Universitat Oberta de Catalunya

uoc.edu

Análisis de herramientas bioinformáticas para la detección de expansiones y alteraciones estructurales mediante secuenciación de genoma completo

Irene Hidalgo Mayoral
Máster Universitario en Bioinformática y Bioestadística
Área del trabajo final: Bioinformática clínica

Consultor
Joan Maynou Fernández

Profesor responsable de la asignatura
David Merino Arranz

Tutor
Jose Miguel Lezana Rosales

Fecha de entrega
06/2021

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons
Copyright © 2021 Irene Hidalgo Mayoral

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU Free Documentation License".

Copyright

© (Irene Hidalgo Mayoral)
Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis de herramientas bioinformáticas para la detección de expansiones y alteraciones estructurales mediante secuenciación de genoma completo
Nombre del autor:	Irene Hidalgo Mayoral
Nombre del consultor/a:	Joan Maynou Fernández
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	Master Universitario en Bioinformática y Bioestadística
Área del Trabajo Final:	Bioinformática clínica
Idioma del trabajo:	Castellano
Número de créditos:	15
Palabras clave	*Benchmarking, WGS, SVs, expansiones*

Resumen del Trabajo

La secuenciación del genoma completo (WGS) resulta una aproximación de gran interés al permitir secuenciar el genoma completo de un individuo y posibilitar la detección de una mayor variedad de alteraciones genómicas en un único estudio. A día de hoy la tecnología más extendida por su antigüedad y coste es la *short-read sequencing, que* presenta un elevado rendimiento en la detección de variaciones puntuales e inserciones/deleciones de pequeño tamaño. En este trabajo se ha evaluado el potencial de distintas herramientas bioinformáticas en la detección de variaciones estructurales (SVs) y expansiones (STRs) a partir de datos de WGS *short-read*.

Se han utilizado muestras reales obtenidas de repositorio con SVs y expansiones caracterizadas previamente. Se seleccionaron las herramientas Manta, Delly y Lumpy para la detección de SVs y ExpansionHunter, GangSTR y TREDParse para la detección de expansiones, escogidas en base al uso de métodos de detección combinados, un mayor rendimiento descrito en la literatura y/o un uso extendido en la comunidad científica. El rendimiento de las herramientas se ha valorado en términos de sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y grado de concordancia entre detectores frente a un fichero *gold-standard*.

Tras la evaluación se ha observado un rendimiento limitado de los detectores. En el caso de las SVs el desempeño es dependiente del tipo de alteración y de su tamaño, mientras que en el caso de las STRs su desempeño está en relación con el tamaño de la expansión y el contexto genómico de la región que la contiene.

Abstract

Whole genome sequencing (WGS) is an interesting approach as it allows the complete genome of an individual to be sequenced and allows the possibility of detecting a greater variety of genomic alterations in a single study. Today, the most widespread technology due to its age and cost is short-read sequencing, which has a high performance in detecting point variations and small insertions/deletions. In this work, the potential of different bioinformatics tools in the detection of structural variations (SVs) and expansions (STRs) from WGS short-read data has been evaluated.

Real samples obtained from the repository with previously characterized SVs and STRs have been used. The selected tools were Manta, Delly and Lumpy for SVs and ExpansionHunter, GangSTR and TREDParse for STRs, which were chosen based on the use of combined detection methods, higher performance described in the literature, and widespread use in the scientific community. Their performance has been assessed in terms of sensitivity, specificity, positive predictive value, negative predictive value and degree of agreement between detectors against a gold standard file.

After evaluation, limited performance of the detectors has been observed. For SVs, its performance is dependent on the type and size of the event, while in the case of STRs, its performance is related to the size of the expansion and the genomic context of the region that contains it.

Índice

Lista de figuras

Lista de tablas

1 Introducción

Los trastornos genéticos representan un importante grupo de patologías causadas por alteraciones en el material genético. Se caracterizan por presentar un espectro fenotípico muy heterogéneo, que abarca desde enfermedades con afectación aislada de debut tardío hasta desórdenes congénitos complejos. Como entidades aisladas su prevalencia es baja, pero de forma conjunta los trastornos genéticos suponen hasta un 80% de las enfermedades raras[(1)], habiéndose descrito en la actualidad cerca de 7.000 trastornos con causa molecular conocida de acuerdo al catálogo OMIM[(2)] (*Online Mendelin Inheritance in Man*).

El diagnóstico genético de los individuos afectos permite identificar la causa del trastorno, prever su evolución, adecuar el tratamiento y conocer los riesgos de transmisión a la descendencia. La utilidad clínica es extensible a los miembros de la familia, al permitir la identificación de los familiares en riesgo de desarrollar enfermedad, sobre los que puede instaurarse un seguimiento o tratamiento precoz [(3)]. En los últimos años, los avances en el conocimiento de los mecanismos involucrados en las patologías y el desarrollo de la tecnología disponible han hecho posible un aumento del rendimiento diagnóstico. Sin embargo, en muchos casos el diagnóstico sigue siendo un proceso complejo y dilatado en el tiempo.

La secuenciación del ADN se hizo posible a finales de la década de los 70 con la llegada de la secuenciación Sanger [(4)] y su posterior automatización, que permite secuenciar fragmentos de ADN de aproximadamente 600-800 pares de bases (pb) mediante el uso de dideoxinucleótidos marcados con fluorescencia. Posteriormente irrumpieron las técnicas de secuenciación masiva (NGS, del inglés *Next-generation sequencing*)[(5)], capaces de secuenciar millones de fragmentos de ADN de forma simultánea y que permitieron el estudio de múltiples genes en un único ensayo, reduciendo los tiempos de procesado ylos costes de la secuenciación. Inicialmente, el abordaje consistía en el estudio de paneles acotados a un número determinado de genes enfocados a la detección de patologías específicas. Los paneles presentan la ventaja de tener un diseño muy optimizado, que permite la secuenciación de los genes de interés con una elevada profundidad de lectura, y además evita la detección de variantes no relacionadas con lapatología de estudio (hallazgos incidentales) [(6)]. Sin embargo, el descubrimiento constante de nuevas relaciones genotipo-fenotipo requiere de la actualización periódica de los paneles, limitando su rendimiento, por lo que la secuenciación del exoma completo (WES, del inglés *whole exome sequencing*) se ha impuesto como técnica de elección en la rutina asistencial de muchos Servicios de Genética a nivel hospitalario [(7)]

El exoma se define como la región codificante del genoma, cuya información permite la síntesis de proteínas. Si bien el exoma supone el 2% del genoma, el 85% de las variantes genéticas deletéreas se encuentran en él [(8)]. El estudio mediante WES resulta de especial interés en el estudio de desórdenes complejos con un fenotipo poco definido y/o en los que se han identificado un gran número de genes potencialmente implicados. Las aproximaciones en su estudio son variadas e incluyen el análisis de paneles virtuales de genes específicos, análisis del exoma clínico (genes con implicación clínica conocida) [(9)] y análisis del exoma completo (conjunto completo de genes del exoma), pudiéndose priorizar genes de interés en base al fenotipo de un individuo mediante el uso de términos HPO (del inglés, Human Phenotipe Ontology) [(10)]. El rendimiento diagnóstico descrito en la literatura científica oscila entre el 24-62% en función de las características de la cohorte estudiada, fundamentalmente en relación a su edad y patología [(11–13)], y puede verse incrementado al utilizar una estrategia de análisis en trío (estudio en paralelo del probando y sus progenitores) [(14)].

En el resto de pacientes, sin embargo, no es posible llegar a un diagnóstico genético y puede ser debido a las limitaciones propias de la técnica utilizada[15–17], que no permite detectar variantes en regiones intrónicas profundas, variaciones en el número de copias (CNVs, del inglés *copy-number variants*) en regiones no capturadas o que involucran a un número reducido de exones, ni identificar determinadas alteraciones estructurales (SVs, del inglés *structural variants*) o expansiones (STRs, del inglés *short tandem repeats)*.

Se entiende por SV los reordenamientos cromosómicos que pueden conllevar una ganancia o pérdida de material genético (SVs desbalanceadas: deleciones, inserciones, duplicaciones y CNVs), o bien un cambio en la orientación o localización del material genético (SVs balanceadas: inversiones y traslocaciones). La detección de este tipo de eventos se realiza habitualmente mediante cariotipo o técnicas de hibridación fluorescente *in situ* (FISH), análisis de *microarrays* cromosómicos (CMA) o MLPA (del inglés, *multiplex ligation-dependent probe amplification*), en función de su tamaño. Por otra parte, se entiende por expansión el aumento del número de repeticiones de secuencias repetitivas del ADN (STRs), que habitualmente son detectadas mediante TP-PCR (del inglés, *triplet repeat Primed-PCR*).

La secuenciación del genoma completo (WGS, del inglés *whole genome sequencing*) permite solventar dichas limitaciones [18] al obtener perfiles de cobertura más uniformes y que abarcan casi la totalidad del genoma [19–21], incluyendo las regiones codificantes y no codificantes del genoma de un individuo. Inicialmente consideradas nofuncionales, las regiones no codificantes desempeñan un importante papel en la regulación de la expresión génica. Con el objetivo de identificar todos los elementos funcionales del genoma humano, surgió en 2003 el proyecto ENCODE [22] (del inglés, *The Encyclopedia of DNA Elements Consortium*), que permitió determinar que aproximadamente el 80% del genoma contiene elementos funcionalmente relevantes. Desde entonces, numerosas investigaciones han contribuido a establecer el papel del ADN no codificante en la regulación de la expresión génica y el correcto plegamiento proteico, y hoy en día está ampliamente descrita su implicación en el desarrollo de ciertas patologías [23,24].

Se ha descrito un aumento del rendimiento diagnóstico del WGS comparado con WES variable, que oscila entre el 1-7% en función de la cohorte consultada [27,28]. En consecuencia, la inclusión de las regiones no codificantes junto con la posibilidad de detección de una mayor variedad de alteraciones genómicas en un único estudio hacen de la WGS una aproximación de gran interés, si bien su uso a día de hoy se limita al ámbito de la investigación.

La secuenciación del genoma completo puede abordarse mediante la utilización de tecnología secuenciación de lecturas cortas (*short-read sequencing)* o tecnología de lecturas largas (*long-read sequencing)*. La principal diferencia reside en la longitud de las lecturas generadas, siendo ésta una de las variables fundamentales a tener en cuenta en función del tipo de alteración genética a detectar. La tecnología *de* lecturas cortas se caracteriza por generar lecturas de hasta 600 pares de bases en funciónde la plataforma utilizada, que permiten la detección de variantes puntuales (SNVs, delinglés *Single Nucleotide Variant*) e inserciones/deleciones de pequeño tamaño (INDELs) con una elevada precisión. Esta aproximación requiere de un proceso de amplificación que introduce un sesgo de secuenciación y es muy dependiente de la cobertura de la región de interés, lo que unido a la dificultad de alinear lecturas cortas frente al genoma de referencia en regiones con una elevada variabilidad interindividual o con pseudogenes, limita su rendimiento en el abordaje de SVs y expansiones. Por su parte,la tecnología de lecturas largas genera lecturas con un tamaño superior a 10 kb en función de la plataforma utilizada obtenidas en ausencia de una amplificación clonal, permitiendo un desempeño superior en la detección de variantes estructurales y en la gestión de regiones repetitivas del genoma [25,26].

A día de hoy, la tecnología más utilizada por antigüedad y coste es la tecnología de lecturas cortas. Las herramientas bioinformáticas disponibles presentan elevados rendimientos en la detección de SNVs e INDELs, pero su desempeño en la detección de variaciones estructurales y STRs es aún limitado y los estudios de comparación entre ellas son escasos. Sin embargo, el progresivo abaratamiento de los costes de secuenciación y el potencial beneficio diagnóstico hacen prever la transición de la WGS de lecturas cortas de su uso en investigación a la rutina asistencial hospitalaria en el futuro [29].

1.1 Contexto y justificación del Trabajo

En este contexto, se está desarrollando en nuestro centro un *pipeline* de análisis de WGS, adaptado del *pipeline* bioinformático actualmente empleado en el estudio de WES. En este Trabajo de Fin de Máster (TFM) se pretende evaluar y comparar herramientas bioinformáticas de detección de expansiones y variaciones estructurales a partir de datos de secuenciación de genomas.

Se ha escogido este tema por su utilidad clínica y traslacional, ya que una vez evaluadas las diferentes herramientas se pretende implementarlas en el *pipeline* general de análisis de WGS de nuestro centro de trabajo para su uso en proyectos de investigación y/o incorporación a la rutina asistencial en el futuro.

1.2 Objetivos del Trabajo

Implementación de un *pipeline* de análisis para la detección de SVs y expansiones a partir de datos de secuenciación de genoma completo en muestras reales obtenidas a partir de repositorios.

<u>-Objetivos específicos</u>

A. Selección de herramientas
   - Búsqueda en la literatura científica de las diferentes aproximaciones para el estudio de las alteraciones de interés y de las diferentes herramientas bioinformáticas existentes.
   - Análisis comparativo de las diferentes herramientas y selección de los detectores candidatos para realizar el estudio de cada una de las alteraciones de interés
   - Búsqueda e instalación de los diferentes *softwares* seleccionados

B. Selección de genomas
   - Búsqueda y obtención de genomas con las alteraciones de interés detectadas y validadas por técnicas de referencia de repositorios públicos. Se seleccionarán de forma preferencial ficheros de alineamiento.

C. Evaluación de las herramientas
   - Preparación de los archivos de trabajo y procesamiento de los ficheros de alineamiento por las diferentes herramientas seleccionadas y obtención de los archivos de resultados.
   - Evaluación de los resultados frente a los archivos con los resultados de referencia (*gold-standard*) de cada tipo de alteración, y evaluación del rendimiento de las herramientas en términos de sensibilidad y especificidad.

1.3 Enfoque y método seguido

A día de hoy, la tecnología más extendida en la WGS es la tecnología de secuenciación de lecturas cortas, que se caracteriza por generar lecturas de hasta 600 pb en función de la tecnología utilizada, presentar una elevada precisión en la detección de SNVs e INDELs, y presentar un coste menor respecto a tecnologías de lecturas largas. En este trabajo, se realizará un estudio comparativo de las diferentes herramientas bioinformáticas disponibles en la detección de SVs y STRs a partir de datos de WGS generados mediante secuenciación de lecturas cortas.

En la selección de los genomas de trabajo, el criterio de selección fue que contuviesen SVs y expansiones conocidas y validadas por técnicas de referencia o por métodos ortogonales. Los genomas se obtuvieron a partir de repositorios públicos, siendo los ficheros de partida de primera elección ficheros de alineamiento (BAM/CRAM).

