSABILIDAD DIAGNÓSTICA

Los servicios de asistencia al diagnóstico genético con secuenciación masiva de DNA están destinados exclusivamente a profesionales de la salud cualificados para su interpretación. Los resultados obtenidos mediante estos estudios y la información que se pueda derivar de los mismos no pueden ser considerados en ningún caso como sustitutivos del consejo diagnóstico o tratamiento médico de un profesional especializado, ni constituyen por sí mismos una consulta médica. Los resultados obtenidos mediante determinaciones analíticas basadas en secuenciación masiva de DNA han de ser interpretadas, junto con otros datos clínicos, dentro del contexto general de una consulta médica que ha de estar dirigida por profesionales especialistas en diagnóstico genético y/o clínico. EL LABORATORIO/HOSPITAL no se hace responsable del uso que haga el contratante de sus servicios ni de los resultados obtenidos mediante sus análisis de estudios, así como tampoco de las eventuales consecuencias perjudiciales derivadas de este uso, haciendo expresa reserva de ejercer las acciones legales oportunas en el supuesto de un uso indebido de los citados estudios y análisis.

Los datos obtenidos en este estudio son confidenciales y se regulan según la ley de protección de datos de carácter personal (Ley Orgánica 15/1999, de 13 de diciembre, revisada el 6 de marzo de 2011).



Babraham Institute. Babraham Bioinformatics - FastQC A Quality Control tool for HighThroughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Published 2016. Accessed September 23, 2017.

Babraham Institute. Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Published 2017 Accessed September 23, 2017.

Novocraft Technologies Sdn Bhd. NovoAlign | Novocraft.<http://www.novocraft.com/products/novoalign/>. Published 2014. Accessed September 23, 2017.

Broad Institute. Picard Tools - By Broad Institute. <http://broadinstitute.github.io/picard/>. Published 2017. Accessed September 23, 2017.

The Regents of the University of California. UCSC Genome Browser Home. <http://genome.ucsc.edu/>. Published 2017. Accessed September 23, 2017.

National Center for Biotechnology Information. NCBI's genome browser for human (Homo sapiens) - Genome Data Viewer. <https://www.ncbi.nlm.nih.gov/genome/gdv/>. Published 2017. Accessed September 23, 2017.

EMBL-EBI. Ensembl genome browser 90. <http://www.ensembl.org/index.html>. Published 2017. Accessed September 23, 2017.

Broad Institute. Downloads | Integrative Genomics Viewer. <http://software.broadinstitute.org/software/igv/download>. Published 2016. Accessed September 23, 2017.

EMBL-EBI. Ensembl Genome Browser. <http://grch37.ensembl.org/index.html>. Published 2017. Accessed September 23, 2017.

Washington U of. NHLBI Exome Sequencing Project (ESP) Exome Variant Server. <http://software.broadinstitute.org/software/igv/download>. Published 2015. Accessed September 23, 2017.

National Center for Biotechnology Information. Home - SNP - NCBI. <https://www.ncbi.nlm.nih.gov/snp>. Published 2017. Accessed September 23, 2017.

National Center for Biotechnology Information. Home - dbVar - NCBI. <https://www.ncbi.nlm.nih.gov/dbvar>. Published 2017. Accessed September 23, 2017.

Exome Aggregation Consortium (ExAC). ExAC Browser. <http://exac.broadinstitute.org/>. Published 2016. Accessed September 23, 2017.

National Center for Biotechnology Information. ClinVar. <https://www.ncbi.nlm.nih.gov/clinvar/>. Published 2017. Accessed September 23, 2017.

Sanger Institut. COSMIC: Catalogue of Somatic Mutations in Cancer - Home Page. <http://cancer.sanger.ac.uk/cosmic>. Published 2017. Accessed September 23, 2017.

International Agency for Research on Cancer. Welcome - IARC TP53 Database. <http://p53.iarc.fr/>. Published 2016. Accessed September 23, 2017.

Vanderbilt-Ingram Cancer Center. Genetically Informed Cancer Medicine - My Cancer Genome. <https://www.mycancergenome.org/>. Published 2017. Accessed September 23, 2017.

The University of Texas MD Anderson Cancer Center. Knowledge Base for Precision Oncology. <https://pct.mdanderson.org/>. Published 2017. Accessed September 23, 2017.

Group TBG. IntOGen - mutational cancer drivers database. <http://www.intogen.org/search>. Published 2014. Accessed September 23, 2017.

St. Jude Children's Research Hospital and Washington University School of Medicine in St. Louis. Pediatric Cancer Genome Project | Explore - St. Jude Children's Research Hospital, Washington University School of Medicine in St. Louis. <http://explorepcgp.org/>. Published 2017. Accessed September 23, 2017.

International Cancer Genome Consortium. Welcome | ICGC Data Portal. <https://dcc.icgc.org/>. Published 2017. Accessed September 23, 2017.