En relación a la **identificación de SVs**, existen diferentes aproximaciones basadas en métodos indirectos de gestión de las lecturas (*reads)* respecto al genoma de referencia (Figura 1): (A) detección de cambios en la profundidad de lectura (*read depth*), (B) detección de cambios en la orientación, orden o longitud del inserto de lecturas pareadas (*paired-read*), (C) detección de bloques de pérdidas de lecturas con un mismopunto de ruptura (*split reads*) y/o (D) ensamblaje de las lecturas *de novo* (*de novo assembly*)[30].

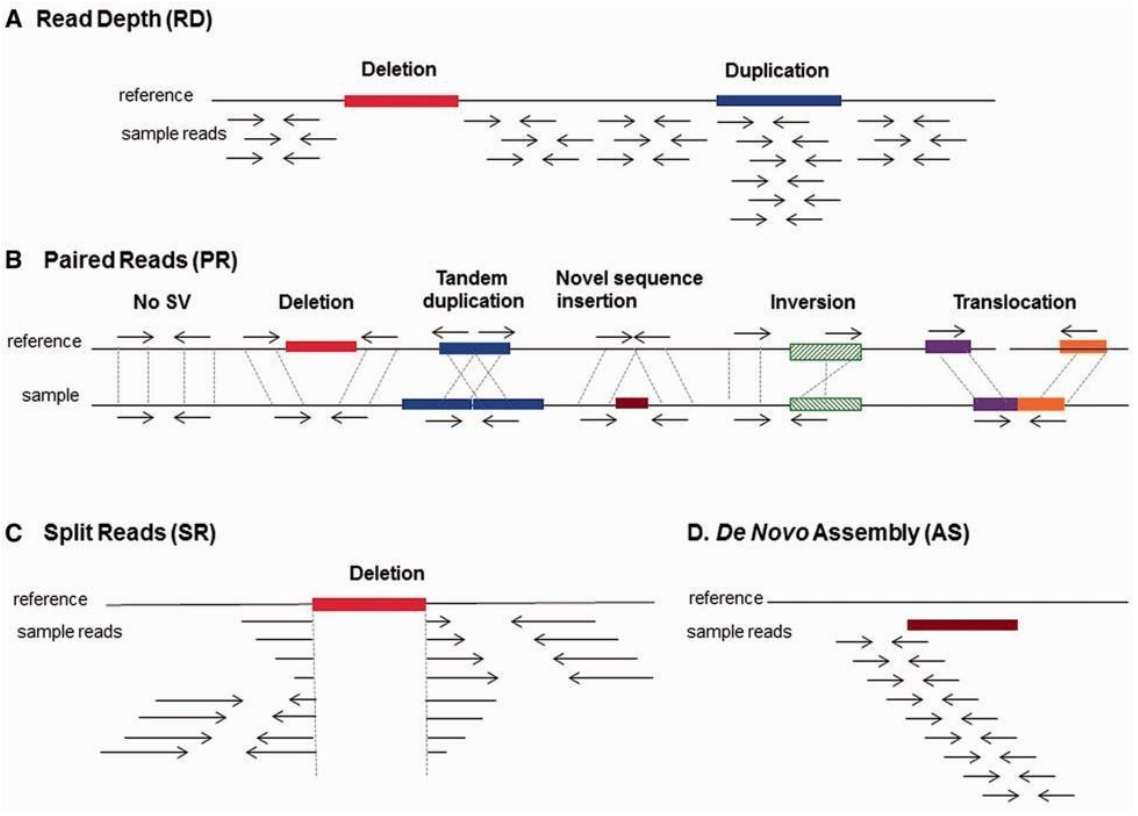


Figura 1. Estrategias en la detección de variaciones estructurales. Tomada de Escaramís G et al (2015). Obtenido con permiso de Copyright Clearance Center's RightsLink®

En relación a la **identificación de expansiones,** existen diferentes aproximaciones que varían en función del número de repeticiones y su localización con respecto a las lecturas (Figura 2): *(A) flanking reads*, la expansión está contenida parcialmente en la *read*, *(B) spanning reads*, la expansión no está contenida en la *read*, pero se encuentra flanqueada por la pareja de *reads, (C) fully repetitive reads* o FRR, la expansión contiene en su totalidad una de las *reads* de la pareja, pero su tamaño es mayor que el de la *read* y/o *(D), enclosing reads,* el tamaño de la lectura es mayor que el de la expansión y la contiene en su totalidad [31,32].

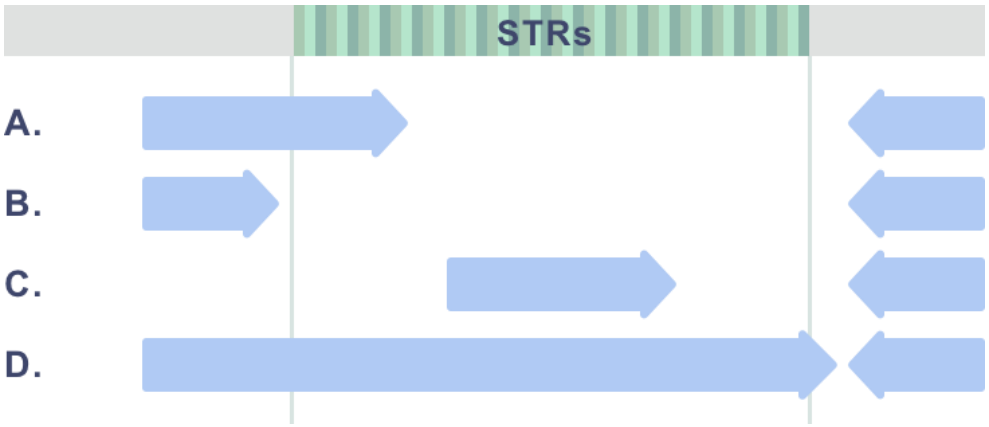


Figura 2. Clases de *paired-read* informativas en el abordaje de la detección de expansiones. (A) *Flanking reads*, (B) *Spanning reads*, (C) *Fully repetitive reads*, (D) *Enclosing reads.* Adaptada de Mousavi N *et al* (2019).

Las herramientas bioinformáticas disponibles para la detección de SVs y expansiones se basanen la combinación de una o varias de las aproximaciones mencionadas. Dado que el uso de una metodología u otra supone una fuente de variabilidad en el rendimiento de las herramientas, su evaluación resulta fundamental. Teniendo en cuenta esta consideración, la aproximación planteada en este trabajo para la selección de las herramientas bioinformáticas a evaluar se basa en la elección de herramientas que utilicen métodos de detección combinados, presenten un mayor rendimiento descrito en la literatura y/o tengan uso más extendido en la comunidad científica.

Una vez seleccionadas las herramientas a evaluar se realizó un procesamiento de los archivos de partida, empleando las diferentes herramientas seleccionadas. Los resultados obtenidos se compararon con los resultados del archivo *gold-standard*. Se realizó un análisis estadístico para evaluar el grado de concordancia entre los programas y su rendimiento en términos de sensibilidad y especificidad.

1.4 Planificación del Trabajo

En primera instancia se definió el ámbito, extensión y los objetivos del Trabajo. La planificación del proyecto se estructuró en dos bloques, uno por cada una de las alteraciones de interés. Cada bloque se constituye de las siguientes etapas: *(i)* búsqueda bibliográfica de las herramientas disponibles, selección e instalación de las herramientas candidatas (dedicación 95 horas), *(ii)* búsqueda de los *dataset* de trabajo, selección y descarga de los ficheros de alineamiento, ficheros *gold-standard* y genomas de referencia (40 horas), *(iii)* desarrollo de un *pipeline* bioinformático para el procesamiento de las muestras y de los ficheros de alineamiento (125 horas) y *(iv)* evaluación y análisis estadístico de los resultados (35 horas). Una vez finalizado el proyecto, se contempla la dedicación de tiempo a la redacción de la memoria (50 horas) y la preparación de la presentación final del trabajo (30 horas).

1.5 Análisis del riesgo

Se llevó a cabo un análisis inicial de los principales problemas que se contemplaron durante la realización del trabajo, recogidos en la Tabla 1:

Riesgos	Solución
Problemas en la obtención de los datos de WGS seleccionados del repositorio público.	En caso de no poder acceder a los datos, se contemplará la posibilidad de cambiar de repositorio.
No disponibilidad de los ficheros de alineamiento.	Se utilizarán ficheros FASTQ, realizándose un análisis de calidad de los datos (FastQC)[33] y un alineamiento de las lecturas frente al genoma de referencia mediante BWA-mem[34] para la obtención de los archivos de alineamiento.

Tabla 1: Análisis del riesgo

1.6 Breve sumario de contribuciones y productos obtenidos

De este trabajo se ha obtenido una memoria detallada de todo el proceso de evaluación de las herramientas bioinformáticas seleccionadas para la detección de SVs y expansiones, incluyendo la metodología, *scripts* bioinformáticos utilizados y los resultados de la evaluación de su rendimiento en términos de sensibilidad, especificidad y grado de concordancia.

Los *scripts* bioinformáticos desarrollados durante el trabajo se han depositado en la plataforma Gitlab, y pueden consultarse a través del siguiente enlace: Link al repositorio Gitlab.

2 Análisis de alteraciones estructurales

2.1 Metodología

2.1.1 Búsqueda de *datasets* de trabajo

Selección de muestras

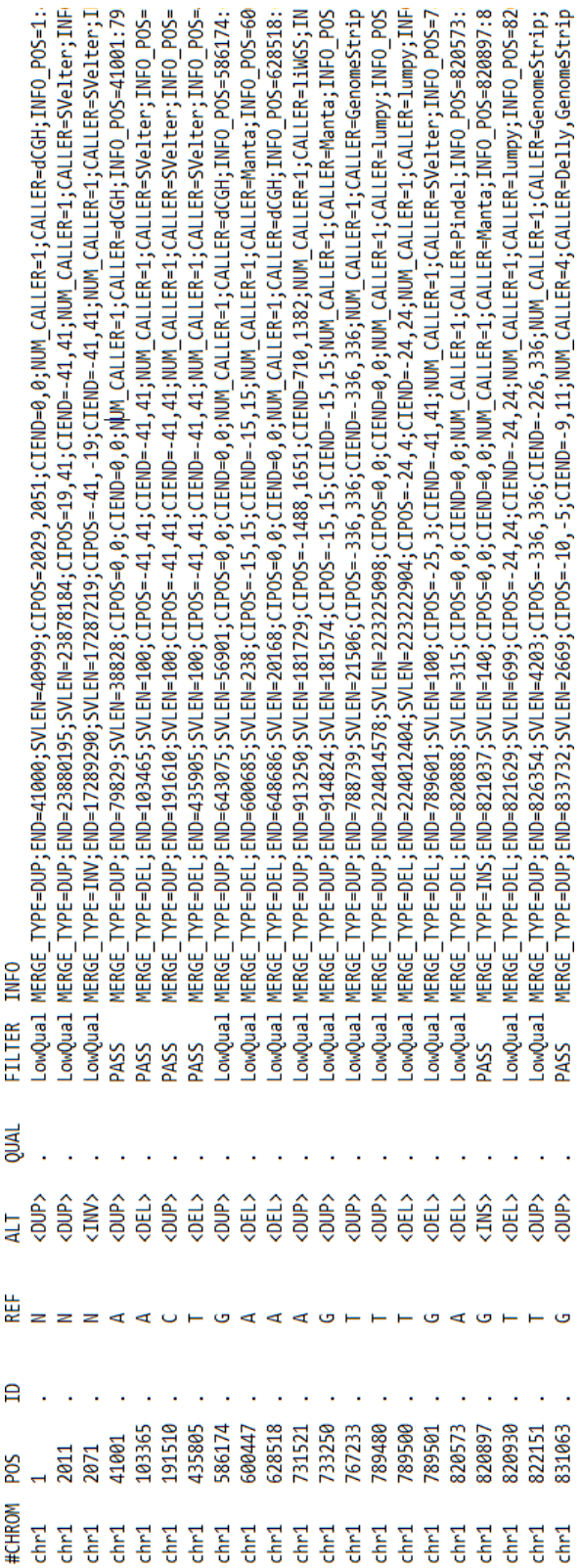
El consorcio internacional HGSVC (del inglés, *The Human Genome Structural Variation Consortium*) es una iniciativa que surge del Instituto Nacional de Investigación del Genoma Humano *(NHGRI, del inglés National Human Genome Research Institute*) con el objetivo de determinar el conjunto de SVs del genoma humano y generar ficheros de resultados de referencia que sirviesen como *gold-standard* en el estudio de las SVs. El proyecto consistió en la secuenciación del genoma de 9 individuos utilizando un amplio espectro de tecnologías entre las que se incluyen: WGS de lectura corta, WGS de lectura larga (SMRT, del inglés *Single Molecule Real-Time*, PacBio®), mapeo óptico y secuenciación StrandSeq (del inglés, *Single-cell DNA template strand sequencing*). Los nueve individuos fueron seleccionados por ser representativos de distintos grados de diversidad genética: trío Yoruban (NA19238, NA19239, NA19240), trío Puerto Rico (HG00731, HG00732, HG00733) y trío China (HG00512, HG00513, HG00514) [36].

Para este trabajo, se seleccionaron como *dataset* de trabajo las nueve muestras mencionadas. En el repositorio público IGSR (del inglés, *The International Genome Sample Resource*) [37], se encuentran depositados los datos de WGS correspondientes al proyecto. Se escogieron los datos generados mediante tecnología de secuencias cortas sin protocolo de PCR y se descargaron los ficheros de alineamiento (CRAM). El alineamiento se había realizado frente al genoma de referencia GRCh38 *plus decoy HLA* (disponible en el servidor FTP de IGSR), que incluye *contigs* alternativos (secuencias alternativas de una región que surgen como consecuencia de la diversidad genética, como por ejemplo los *loci* HLA) y la secuencia del virus del Epstein-Barr (EBV), cuya infección afecta al 90% de la población [38].

Una vez descargados los ficheros de alineamiento y el genoma de referencia, se indexaron mediante la herramienta SAMtools [39] para obtener los ficheros necesarios para el análisis.

Selección del *dataset* de referencia

Como resultado del proyecto HGSVC se generaron diferentes *datasets* de referencia en función de la tecnología de detección. Los *datasets* se encuentran disponibles en el servidor FTP de IGSR, desde donde se descargó el correspondiente a la WGS de lecturas cortas, que recibe el nombre de *Illumina integrate* por la plataforma de secuenciación utilizada. Este archivo, que se considerará en adelante el *gold-standard,* contiene las SVs detectadas mediante las siguientes herramientas bioinformáticas: Delly[40], dCGH, ForestSV[41], GenomeSTRiP[42], Lumpy[43], Manta[44], MELT[45], NovoBreak[46], Pindel[47], SVelter[48], Tardis[49], VariantHunter[50] y Wham[51]. En la Figura 3 se muestra un extracto del fichero, disponible en su totalidad en el anexo del documento.



```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
chr1    1       .       N       <DUP>   .       LowQual MERGE_TYPE=DUP;END=41000;SVLEN=40999;CIPOS=2029,2051;CIEND=0,0;NUM_CALLER=1;CALLER=dCGH;INFO_POS=1:
chr1    2011    .       N       <DUP>   .       LowQual MERGE_TYPE=DUP;END=23880195;SVLEN=23878184;CIPOS=19,41;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;INF
chr1    2071    .       N       <INV>   .       LowQual MERGE_TYPE=INV;END=17289290;SVLEN=17287219;CIPOS=-41,-19;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;I
chr1    41001   .       A       <DUP>   .       PASS    MERGE_TYPE=DUP;END=79829;SVLEN=38828;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=dCGH;INFO_POS=41001:79
chr1    103365  .       A       <DEL>   .       PASS    MERGE_TYPE=DEL;END=103465;SVLEN=100;CIPOS=-41,41;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;INFO_POS=
chr1    191510  .       C       <DUP>   .       PASS    MERGE_TYPE=DUP;END=191610;SVLEN=100;CIPOS=-41,41;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;INFO_POS=
chr1    435805  .       T       <DEL>   .       PASS    MERGE_TYPE=DEL;END=435905;SVLEN=100;CIPOS=-41,41;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;INFO_POS=
chr1    586174  .       G       <DUP>   .       LowQual MERGE_TYPE=DUP;END=643075;SVLEN=56901;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=dCGH;INFO_POS=586174:
chr1    600447  .       A       <DEL>   .       LowQual MERGE_TYPE=DEL;END=600685;SVLEN=238;CIPOS=-15,15;CIEND=-15,15;NUM_CALLER=1;CALLER=Manta;INFO_POS=60
chr1    628518  .       A       <DEL>   .       LowQual MERGE_TYPE=DEL;END=648686;SVLEN=20168;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=dCGH;INFO_POS=628518:
chr1    731521  .       A       <DUP>   .       LowQual MERGE_TYPE=DUP;END=913250;SVLEN=181729;CIPOS=-1488,1651;CIEND=710,1382;NUM_CALLER=1;CALLER=liWGS;IN
chr1    733250  .       G       <DUP>   .       LowQual MERGE_TYPE=DUP;END=914824;SVLEN=181574;CIPOS=-15,15;CIEND=-15,15;NUM_CALLER=1;CALLER=Manta;INFO_POS
chr1    767233  .       T       <DUP>   .       LowQual MERGE_TYPE=DUP;END=788739;SVLEN=21506;CIPOS=-336,336;CIEND=-336,336;NUM_CALLER=1;CALLER=GenomeStrip
chr1    789480  .       T       <DUP>   .       LowQual MERGE_TYPE=DUP;END=224014578;SVLEN=223225098;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=lumpy;INFO_POS
chr1    789500  .       T       <DEL>   .       LowQual MERGE_TYPE=DEL;END=224012404;SVLEN=223222904;CIPOS=-24,4;CIEND=-24,24;NUM_CALLER=1;CALLER=lumpy;INF
chr1    789501  .       G       <DEL>   .       LowQual MERGE_TYPE=DEL;END=789601;SVLEN=100;CIPOS=-25,3;CIEND=-41,41;NUM_CALLER=1;CALLER=SVelter;INFO_POS=7
chr1    820573  .       A       <DEL>   .       LowQual MERGE_TYPE=DEL;END=820888;SVLEN=315;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=Pindel;INFO_POS=820573:
chr1    820897  .       G       <INS>   .       PASS    MERGE_TYPE=INS;END=821037;SVLEN=140;CIPOS=0,0;CIEND=0,0;NUM_CALLER=1;CALLER=Manta;INFO_POS=820897:8
chr1    820930  .       T       <DEL>   .       LowQual MERGE_TYPE=DEL;END=821629;SVLEN=699;CIPOS=-24,24;CIEND=-24,24;NUM_CALLER=1;CALLER=lumpy;INFO_POS=82
chr1    822151  .       T       <DUP>   .       LowQual MERGE_TYPE=DUP;END=826354;SVLEN=4203;CIPOS=-336,336;CIEND=-226,336;NUM_CALLER=1;CALLER=GenomeStrip;
chr1    831063  .       G       <DUP>   .       PASS    MERGE_TYPE=DUP;END=833732;SVLEN=2669;CIPOS=-10,-5;CIEND=-9,11;NUM_CALLER=4;CALLER=Delly,GenomeStrip
```

Figura 3: Extracto del archivo VCF *gold-standard*. En el cabecero se incluye la información proporcionada para cada una de las SV detectadas. CHROM: cromosoma, POS: coordenada de inicio referenciada respecto a cada cromosoma, ID: identificador de dbSNP (si no existe, es "."), REF: nucleótido de referencia, ALT: tipo de SV (DEL: deleción, DUP: duplicación, INV: inversión, INS: inserción), QUAL: calidad de la llamada (no informativa, se referencia como "." en todas las SVs), FILTER: filtro de calidad (PASS: supera el umbral de calidad definido, LowQual: no supera el filtro de calidad), INFO: información sobre el genotipado de la SV. Incluye la coordenada de fin, el tamaño de la alteración, el intervalo de confianza de los puntos de ruptura, el número de herramientas bioinformáticas que la detectaron y cuales fueron.

La construcción del fichero *gold-standard* por el grupo de trabajo del HGSVC se realizó de acuerdo al siguiente esquema: *(i)* procesamiento de las 9 muestras con las herramientas enumeradas previamente, *(ii)* filtrado inicial y exclusión de las SVs pertenecientes a *contigs* alternativos y/o con un tamaño inferior a 50 pb, *(iii)* validación de los resultados con los resultados obtenidos mediante secuenciación de lecturas largas, descartándose las SVs con un grado de solapamiento inferior al 50%, *(iv)* evaluación entre sí de las SVs restantes e integración de las SVs candidatas en un único fichero y *(v)* agrupación de las alteraciones llamadas por más de una herramienta (*clusters*) y refinamiento de los puntos de ruptura para la determinación de una única SV de cada *cluster*. Finalmente, se catalogaron las llamadas en función de su calidad, identificándose como de baja calidad (*LowQual*):

-SVs localizadas en una región centromérica o telomérica
-SVs con un tamaño superior a 1 Mb
-SVs llamadas por un único programa que solapaban con otra alteración detectada por múltiples programas (*cluster*)

El archivo *gold-standard* se encuentra en formato VCF (del inglés, *Variant call format*) e incluye información sobre la localización y el tipo de SV detectada, su genotipo, muestras en las que se identificó y las herramientas bioinformáticas que la detectaron.

En un primer análisis descriptivo del *dataset,* se ha determinado que contiene 44505 SVs distribuidas en deleciones, duplicaciones, inserciones e inversiones (Figura 4). Los eventos de tipo traslocación, en los que se produce una transferencia de material genético entre dos cromosomas, no están contemplados en este fichero y no serán objeto de evaluación en este trabajo.

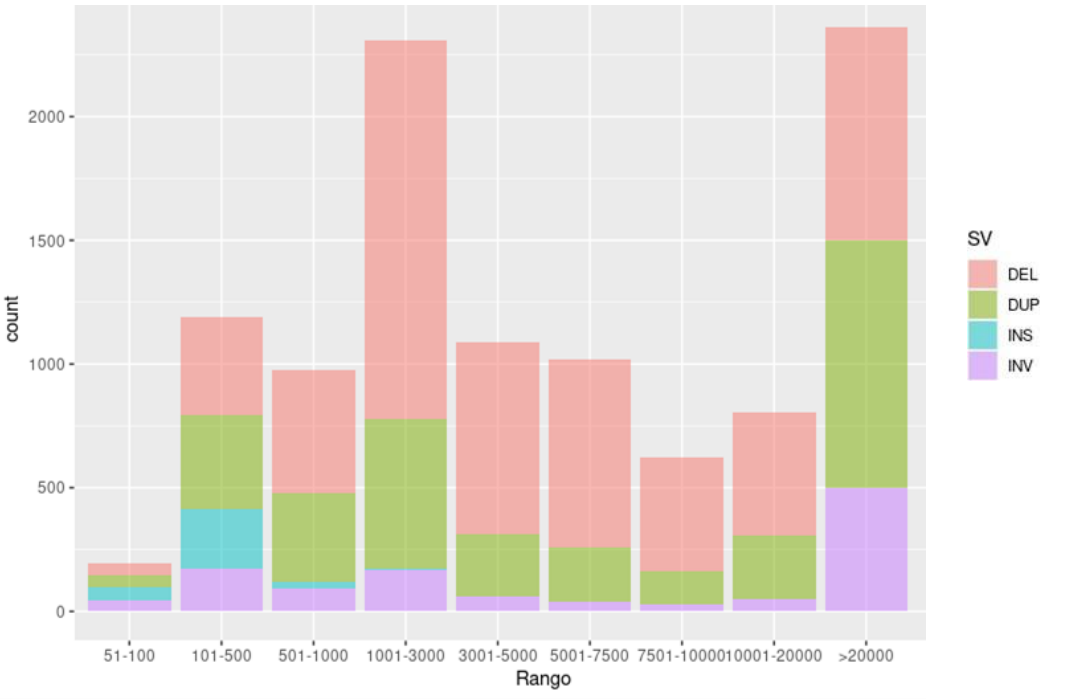


Figura 4: Distribución del tipo y número de SVs presentes en el *dataset* de referencia en función de su tamaño.DEL: deleción, DUP: duplicación, INS: inserción, INV: inversión.

No existe una representación homogénea de las diferentes SVs en el *dataset*, observándose una mayor prevalencia de deleciones frente al resto de alteraciones y una escasa prevalencia de inserciones. Asimismo, la distribución de tamaños no es uniforme, destacando fundamentalmente la ausencia de inserciones de gran tamaño. Teniendo en consideración este hecho y acorde con lo descrito en la literatura [52], se plantea la implicación del tipo y tamaño del evento en el rendimiento de las herramientas bioinformáticas utilizadas para su detección, y se tendrá en cuenta durante la evaluación.

Para ello y en base a la distribución de tamaños observada, se dividirán en tres grupos de trabajo independientes las deleciones, duplicaciones e inversiones: *(i)* SVs de tamaño comprendido entre 50-3000pb, *(ii)* SVs de tamaño comprendido entre 3001-7500 pb y *(iii)* SVs de tamaño >7500pb, y en dos grupos independientes las inserciones: *(i)* inserciones con un tamaño <200pb e *(ii)* inserciones con un tamaño ≥200 pb.

Selección de herramientas

En 2019 Cameron D.L *et al* [2] publicaron una evaluación de las principales herramientas bioinformáticas disponibles en la detección de SVs en genomas generados a partir de *short-read sequencing.* Se ha tomado como punto de partida las herramientas caracterizadas en dicha revisión, así como las incluidas en el *dataset* de referencia para realizar la búsqueda de las herramientas candidatas a evaluar en este trabajo.

En primera instancia se han seleccionado las herramientas con un uso más extendido en la comunidad científica. Para ello, se ha realizado una búsqueda del número de citaciones de cada herramienta en la base de datos *Web of Science*, siendo las más citadas: Pindel (1145 citaciones), BreakDancer (903 citaciones), Delly (770 citaciones), Lumpy (493 citaciones) y Manta (390 citaciones). Se ha descrito en la literatura un mayor rendimiento en términos de precisión de las herramientas basadas en métodos de detección combinados [53], por lo que se han excluido los detectores BreakDancer y Pindel basados en detección de cambios en la profundidad de lectura (*read depth*) y en la detección de cambios en la orientación, orden o longitud del inserto de lecturas pareadas (*paired-read*), respectivamente.

Con ello, las herramientas seleccionadas son Lumpy, Delly y Manta. Las dos primeras presentan un algoritmo basado en la integración de la información procedente de lecturas *paired-end* y *split-reads*, mientras que Manta incorpora además un ensamblaje local *de novo*. Las tres se encuentran disponibles en el repositorio público github, desde donde se han descargado e instalado.

2.2 Análisis bioinformático

Una vez en disposición de las herramientas bioinformáticas a evaluar, los ficheros de alineamiento de los genomas seleccionados y el fichero *gold-standard*, se diseñó el flujo de trabajo a seguir. El objetivo final es la realización de un análisis comparativo del número de SVs detectadas por las herramientas Manta, Delly y Lumpy frente al dataset de referencia, evaluándose la influencia del tipo de SV, su tamaño y la calidad de la llamada en el rendimiento obtenido. Para ello, se plantea un *pipeline* de trabajo dividido en dos bloques.

Procesamiento del *gold-standard*

Como se comentó previamente, el *dataset* de referencia es un archivo global que recoge las SVs identificadas en el conjunto de las nueve muestras, incluyendo información sobre la localización y el tipo de SV detectada, su genotipo, muestras en las que se identificó y las herramientas bioinformáticas que la detectaron.

En base a esta información, se descompuso el *dataset* original en 9 ficheros con las SVs identificadas en cada una de las muestras. De cada uno de ellos, se generaron ficheros individuales en función del tipo, calidad y tamaño de la alteración aplicando los puntos de corte previamente establecidos: deleciones, duplicaciones e inversiones: *(i)* SVs de tamaño comprendido entre 50-3000pb, *(ii)* SVs de tamaño comprendido entre 3001-7500 pb y *(iii)* SVs de tamaño >7500pb; inserciones: *(i)* tamaño <200pb y *(ii)* tamaño ≥200 pb. Los ficheros resultantes son ficheros VCF. La evaluación de los detectores se llevó a cabo mediante la herramienta *bedtools* [54], que utiliza como *input* archivos codificados en formato BED (Figura 5). Como consecuencia, en última instancia se transformaron los ficheros VCF en los ficheros BED finales.

#CHROM	POS	END
chr1	820573	820888
chr1	831213	833732
chr1	853415	853807
chr1	934003	934879
chr1	934331	934879
chr1	934331	934931
chr1	1017781	1017939
chr1	1207049	1207760
chr1	1350034	1351399
chr1	1350047	1351015
chr1	1366934	1367404
chr1	1477838	1478002
chr1	1530529	1530850

Figura 5: Extracto del fichero de deleciones con un tamaño entre 50- 3000pb de la muestra HG00512 (formato BED). Consta de tres columnas separadas por tabuladores: CHROM: cromosoma, POS: coordenada de inicio referenciada al cromosoma, END: coordenada final referenciada al cromosoma.

Otro punto importante de la evaluación es la valoración del *dataset* de referencia. Las SVs incluidas en el *dataset* de referencia no están validadas experimentalmente, de manera que cabe la posibilidad de que no todas las llamadas sean reales y ésto repercuta en el rendimiento de los detectores. Con el objetivo de valorar este escenario, se pretende evaluar el impacto de trabajar con un *dataset* con las SV de mayor confianza en el rendimiento de las herramientas. Para ello, se filtra el *dataset* de referencia en base al número de herramientas bioinformáticas que detectan cada alteración (criterio de selección: nº herramientas ≥3), realizándose de forma paralela el mismo flujo de trabajo que con el *dataset* original (Figura 6).