National Cancer Institute. Genomic Data Commons Data Portal. <https://portal.gdc.cancer.gov/>. Published 2017. Accessed September 23, 2017.

Center for Molecular Oncology. cBioPortal for Cancer Genomics. <http://www.cbioportal.org/>. Accessed September 23, 2017.

Cardiff University. HGMD® home page. <http://www.hgmd.cf.ac.uk/ac/index.php>. Published 2017. Accessed September 23, 2017.

Leiden University Medical Center. LOVD v.3.0 - Leiden Open Variation Database. Online gene-centered collection and display of DNA variations. <http://www.lovd.nl/3.0/home>. Published May 2017. Accessed September 23, 2017.

European Medicines Agency. Clinical Trials Register. <https://www.clinicaltrialsregister.eu/ctr-search/search>. Published 2017. Accessed September 23, 2017.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS SS. PolyPhen-2: prediction of functional effects of human nsSNPs. <http://genetics.bwh.harvard.edu/pph2/>. Published 2016. Accessed September 23, 2017.

Craig Vente Institute J. SIFT. <http://sift.jcvi.org/>. Published 2012. Accessed September 23, 2017. Computational Biology Center | Memorial Sloan Kettering Cancer Center.

MutationAssessor.org /// functional impact of protein mutations. <http://mutationassessor.org/r3/>. Published 2015. Accessed September 23, 2017.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. de Brevern AG, ed. PLoS One. 2012;7(10):e46688. doi:10.1371/journal.pone.0046688. Universitaetsmedizin Berlin - Charite. PROVEAN. <http://provean.jcvi.org/index.php>. Published 2015.

González-Pérez A, López-Bigas N. FannsDB is a database for Functional ANnotations for Non Synonymous SNVs which contains precalculated scores for several predictors. <http://bbglab.irbbarcelona.org/fannsdb/>. Published 2014. Accessed September 23, 2017.

Slashdot Media. Consensus Variant Effect Classification (CoVEC) - Browse Files at SourceForge.net. <https://sourceforge.net/projects/covec/files/>. Published 2017. Accessed September 23, 2017.

University of Washington and Hudson-Alpha Institute for Biotechnology. Combined Annotation Dependent Depletion (CADD). <http://cadd.gs.washington.edu/>. Published 2014. Accessed September 23, 2017.

Leland Stanford Junio University. Sidow Lab - GERP. <http://mendel.stanford.edu/sidowlab/downloads/gerp/>. Published 2011. Accessed September 23, 2017.

Human Splicing Finder - Genetics and Bioinformatics Team. Human Splicing Finder - Version 3.0. <http://www.umd.be/HSF3/>. Published 2017. Accessed September 23, 2017.

National Science Foundation. MaxEntScan::scoresplice. http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html. Published 2004. Accessed September 23, 2017.

Center for Biological Sequence Analysis (CBS). NetGene2 Server. <http://www.cbs.dtu.dk/services/NetGene2/>. Published 2017. Accessed September 23, 2017.

The Berkeley Drosophila Genome Project (BDGP) is a consortium of the Drosophila Genome Center, funded by the National Human Genome Research Institute, National Cancer Institute, and Howard Hughes Medical Institute, through its support of work in the Susan and GK laboratories. BDGP: Splice Site Prediction by Neural Network. http://www.fruitfly.org/seq_tools/splice.html. Published 2017. Accessed September 23, 2017.

Center for Bioinformatics & Computational Biology (CBCB) - University of Maryland. GeneSplicer web interface. http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml. Published 2017. Accessed September 23, 2017.

© 2017 GESMD

Reservados todos los derechos. El contenido de la presente publicación no puede ser reproducido, ni transmitido por ningún procedimiento electrónico o mecánico, incluyendo la fotocopia o grabación magnética, ni registrado por ningún sistema de recuperación de información, en ninguna forma, ni por ningún medio, sin la previa autorización por escrito del propietario del Copyright.

Depósito legal: B 24292-2017

ISBN: 978-84-697-6291-2

Editado por:



**MARKETING FARMACÉUTICO &
INVESTIGACIÓN CLÍNICA, S.L.**

Secretari Coloma, 64-68, esc. B, entlo. 5a

08024 Barcelona

Tel.: (34) 93 434 44 12

Fax.: (34) 93 253 11 68

El contenido de este manual es el resultado de la libre opinión científica de los miembros del Grupo de Trabajo que la suscriben.





GRUPO ESPAÑOL DE
SÍNDROMES MIELODISPLÁSICOS



**GUÍA DE APLICACIÓN CLÍNICA
DE LA SECUENCIACIÓN MASIVA EN
SÍNDROMES MIELODISPLÁSICOS Y
LEUCEMIA MIELOMONOCÍTICA CRÓNICA**



**GRUPO ESPAÑOL DE
SÍNDROMES MIELODISPLÁSICOS**

Con el aval científico de:



Sociedad Española de
Hematología y Hemoterapia