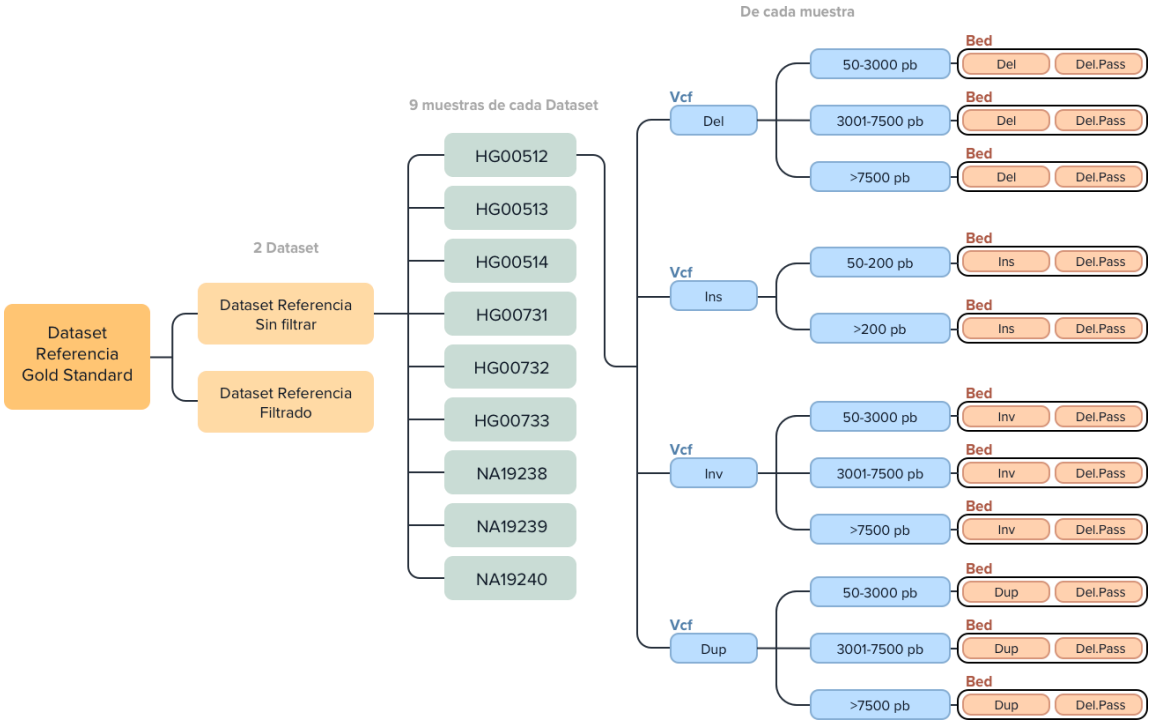


Figura 6: *Pipeline* de procesamiento del *dataset* de referencia.

Tomando como punto de partida el *dataset gold standard* (i) se descompone el fichero original en 9 archivos VCF individuales con el conjunto de SVs identificadas en cada una de las muestras. De forma secuencial, dichos ficheros se desglosan según el tipo de SVs (deleciones, inserciones, inversiones y duplicaciones), y su tamaño. De Los ficheros VCF resultantes se obtiene una copia del fichero en formato BED y un segundofichero filtrado por la calidad de la llamada en formato BED (ii) se filtra el fichero original, obteniéndose un dataset de referencia filtrado con las SV de mayor confianza, que se procesa de forma paralela el *dataset* de referencia original.

Procesamiento de las muestras

Los archivos de partida son los ficheros de alineamiento de los nueve genomas seleccionados y la secuencia del genoma de referencia, previamente indexados. Tras su procesamiento por las diferentes herramientas bioinformáticas se obtuvieron los VCFs iniciales de trabajo, que recogen las alteraciones identificadas en cada muestra, incluyendo información sobre el tipo de SV, localización, coordenada de inicio y fin, calidad de la llamada y precisión en la estimación del punto de ruptura de la alteración.

A partir de ahí, el *pipeline* de trabajo es similar al del procesamiento del *dataset* de referencia (Figura 7). Para cada una de las herramientas y de cada una de las muestras, se generaron archivos individuales desglosados en función del tipo, calidad, precisión en la estimación del punto de ruptura y tamaño de la alteración, aplicando los puntos de corte previamente establecidos: deleciones, duplicaciones e inversiones: *(i)* SVs de tamaño comprendido entre 50-3000pb, *(ii)* SVs de tamaño comprendido entre 3001-7500 pb y *(iii)* SVs de tamaño >7500pb; inserciones: *(i)* tamaño <200pb y *(ii)* tamaño ≥ 200 pb. Los ficheros resultantes son ficheros VCF que se codificaron en formato BED.

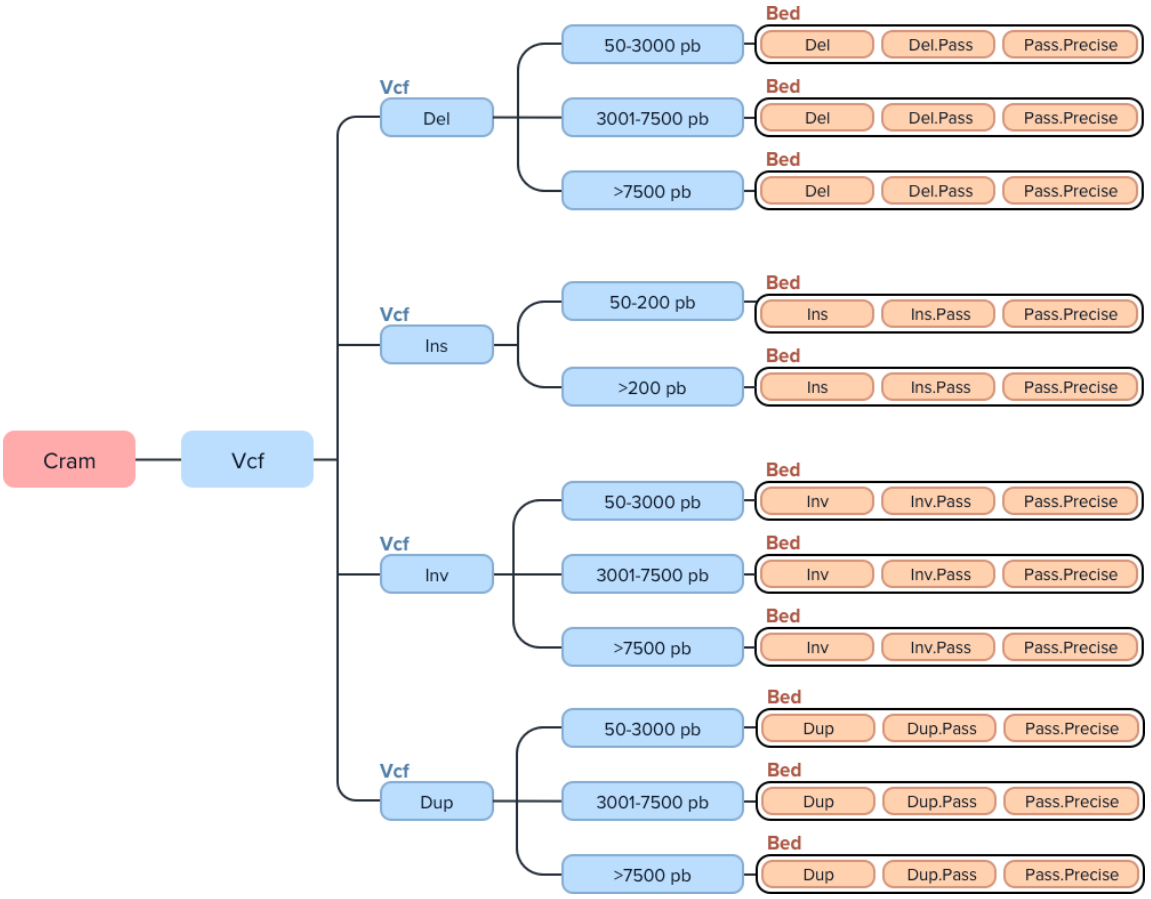


Figura 7: *Pipeline* de procesamiento de los ficheros de trabajo, común para Manta, Delly y Lumpy Se procesa el fichero de alineamiento de cada muestra (CRAM) para obtener los VCFs iniciales de trabajo. De forma secuencial, dichos ficheros se desglosan según el tipo de SVs (Del: deleción, Ins: inserción, Inv: inversión, Dup: duplicación) y su tamaño. De los ficheros VCF resultantes se obtienen los ficheros finales de trabajo: (i) una copia del fichero en formato BED (Del, Ins,Inv,Dup), (ii) un fichero filtrado con las SVs que presentan calidad PASS (.Pass) en formato BED y (iii) un fichero filtrado con las SVs que presentan calidad PASS y en las que se conoce el punto de ruptura exacto de la alteración (.Pass.Precise)

A continuación se desglosan las particularidades del procesado de cada una de las herramientas bioinformáticas para la obtención del VCF inicial.

-Delly

El *script* utilizado se encuentra disponible en el siguiente Link. Tras el procesamiento individual de los ficheros de alineamiento se obtuvo un fichero binario BCF, que se transformó en un archivo VCF mediante la herramienta *bcftools* [55]. Con el objetivo de mejorar la comparación de resultados con el *dataset* de referencia, se excluyeron las alteraciones detectadas en *contigs* alternativos y las SV con un tamaño inferior a 50pb.

El VCF generado incluye inserciones, deleciones, duplicaciones, inversiones y traslocaciones. De acuerdo al *pipeline* de trabajo establecido, se procesó el fichero VCF y se obtuvieron los ficheros BED finales desglosados en función del tipo, tamaño, calidad y precisión en la estimación de los puntos de ruptura de la SV. No se trabajaron las traslocaciones.

-Manta

El *script* utilizado se encuentra disponible en el siguiente Link y consta de dos pasos: en primera instancia se definen las condiciones del análisis (creación del archivo runWorkflow.py) y en un segundo paso se ejecuta dicho fichero.

Como resultado, se obtienen 3 ficheros VCF:

-Archivo *candidateSV*: incluye todas las alteraciones detectadas
-Archivo *diploidSV*: incluye todas las alteraciones puntuadas en base a su calidad y genotipadas bajo un modelo diploide
-Archivo *candidateSmallIndels:* incluye inserciones y deleciones <50pb sin puntuar.

Se ha trabajado exclusivamente con el fichero *diploidSV*, del que se excluyeron las alteraciones detectadas en *contigs* alternativos y las SVs con un tamaño inferior a 50pb. El VCF generado incluye inserciones, deleciones, duplicaciones y BNDs (del inglés, *break-ends*), que fueron desglosadas en archivos VCF individuales en función del tipo de evento, de acuerdo al *pipeline* de trabajo establecido.

A diferencia de Delly, Manta recoge las inversiones y traslocaciones en la misma categoría bajo el nombre de BNDs por lo que fue necesario procesar el VCF correspondiente a las BNDs para separarlas: las alteraciones con puntos de ruptura en el mismo cromosoma se catalogaron como inversiones mientras que las alteraciones con puntos de ruptura en cromosomas diferentes fueron catalogados como traslocaciones. Una vez hecha la distinción, se almacenaron en ficheros VCF independientes: inversiones por un lado y traslocaciones por otro lado.

Una vez creado el fichero de inversiones, se tuvo en cuenta que Manta aborda de forma particular este tipo de eventos: por cada inversión genera dos llamadas, una por cada punto de ruptura (Figura 8). En el caso de las inversiones recíprocas, se generan cuatro llamadas. Para corregir este hecho, la herramienta cuenta con un *script* propio denominado *convertInv.py,* que se utilizó para el post-procesamiento de las inversiones.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00512
chr1	1546876	MantaBND:102:0:1:0:0:0:0		A		A	A[chr19:46541274[	136	PASS	SVTYPE=BND;MATEID=MantaBND:102:0:1:0:0:0
chr19	46541274	MantaBND:102:0:1:0:0:0:1		A		A	]chr1:1546876]A	136	PASS	SVTYPE=BND;MATEID=MantaBND:102:0:1:0:0:0

Figura 8. Ejemplo de inversión detectada por Manta en la muestra HG00512. Manta realiza dos llamadas independientes para cada uno de los puntos de ruptura la inversión, que comparten el mismo identificador MATEID.

Finalmente, y de acuerdo al *pipeline* de trabajo establecido, se procesaron los ficheros VCF de cada tipo de SV y se obtuvieron los ficheros BED finales desglosados en función del tipo, tamaño, calidad y precisión en la estimación de los puntos de ruptura de la SV. No se trabajaron las traslocaciones.

-Lumpy

Desde el repositorio de descarga de Lumpy se recomienda utilizar el *software* Smoove (https://github.com/brentp/smoove) para mejorar el desempeño de la herramienta. Se trata de un programa externo que simplifica y disminuye los tiempos de procesado de otros detectores de SVs que parten de tecnología de secuenciación de lecturas cortas. Por ello, se descargó el programa Smoove a través de github y se utilizó como herramienta de trabajo.

El algoritmo de trabajo de Lumpy está diseñado para realizar un análisis de SVs de múltiples muestras de forma simultánea, integrando la información de cada muestra con el objetivo de resolver las alteraciones con una mayor precisión[43,56].En base a ello, y a diferencia de Manta y Delly, se realizó el procesamiento conjunto de los nueve ficheros de alineamiento.

El *script* utilizado se encuentra disponible en el siguiente enlace Link. Smoove recomienda trabajar excluyendo regiones problemáticas para facilitar el procesado de los datos. De acuerdo a sus recomendaciones, se descargó el fichero de exclusión proporcionado por Smoove (que incluye el *loci* HLA y los *contigs* alternativos, entre otras regiones). Como resultado del procesamiento, se obtuvo un fichero VCF global que recoge las SVs identificadas en el conjunto de las nueve muestras, pre filtradas en base a su calidad (todas las llamadas tienen calidad PASS). Se generaron archivos individuales para cada una de las muestras, que constituyen los archivos VCF de trabajo inicial y en los que se incluyen deleciones, duplicaciones, inversiones y BNDs.

De acuerdo al *pipeline* de trabajo establecido, se procesó el fichero VCF inicial y se obtuvieron los ficheros BED finales desglosados en función del tipo, tamaño, calidad y precisión en la estimación de los puntos de ruptura de la SV. No se trabajaron las traslocaciones.

Cabe reseñar que el *pipeline* de trabajo de las muestras procesadas por Lumpy presenta particularidades. Por un lado, al no detectar eventos del tipo inserción, no se ha realizado el procesamiento correspondiente a esta rama de alteraciones. Por otro lado,dado que todas las SVs detectadas por Lumpy presentan calidad PASS, no existe diferencia entre el fichero original y el fichero PASS generado. Sin embargo, se realiza este paso con el objetivo de mantener la misma estructura para las tres herramientas yfacilitar el trabajo.

2.3 Preparación de los resultados

En resumen, como resultado del análisis bioinformático se han generado:

-Del *gold standard:* para cada muestra *(i)* el conjunto de SVs detectadas y *(ii)* las SVs de mayor confianza, desglosados en función del tipo de alteración, tamaño y calidad de la llamada.

-De cada muestra: *(i)* el conjunto de SVs detectadas por Delly, *(ii)* el conjunto de SVs detectadas por Manta y *(iii)* el conjunto de SVs detectadas por Lumpy, desglosados en función del tipo de alteración, tamaño, calidad de la llamada y precisión en la estimación de su punto de ruptura.

Todos los archivos se encuentran codificados en formato BED. Se realizó un análisis comparativo del número de SVs detectadas por cada una de las herramientas frente al *dataset* de referencia original y el de mayor confianza, evaluándose la influencia del tipo de SV, tamaño, calidad de la llamada y precisión en el rendimiento obtenido. Así mismo, se valoró la implicación del uso de los detectores en combinación, evaluándose de forma independiente las SVs comunes entre parejas de detectores y entre los tres en conjunto frente al *dataset* de referencia. El *script* utilizado se encuentra disponible en el siguiente enlace Link.

La comparación se ha llevado a cabo mediante la herramienta *bedtools* y la opción *intersect*, que permite comparar el *output* de cada herramienta frente al *dataset* de referencia teniendo en cuenta un grado de solapamiento mínimo. Se ha evaluado el rendimiento de los detectores para cada uno de los escenarios en términos de sensibilidad y valor predictivo positivo (VPP), asumiendo grados de solapamientos superiores al 50% (0.5), 60% (0.6), 70% (0.7), 80% (0.8) y 90% (0.9). Los resultados se integraron en un archivo global de rendimiento en formato txt (Tabla 2, disponible en su totalidad en el anexo del documento).

Muestra	Herramienta	Gold.Standard	SV	Size	Calidad	Solapamiento	TP	FP	FN	Sensibilidad	VPP
HG00512	Manta	results	DEL	50-3000	Todas	0.5	3716	857	5728	39	81
HG00512	Manta	results	DEL	50-3000	PASS	0.5	3302	701	4779	40	82
HG00512	Manta	results	DEL	50-3000	PASS.PRECISE	0.5	3248	689	4848	40	82
HG00512	Manta	results	DEL	50-3000	Todas	0.6	3659	914	5878	38	80
HG00512	Manta	results	DEL	50-3000	PASS	0.6	3249	754	4839	40	81
HG00512	Manta	results	DEL	50-3000	PASS.PRECISE	0.6	3195	742	4908	39	81
HG00512	Manta	results	DEL	50-3000	Todas	0.7	3581	992	6083	37	78
HG00512	Manta	results	DEL	50-3000	PASS	0.7	3179	824	4919	39	79
HG00512	Manta	results	DEL	50-3000	PASS.PRECISE	0.7	3128	809	4986	38	79
HG00512	Manta	results	DEL	50-3000	Todas	0.8	3486	1087	6320	35	76
HG00512	Manta	results	DEL	50-3000	PASS	0.8	3097	906	5017	38	77
HG00512	Manta	results	DEL	50-3000	PASS.PRECISE	0.8	3048	889	5082	37	77
HG00512	Manta	results	DEL	50-3000	Todas	0.9	3279	1294	6715	32	71
HG00512	Manta	results	DEL	50-3000	PASS	0.9	2921	1082	5211	35	72
HG00512	Manta	results	DEL	50-3000	PASS.PRECISE	0.9	2885	1052	5257	35	73
HG00512	Manta	results_highConf	DEL	50-3000	Todas	0.5	2387	2186	516	82	52
HG00512	Manta	results_highConf	DEL	50-3000	PASS	0.5	2259	1744	1029	68	56
HG00512	Manta	results_highConf	DEL	50-3000	PASS.PRECISE	0.5	2218	1719	1071	67	56

Tabla 2: Extracto del fichero global de rendimiento. El fichero recoge información de los resultados de la comparativa de cada muestra frente al *gold-standard* escogido (results: *gold-standard* original sin filtrar; results_highConf: *gold-standard* con las SVs de mayor confianza), en función de la herramienta utilizada (Manta, Delly y Lumpy), SV (DEL, INS, INV,DUP), tamaño, calidad y grado de solapamiento. Como resultado se obtiene el número de verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) a partir de los cuales se calculó la sensibilidad y VPP de la herramienta. Los valores de verdaderos negativos (VN) no se recogen en la tabla para no dificultar su lectura.

Los resultados obtenidos se han evaluado mediante la herramienta RStudio [57]. Se ha desarrollado un *script* de trabajo con una función que permite evaluar el rendimiento medio obtenido según el tipo SV para cada una de las herramientas en base al grado de solapamiento. La función es dependiente del fichero de *dataset* de referencia, el tipo y tamaño de SV y la calidad de la llamada. El *script* se encuentra disponible en enlace Link.

2.4 Resultados

2.4.1 Evaluación de las herramientas según tipo de SV

Sin restricciones

Por un lado se evaluó el rendimiento de Manta, Delly, Lumpy y de las posibles combinaciones de detectores frente al *dataset* de referencia sin aplicar ninguna restricción (Figura 9). Las tres herramientas presentaron un rendimiento limitado en la detección de SVs, con una sensibilidad inferior al 50% en términos generales que disminuye a medida que el grado de solapamiento requerido se hace más restrictivo. Este hecho es especialmente significativo en el caso de las deleciones, donde el rendimiento del detector oscila hasta en un 15% en términos de sensibilidad.

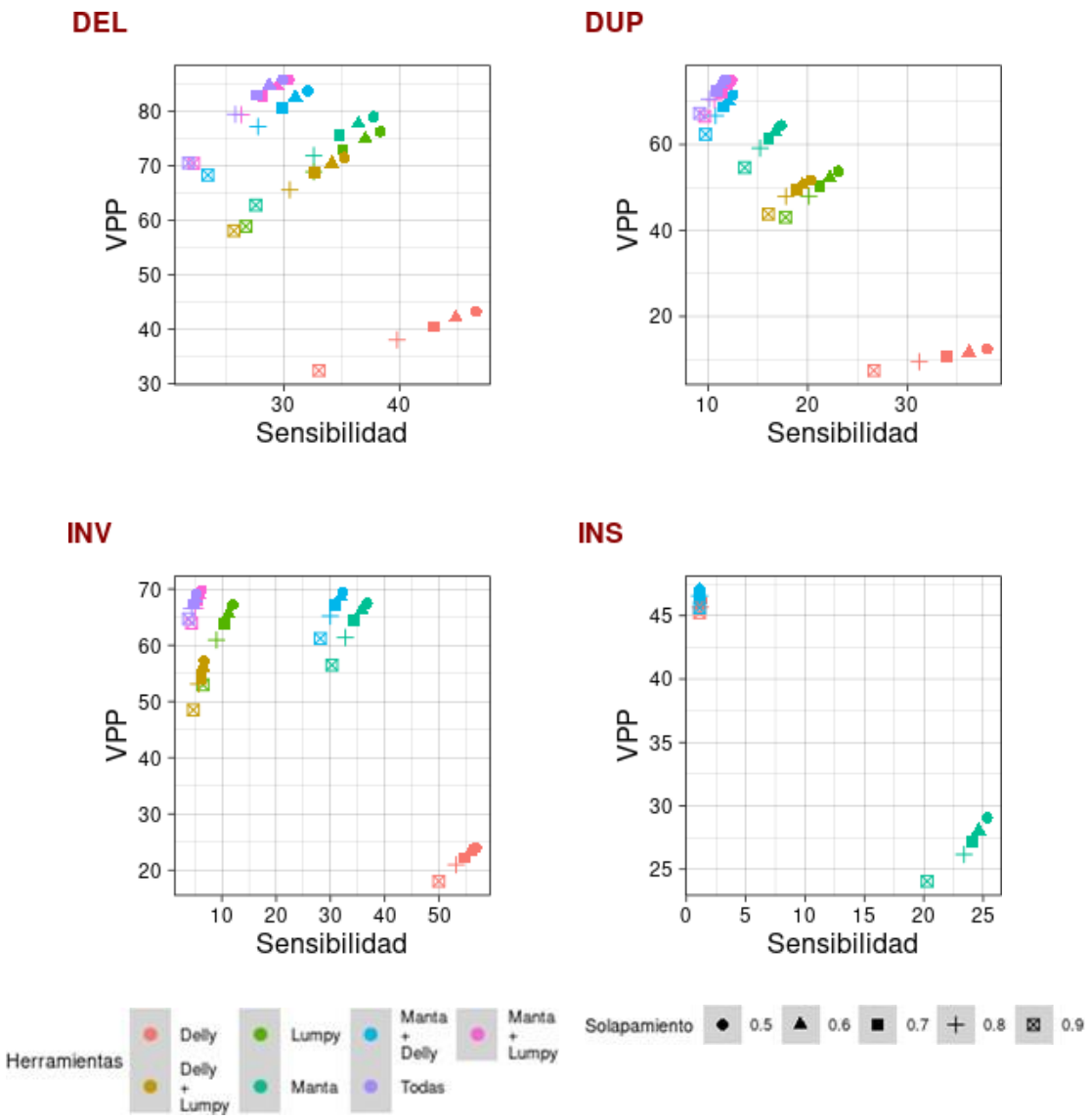


Figura 9: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones frente al *Dataset* de referencia original, en función del grado de solapamiento.

Asimismo, se valoró el impacto de la unión de detectores con el objetivo de obtener llamadas de mayor confianza, evaluándose el rendimiento conjunto de Manta y Delly, Manta y Lumpy, Delly y Lumpy y de las tres herramientas. Se observa una mejora de la precisión a costa de una disminución de la sensibilidad.

Restringido por la calidad de llamada

Por otro lado se evaluó el impacto de la calidad de la llamada de la SVs en el rendimiento de las herramientas, comparándose los resultados de las llamadas realizadas con un filtro de calidad PASS frente a las llamadas con calidad PASS del *gold standard* (Figura 10).

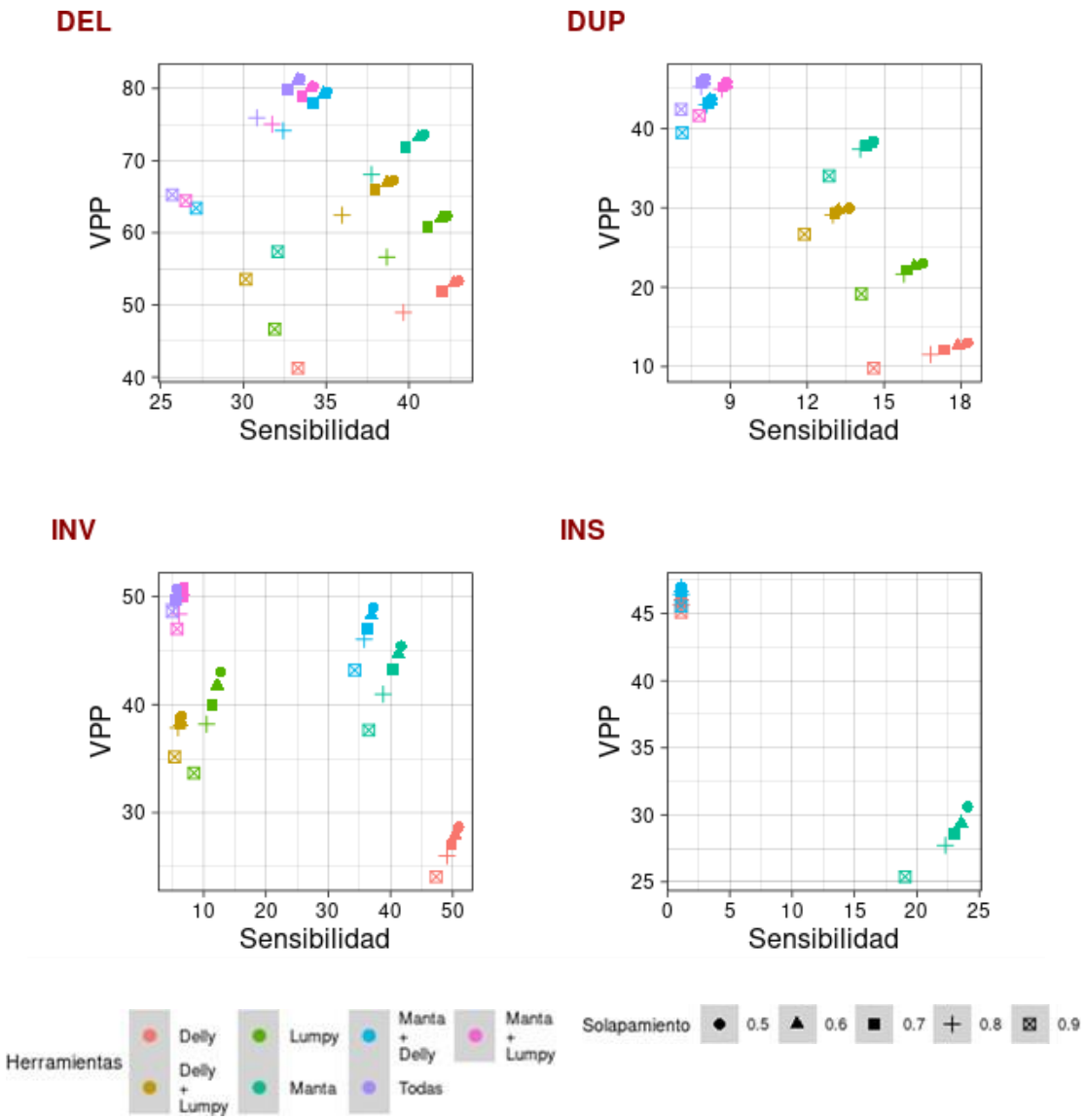


Figura 10: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones con calidad PASS frente al Dataset de referencia con las SVs catalogadas con calidad PASS, en función del grado de solapamiento.

Restringido por la calidad y precisión en la estimación de los puntos de ruptura

Se evaluó el rendimiento de las herramientas teniendo en cuenta el grado de precisión de los puntos de ruptura de las SVs (Figura 11). Se observa una ligera disminución de la sensibilidad y un aumento del valor predictivo positivo respecto a la aproximación anterior, especialmente significativo al trabajar con grados de solapamientos menos restrictivos.

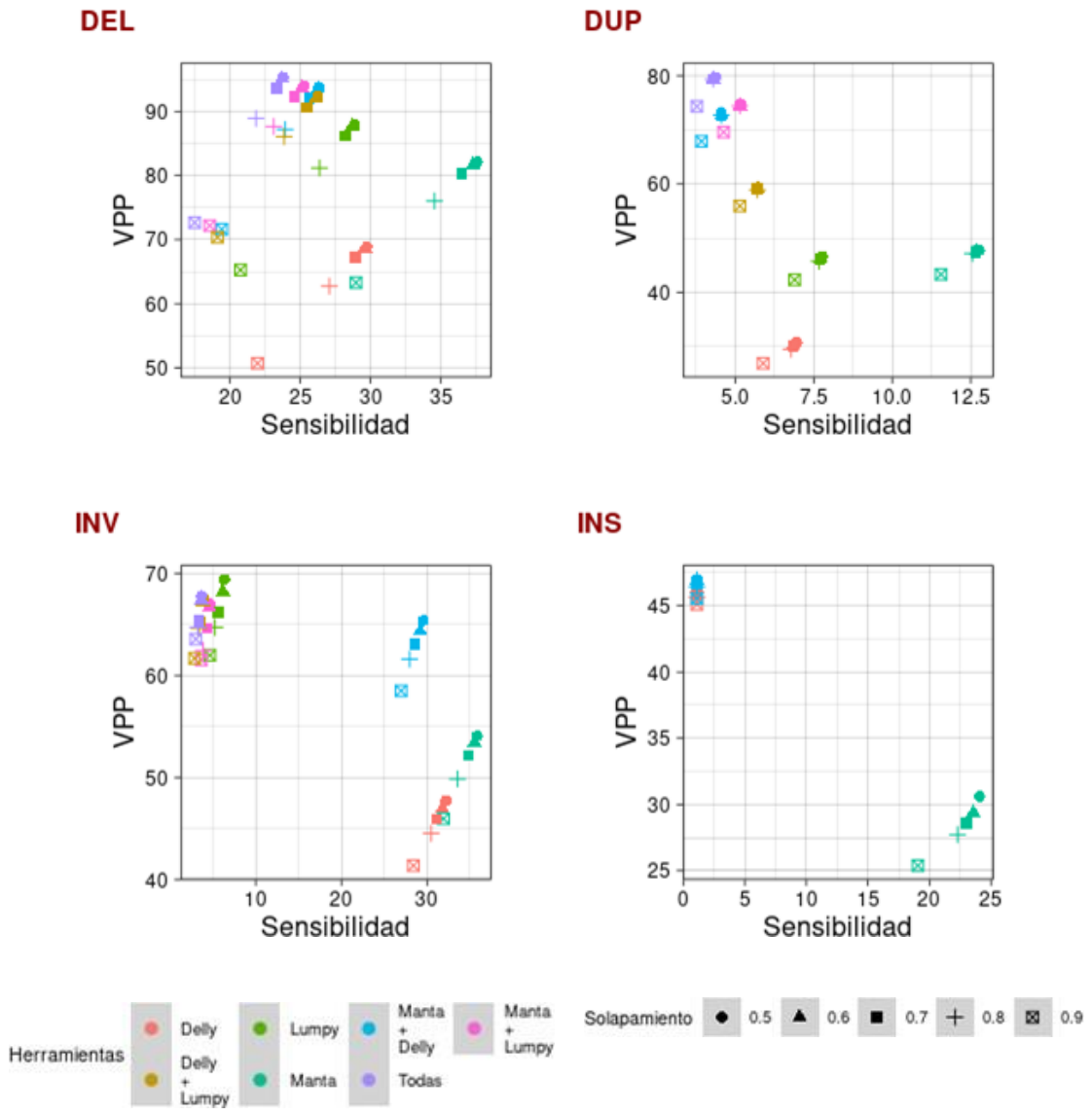


Figura 11: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones con calidad PASS y precisión PRECISE frente al *Dataset* de referencia, en función del grado de solapamiento.

Uso del *gold standard* filtrado por las SVs de mayor confianza

Por último se ha evaluado el rendimiento de Manta, Delly, Lumpy y de las posibles combinaciones de detectores frente al segundo *dataset* de referencia constituido por las SVs de mayor confianza, que han sido detectadas por al menos 3 herramientas bioinformáticas (Figura 12).

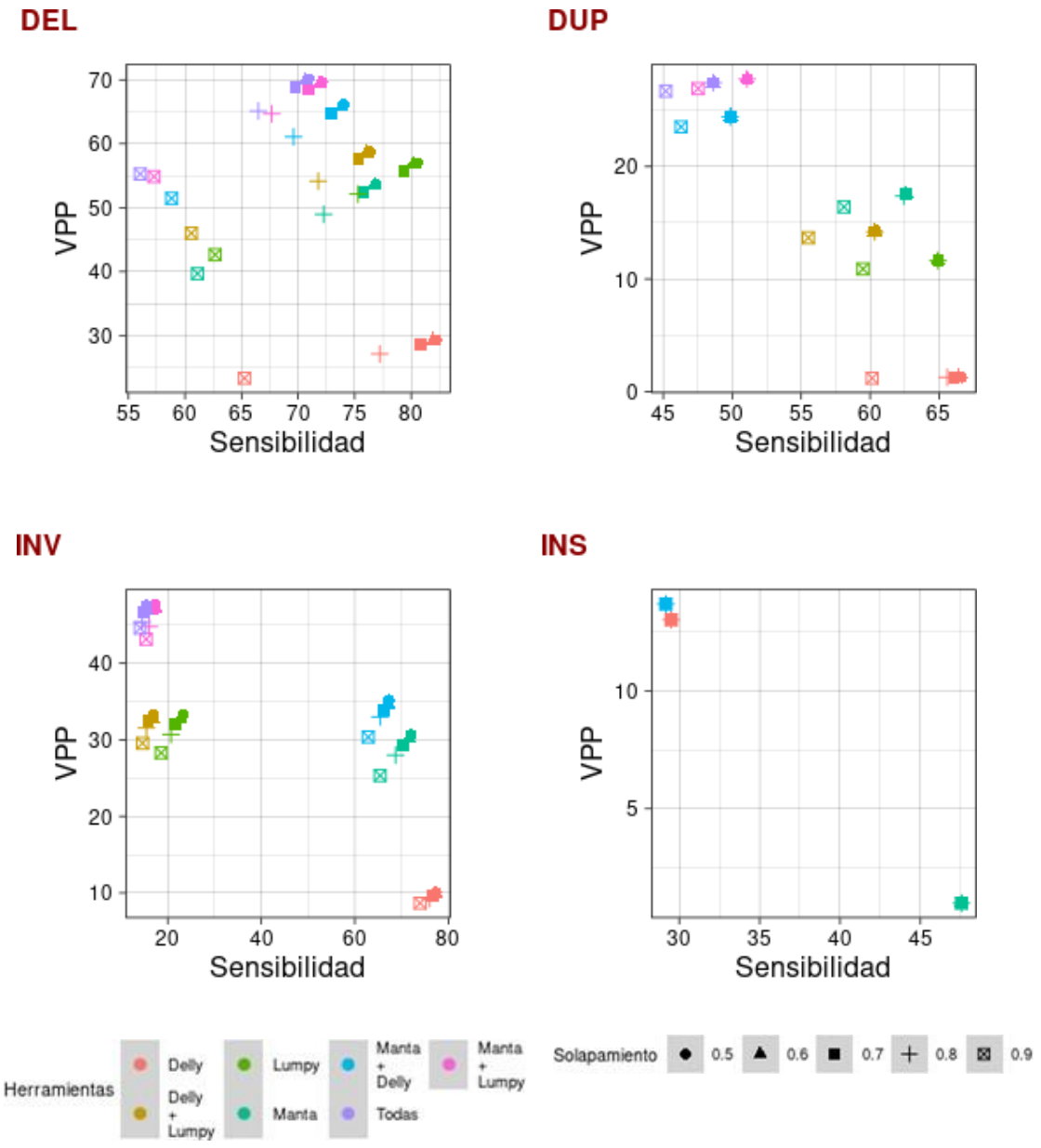


Figura 12: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones frente al *Dataset* de SVs de mayor confianza, en función del grado de solapamiento.

2.4.2 Evaluación de las herramientas según tipo y tamaño de SV

Sin restricciones

Otro punto que se ha querido evaluar es el impacto del tamaño de las SVs en el rendimiento de los detectores. Por un lado se evaluó el rendimiento de Manta, Delly, Lumpy y de las posibles combinaciones de detectores frente al *dataset* de referencia sin aplicar ninguna restricción (Figura 13).

Las tres herramientas presentan un rendimiento limitado en la detección de las SVs de interés independientemente del tamaño de la SV. En relación con el tamaño del evento, en general se observa una disminución de la precisión a medida que aumenta el tamaño la SV, especialmente significativo en la detección de eventos del tipo inserción, donde el VPP oscila de un 40% a un 10% para la herramienta Manta en grados de solapamiento superiores al 90%.

Restringido por la calidad de llamada

Por otro lado se evaluó el impacto de la calidad de la llamada de la SVs en el rendimiento de las herramientas. Para ello se compararon los resultados de las llamadas realizadas con un filtro de calidad PASS frente a las llamadas con calidad PASS del *gold standard* (Figura 14).

Se observa una disminución de la precisión de las herramientas, especialmente notoria en SVs con un tamaño superior a 7500 pb.

Restringido por la calidad y precisión en la estimación de los puntos de ruptura

Se evaluó el rendimiento de las herramientas teniendo en cuenta el grado de precisión de los puntos de ruptura de las SVs. Para ello se compararon los resultados de las llamadas realizadas con un filtro de calidad PASS y puntos de ruptura PRECISE frente a las llamadas con calidad PASS del *gold standard* (Figura 15).

Como consecuencia, se observa una disminución de la sensibilidad de los detectores especialmente significativo en SVs con un tamaño superior a 7500 pb y un aumento de su precisión con respecto a la aproximación anterior.

Uso del *gold standard* filtrado por las SVs de mayor confianza

Se ha evaluado el rendimiento de Manta, Delly, Lumpy y de las posibles combinaciones de detectores frente al segundo *dataset* de referencia constituido por las SVs de mayor confianza, que han sido detectadas por al menos 3 herramientas bioinformáticas (Figura 16).

De forma generalizada, se observa un aumento de la sensibilidad de los detectores respecto al uso del *gold standard* original. Destaca el cambio de comportamiento de Manta en la detección de inserciones, no habiéndose detectado eventos superiores a 200 pb.

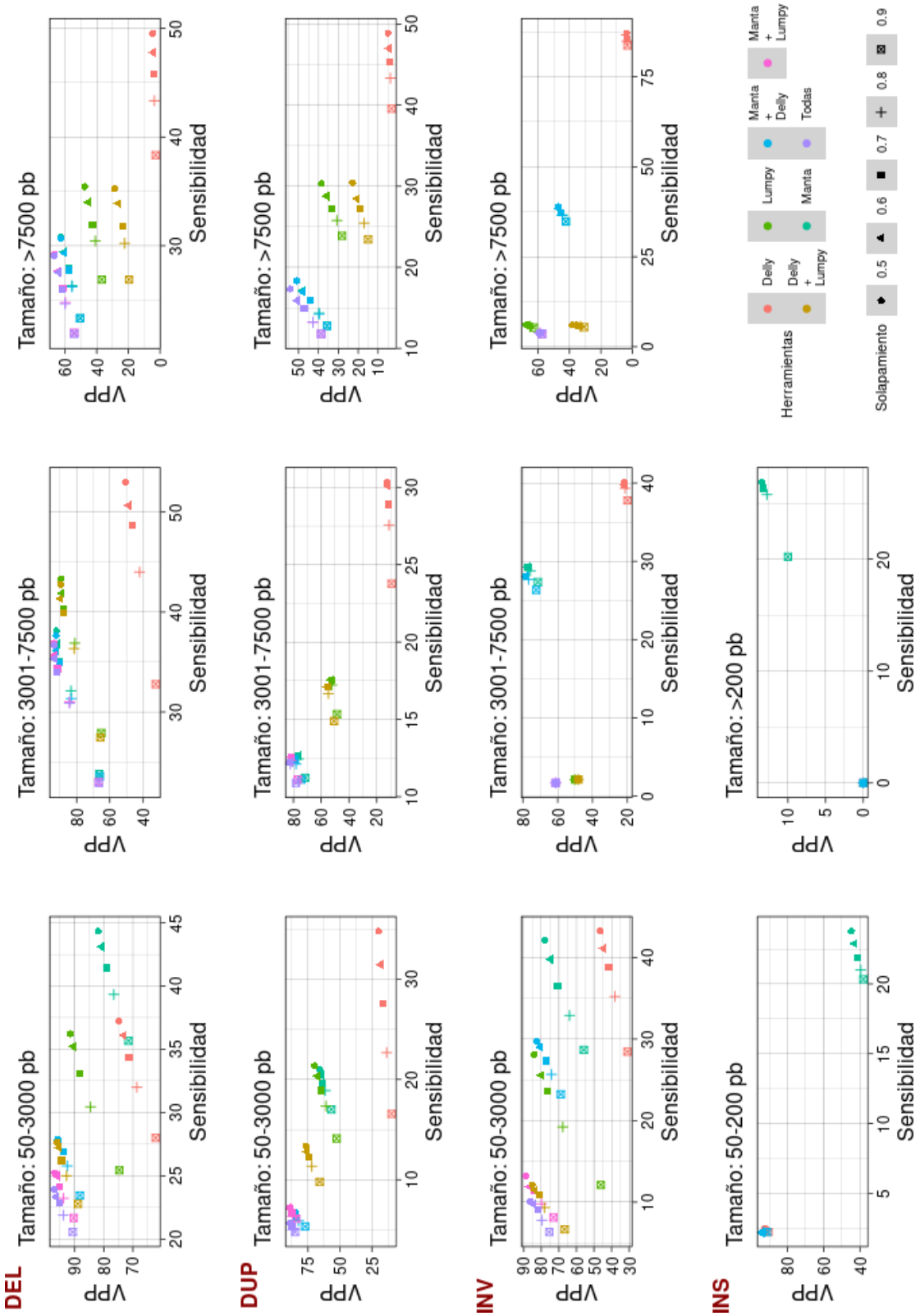


Figura 13: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones frente al *Dataset* de referencia, en función del grado de solapamiento y del tamaño de la SV.

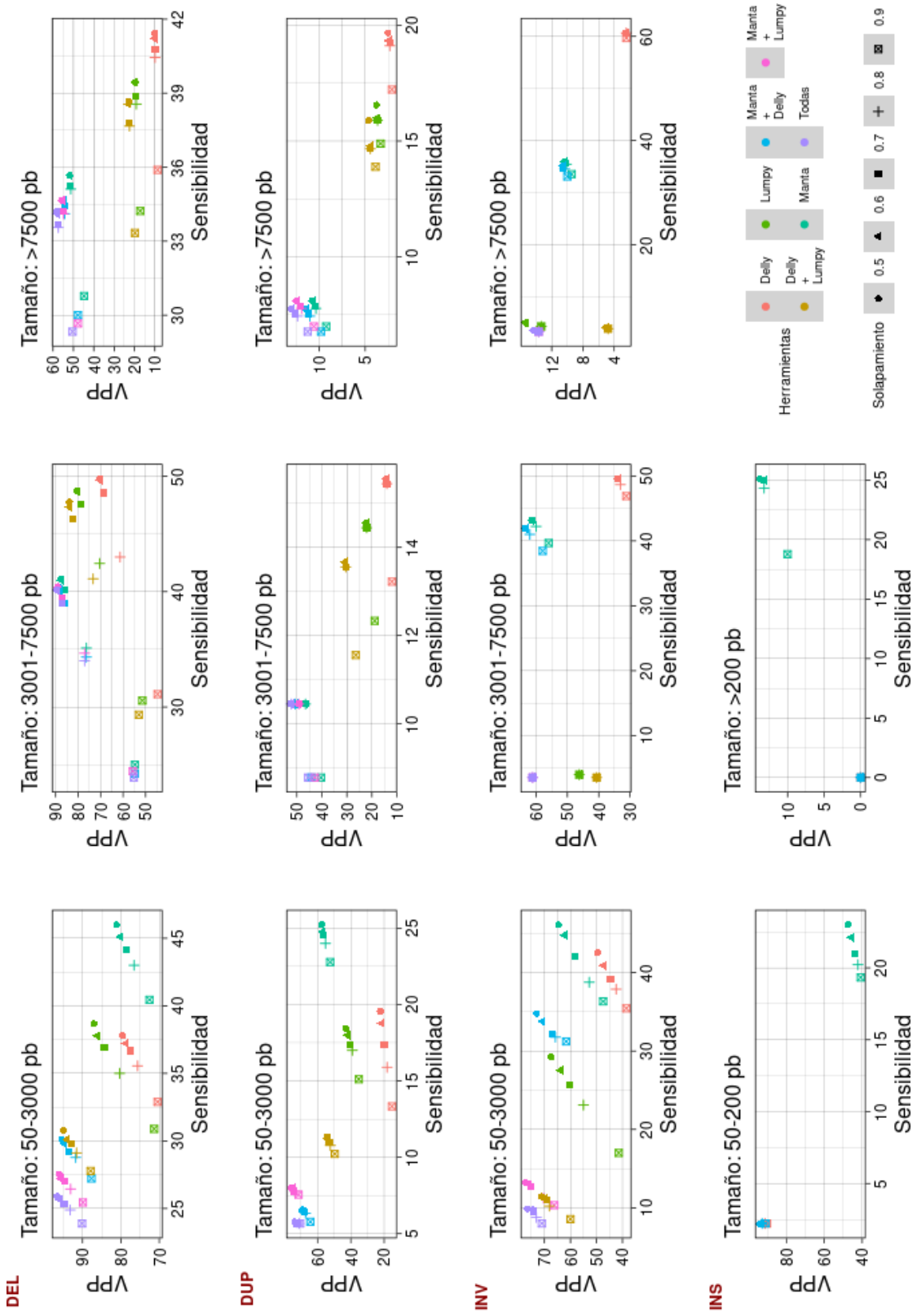


Figura 14: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones con calidad PASS frente al *Dataset* de SVs con calidad PASS en función del grado de solapamiento y del tamaño de la SV.

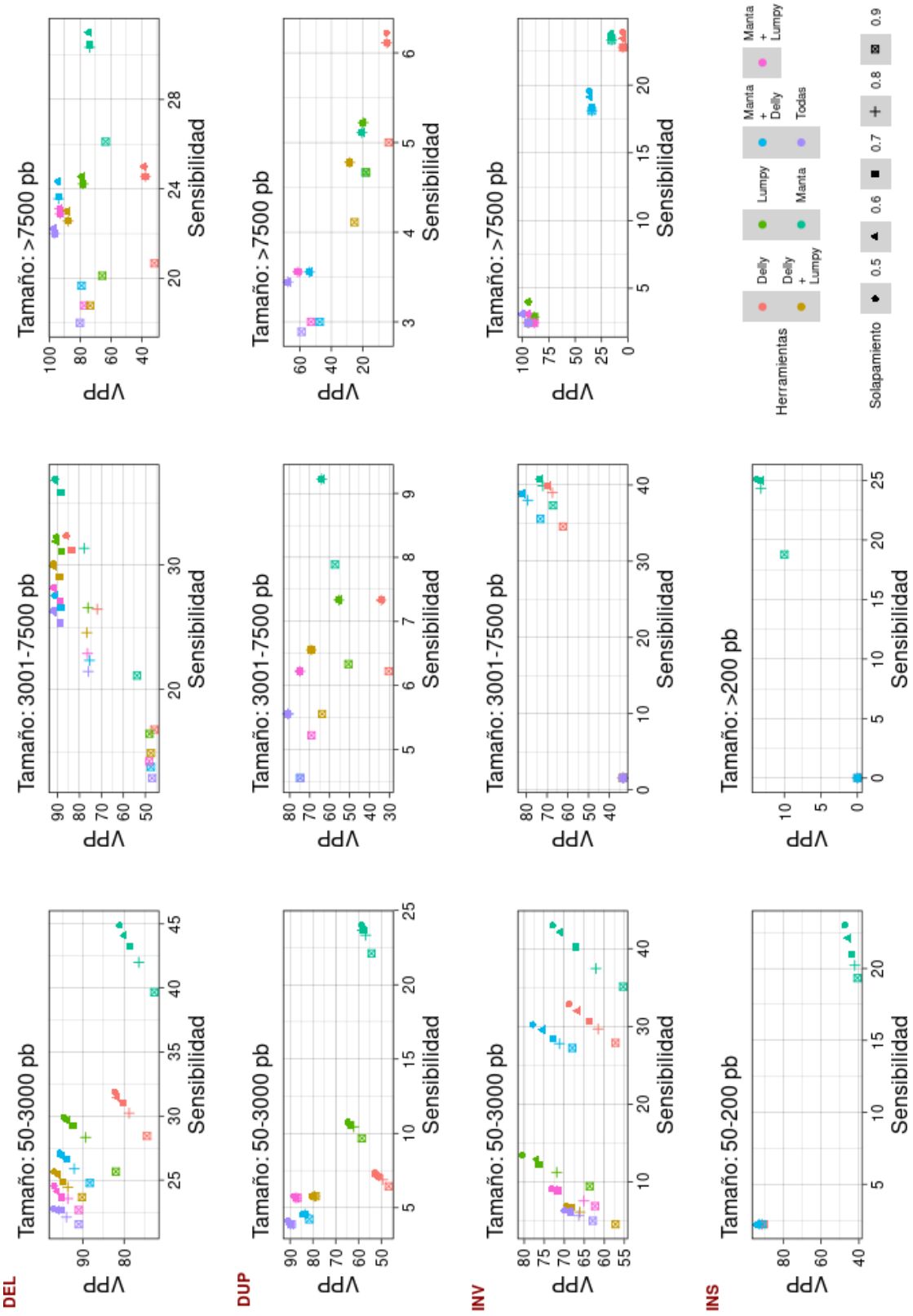


Figura 15: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones con calidad PASS y PRECISE frente al *Dataset* de referencia en función del grado de solapamiento y del tamaño de la SV.

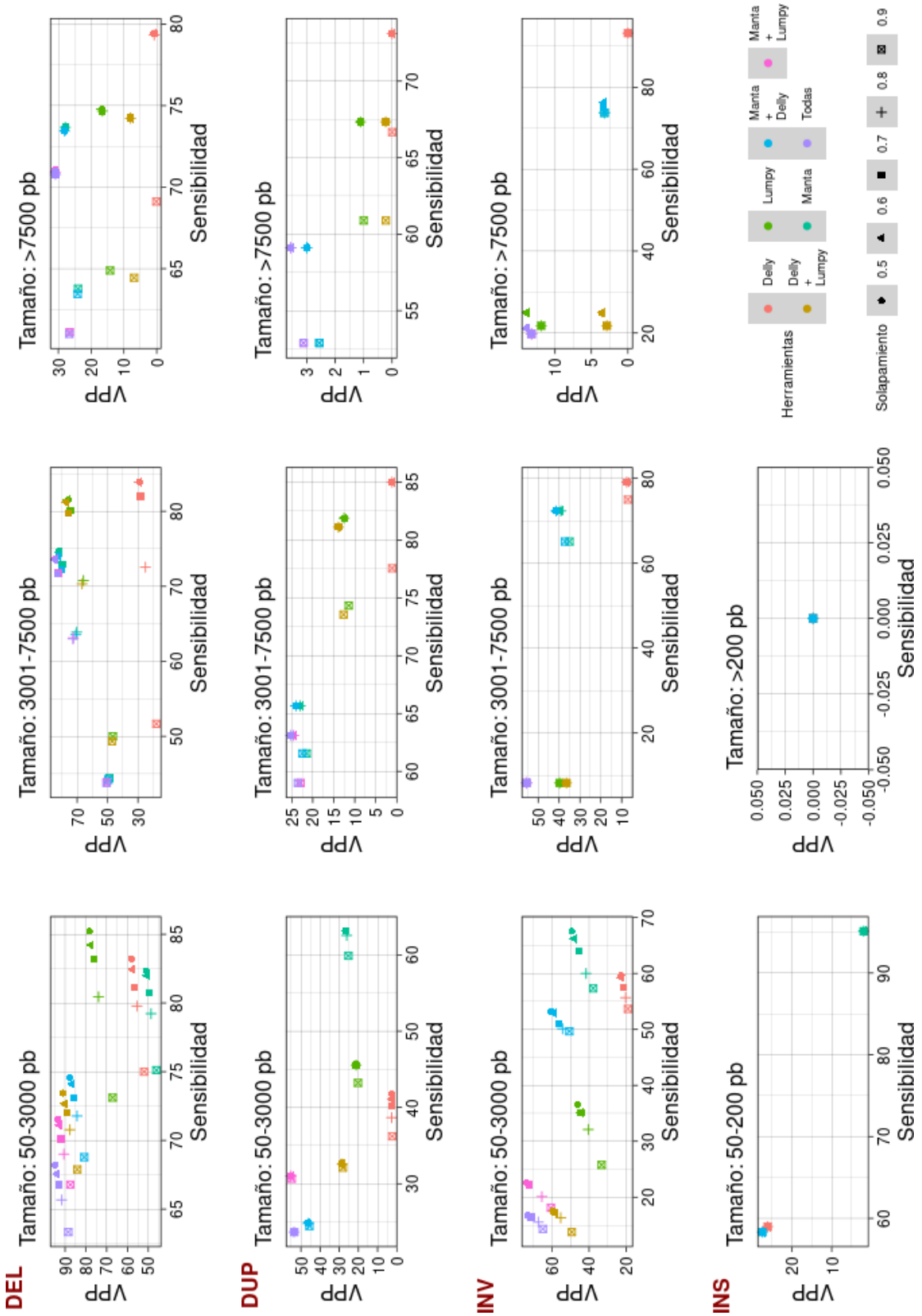


Figura 16: Comparativa del rendimiento medio de las herramientas Manta, Delly y Lumpy, y de las diferentes combinaciones posibles de la unión de los mismos en la detección de deleciones, inserciones, duplicaciones e inversiones frente al *Dataset* de SVs de mayor confianza en función del grado de solapamiento y del tamaño de la SV.

2.5 Discusión

En este capítulo se ha llevado a cabo una evaluación del rendimiento de tres herramientas bioinformáticas dirigidas a la detección de SVs a partir de datos de WGS. La evaluación surge con el objetivo de valorar el potencial de la WGS en el abordaje de este tipo de alteraciones genéticas, habitualmente estudiadas mediante técnicas alternativas.

2.5.1 Evaluación del rendimiento de las herramientas según el tipo de SV

Para ello, se ha realizado un análisis comparativo del número de SVs detectadas por las herramientas Manta, Delly y Lumpy desglosadas por tipo de SV frente a un *dataset* de referencia en términos de sensibilidad y precisión. Se ha valorado el posible impactodel tamaño de la alteración en el rendimiento del detector, así como la fiabilidad de la llamada de acuerdo al criterio de calidad y de precisión en la estimación de su punto de ruptura.

A raíz de los resultados, el rendimiento de las herramientas es limitado, siendo la sensibilidad inferior al 50% en términos generales. Su desempeño es dependiente del tipo de alteración y de su tamaño, y se ve modulado por la calidad y precisión de las llamadas, como puede observarse en las Figuras 9, 10 y 11. A continuación se comentan los resultados obtenidos desglosados según el tipo de evento.

2.5.1.1 Deleciones

Se observa un rendimiento superior en la detección de deleciones en comparación con el resto de eventos. Todos los detectores presentan una sensibilidad comprendida entre el 25 y el 35% para solapamientos superiores al 90%, con diferencias significativas en términos de precisión. Manta y Lumpy presentan un comportamiento similar con valores de precisión en torno al 60%, mientras que Delly presenta unos valores de sensibilidad superiores a costa de un VPP bajo. En todos los casos, el rendimiento es variable en función del grado de solapamiento escogido. A este respecto, es conveniente señalar que Delly presenta una sensibilidad para solapamientos superiores al 90% comparable con la sensibilidad de Manta y/o Lumpy para solapamientos superiores al 80%, con diferencias significativas en el VPP observado, como puede apreciarse en la Figura 9.

La restricción de las llamadas a aquellas con calidad adecuada o PASS supone un ligero incremento en la sensibilidad mínima observada (todos los algoritmos presentan una sensibilidad superior al 30% para niveles de solapamiento del 90%), que repercute significativamente en la precisión de los detectores (ver Figura 10). Se observa una disminución del VPP de Manta y Lumpy y un aumento del VPP de Delly (41.26%), reduciéndose en consecuencias las diferencias de precisión entre detectores.

Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se observa un aumento de la precisión de las herramientas. Manta y Lumpy presenta VPPs similares a los obtenidos en la primera comparativa, mientras que Delly presenta VPPs superiores en torno al 50%. En términos de sensibilidad, se produce una disminución del rendimiento de las herramientas, especialmente significativo en Delly y Lumpy, con una sensibilidad inferior al 25%, como se observa en la Figura 11.

Desglosadas por tamaño

Al desglosar el **rendimiento por tamaño** se observa la influencia de la longitud de la alteración en el desempeño de las herramientas (ver Figura 13), apreciándose una tendencia de disminución de la precisión a medida que aumenta el tamaño de la deleción. En deleciones con un tamaño inferior a 3000pb la herramienta que mejor rendimiento presenta es Manta, con una sensibilidad en torno al 35% y un VPP del 70%. Sin embargo, su rendimiento disminuye notablemente en deleciones de mayor tamaño, siendo Lumpy el detector de elección con una mejor relación sensibilidad/VPP. Por su parte, Delly presenta una precisión inferior al resto de herramientas en cualquiera de los escenarios, que mejora al utilizarse en combinación con Lumpy. Para concluir, resaltar el escaso impacto de la utilización conjunta de múltiples detectores en términos de precisión en deleciones superiores a 3000pb, que presentan VPP semejantes a los de las herramientas de forma aislada.

La restricción de las llamadas a aquellas con calidad adecuada o **PASS** supone un incremento en la sensibilidad mínima observada para el mismo grado de solapamiento de Manta y Lumpy, independientemente del tamaño de la deleción, como se aprecia en la Figura 14. En términos de precisión, destaca la disminución de VPP de Lumpy, especialmente significativo en deleciones con un tamaño superior a 7500 pb.

Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se aprecia una disminución de la sensibilidad de las herramientas. Se acentúan las diferencias de Manta frente al resto de detectores en deleciones con un tamaño inferior a 3000 pb y/o superiores a 7500 pb, y un incremento de la precisión de los tres detectores en deleciones superiores a 7500pb respecto a la aproximación anterior (ver Figura 15).

2.5.1.2 Duplicaciones

Todos los detectores presentan una sensibilidad inferior al 30% para solapamientos superiores al 90%, con diferencias significativas en términos de precisión. Manta y Lumpy presentan un comportamiento similar en términos de sensibilidad, aunque Manta presenta un VPP un 10% superior. El rendimiento es variable en función del grado de solapamiento escogido, siendo el la sensibilidad observada en Lumpy para grados de solapamientos superiores al 90% comparable al de Manta para un solapamiento más permisivo del 50%. Por su parte, Delly presenta unos valores de sensibilidad superiores al resto de detectores, aunque a costa de un VPP por debajo del 10%, como puede apreciarse en la Figura 9.

La restricción de las llamadas a aquellas con calidad adecuada o **PASS** limita el rendimiento de las herramientas (todos los algoritmos presentan una sensibilidad inferior al 20% para niveles de solapamiento del 90%, ver Figura 10). Se observa una disminución de la precisión de los detectores Manta y Lumpy, sin que se vea afectada su sensibilidad. El descenso del VPP gira en torno a una disminución del 20% para ambos detectores, que presentan valores comprendidos entre el 20 y el 40%, respectivamente. Por su parte, Delly sufre una ligera caída de su precisión y una marcada disminución de su rendimiento en términos de sensibilidad. Para grados de solapamiento superiores al 90% presenta una sensibilidad inferior al 15%, comparable a la observada en la herramienta Lumpy.

Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se acentúan las diferencias entre detectores respecto a la aproximación anterior, como se observa en la Figura 11. Lumpy y Delly presentan una disminución de la sensibilidad por debajo del 7.5%, mientras que se observa un incremento significativo de su precisión, que en el caso de Lumpy alcanza niveles comparables a los observados en el fichero original sin filtrar. Por su parte, Manta disminuye ligeramente su sensibilidad, que se sitúa aun por encima del 10%, y presenta un incremento de su VPP superior al 40%, distinguiéndose del resto de detectores.

Desglosadas por tamaño

Al desglosar el **rendimiento por tamaño** se aprecia una tendencia de disminución de la precisión a medida que aumenta el tamaño de la duplicación, como puede observarse en la Figura 13. En duplicaciones con un tamaño inferior a 3000pb las herramientas Manta y Lumpy presentan un comportamiento semejante, con un nivel de sensibilidad en torno al 15% y una precisión con un VPP que oscila entre el 50-55%, siendo el rendimiento de Manta ligeramente superior en grados de solapamientos superiores al 90%. Delly, por su parte, presenta unos valores de sensibilidad similares con diferencias significativas del VPP. En duplicaciones con un tamaño superior, se produce una diferenciación en términos principalmente de sensibilidad del rendimiento de Manta con respecto al resto de detectores. Mientras Manta presentan valores de sensibilidad inferiores al 20% independientemente del tamaño de la duplicación, Lumpy y Delly incrementan su sensibilidad en duplicaciones superiores a 3000pb, destacando el desempeño de la herramienta Lumpy.

La restricción de las llamadas a aquellas con calidad adecuada o **PASS** supone una disminución del rendimiento de las herramientas, tanto en términos de sensibilidad como de precisión. La excepción es el desempeño de Manta en duplicaciones con un tamaño inferior a 3000pb, que presenta un incremento significativo de su sensibilidad con respecto al *dataset* original sin disminuir su precisión, destacando sobre el resto de detectores (ver Figura 14).

Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se aprecia una disminución de la sensibilidad y un incremento del VPP de las herramientas en todos los tramos de tamaño con respecto a la aproximación anterior (ver Figura 15). Es conveniente resaltar que los tres programas presentan una sensibilidad inferior al 10% en cualquier rango a excepción de Manta que presenta un rendimiento similar al de la aproximación previa, como puede observarse en la Figura 13. Se observa un aumento de las diferencias entre Manta y el resto de detectores en duplicaciones con un tamaño inferior a 3000 pb, y un rendimiento muy pobre de todos los algoritmos en duplicaciones con un tamaño superior a 7500 pb.

2.5.1.3 Inversiones

En la detección de inversiones se observan importantes diferencias de sensibilidad entre herramientas, que varían entre aproximadamente el 10% en Lumpy, el 30% en Manta y el 50% en Delly para solapamientos superiores al 90%, como puede verse en la figura 7 (ver Figura 9). La precisión mínima observada es superior al 15% y corresponde al detector Delly, mientras que la precisión máxima observada corresponde a Manta con un VPP en torno al 55%. Cabe resaltar el comportamiento de Lumpy, con una sensibilidad muy por debajo de Manta para este tipo de alteraciones.

La restricción de las llamadas a aquellas con calidad adecuada o **PASS** limita el rendimiento de las herramientas, principalmente en términos de precisión, como se observa en la Figura 10. Se aprecia una disminución del VPP de los detectores, especialmente significativo en Lumpy y Manta, en los que disminuye en torno a un 20%. Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se incrementa el rendimiento de las herramientas en términos de precisión con respecto a la aproximación anterior. El incremento es especialmente significativo en el detector Lumpy, que pasa de un VPP del 36.71% al 61.96%. Por otro lado, se observa una disminución de la sensibilidad de Delly, con valores con un grado de solapamiento superior a 70% equiparables a los observados en Manta para grados de solapamientos superiores al 90% (ver Figura 11). Con todo, el algoritmo que mejor ratio sensibilidad/VPP presenta es Manta, cuyo rendimiento mejora significativamente al utilizarse en combinación con el detector Delly.

Desglosadas por tamaño

Al desglosar el **rendimiento por tamaño** se observa la influencia de la longitud de la alteración en el desempeño de las herramientas (ver Figura 13). En tamaños inferiores a 3000pb se observa una sensibilidad variable que oscila entre aproximadamente el 12% en Lumpy y el 28% en Delly y Manta para grados de solapamientos del 90% Es reseñable que a igualdad de rendimiento de estos dos últimos, presentan una diferencia en términos de precisión del 25%. En inversiones de mayor tamaño, Lumpy aumenta su precisión sin variar su sensibilidad, mientras que Delly tiende a incrementar su sensibilidad en detrimento de su precisión. Por su parte, Manta presenta un mayor rendimiento en inversiones con un tamaño entre 3001-7500 pb y un rendimiento intermedio entre Delly y Lumpy en alteraciones de un tamaño superior.

La restricción de las llamadas a aquellas con calidad adecuada o **PASS** supone un aumento de la sensibilidad de las herramientas en inversiones menores a 3000 pb y una disminución de su precisión especialmente significativa en inversiones mayores de 7500 pb, en las que se observa un VPP inferior al 20%. No se observan modificaciones en la tendencia de los detectores(ver Figura 14).

Si de las llamadas con calidad PASS se seleccionan únicamente aquellas con un grado de precisión de los puntos de ruptura **PRECISE**, se aprecia un importante aumento de la precisión de las herramientas respecto a la aproximación anterior, como puede observarse en la Figura 15. Por otro lado se observa una disminución en torno al 5% de la sensibilidad de Delly en inversiones inferiores a 3000 pb con respecto a la aproximación anterior, destacando el rendimiento de Manta del resto de detectores (ver Figura 15).

2.5.1.4 Inserciones

Se observa un rendimiento inferior en la detección de inserciones en comparación con el resto de eventos, como se observa en la Figura 9. Sólo Manta presenta una sensibilidad superior al 20% para grados de solapamientos superiores al 90%, con una precisión inferior al 25%. La restricción de las llamadas a aquellas con calidad adecuada o **PASS** supone un ligero de la %, que se mantiene constante independientemente de la precisión de su punto de ruptura (Figura 10 y 11). Por otro lado, Delly presenta una sensibilidad inferior al 2% y una precisión en torno al 45% que no varía con la calidad de la llamada. Lumpy, como se vio previamente, no detecta inserciones.

Desglosadas por tamaño

Al desglosar el **rendimiento por tamaño** se observa la influencia de la longitud de la alteración en el desempeño de las herramientas (Figura 13). Para el detector Manta, la sensibilidad varía en función del tamaño de la inserción, pero sí lo hace su precisión, que oscila entre un 40% en inserciones con un tamaño inferior a 200 pb y un 10% en inserciones con un tamaño superior. Este rendimiento no está en relación con la calidad de la llamada ni con la precisión en la estimación de los puntos de ruptura, manteniéndose constante independientemente de dichos parámetros como se observa en las Figuras 14 y 15, respectivamente. En el caso de Delly, se observa su incapacidad en la detección de inserciones con un tamaño superior a 200 pb, que puede estar relacionada con el tamaño de la lectura generada durante la secuenciación.

2.5.2 Tipo de *dataset* de referencia

Otro punto importante de la evaluación es la valoración del *dataset* de referencia, mediante la evaluación de las herramientas frente a un segundo dataset de referencia filtrado con las SVs de mayor confianza. En líneas generales, su uso repercute en un aumento de la sensibilidad de los detectores y un descenso de su VPP (ver Figura 12 y 16). El incremento de sensibilidad es coherente con el filtrado de todas aquellas alteraciones identificadas por una única herramienta bioinformática diferente a Manta, Delly y/o Lumpy, mientas que la disminución del VPP es consecuente con la exclusión de las alteraciones identificadas exclusivamente por Manta, Delly y/o Lumpy.

2.5.3 Valoración global de las herramientas

Respecto al desempeño de las herramientas, en líneas generales Manta ha demostrado un rendimiento superior en la detección de SVs. Su superioridad es independiente del tipo de alteración y se ve influenciada por la calidad y precisión de la llamada, siendo especialmente notorias las diferencias con el resto de detectores al trabajar con SVs con calidad PASS y puntos de ruptura PRECISE. Con todo, el desempeño observado es limitado, especialmente en la detección de inserciones con un tamaño superior a 200pb y de duplicaciones, y está en línea con lo descrito (https://github.com/Illumina/manta). Lo más destacable es que a diferencia de Delly y Lumpy, Manta es capaz de gestionar la detección de inserciones, por lo que resulta la herramienta de elección para este tipo de eventos.

Conviene destacar que durante la evaluación del *dataset* de referencia, se han detectado SVs identificadas por Manta que no han sido llamadas durante el procesamiento de los ficheros de alineamiento. Dichas variantes están catalogadas como *LowQual* en el *dataset* de referencia. Se desconoce la configuración utilizada en el procesamiento de los ficheros de alineamiento del *gold-standard*, por lo que las discrepancias observadas podrían deberse a la utilización de una configuración distinta a la predefinida.

En relación a la evaluación de la herramienta Delly, es llamativa la diferencia de SVs llamadas en comparación con el resto de detectores en términos cuantitativos. Delly realiza un mayor número de *callings*, que se traduce en un incremento de la sensibilidad acompañado de unos valores predictivos positivos bajos que correlacionan inversamente con el tamaño de la SV. De forma particular, es destacable el bajo rendimiento obtenido en la detección de inserciones mayores de 200 pb [58], lo que está

en línea con los resultados obtenidos. El bajo rendimiento puede deberse al algoritmo del detector [52], basado en un flujo de análisis secuencial y no integrativo. Delly utiliza una aproximación inicial filtrando las lecturas discordantes en tamaño u orientación mediante un abordaje *paired-end* y posteriormente, sobre las lecturas seleccionadas, realiza un análisis en bloque de sus puntos de ruptura (*split-read*) [40].

Por su parte, la herramienta Lumpy presenta en líneas generales unos niveles de sensibilidad menores. De manera particular, destaca su incapacidad para detectar inserciones, independientemente de su tamaño: no gestiona las grandes inserciones *de novo* ni las pequeñas inserciones, y las inserciones de gran tamaño intra e inter cromosómicas son catalogadas por la herramienta como BND (https://github.com/arq5x/lumpy-sv/issues/160). Cabe señalar que bajo la categoría de BND Lumpy engloba diferentes eventos: traslocaciones, inversiones de las que se conoce un único punto de ruptura o las ya referenciadas inserciones de gran tamaño intra/inter cromosómicas. Por tanto, es probable que el rendimiento de Lumpy en la detección de inversiones e inserciones sea mayor y que se hayan perdido SVs catalogadas como BND que requieren de un post-procesamiento para su utilización.

Para acabar, se ha valorado la implicación del uso de los detectores en combinación, evaluándose de forma independiente las SVs comunes entre parejas de detectores y entre los tres en conjunto frente al *dataset* de referencia. Esta aproximación permite niveles de precisión superiores en la detección de SVs, por lo que resulta un abordaje muy interesante especialmente en la evaluación de la herramienta Delly, con vistas a mejorar su bajo VPP. Sin embargo, la consecuente disminución de la sensibilidad supone una limitación y en conjunto el rendimiento global de la unión de Delly con un segundo o tercer detector es inferior al del resto de herramientas de forma aislada. Su uso en las inversiones con un tamaño inferior a 3000 pb constituye la excepción, ya que la unión de Delly junto a Manta permite mejorar significativamente la precisión de las llamadas con una disminución de la sensibilidad valorable, como se observa en las Figuras 13 y 14.

3 Análisis de expansiones

3.1 Metodología

3.1.1 Búsqueda de *datasets* de trabajo

Selección de muestras

El repositorio EGA [59] (del inglés, *European Genome-Phenome Archive*) es un repositorio alojado que recoge datos genéticos y fenotípicos de identificación personal secundarios a proyectos de investigación biomédica. Se han seleccionado los datos de secuencia depositados con el número de acceso EGAS00001002462[32], que incluye los datos de WGS de 118 individuos afectos de enfermedades genéticas causadas por expansiones de tripletes con diagnóstico genético (Síndrome de X frágil, Enfermedad de Huntington, Ataxia de Friedreich, Esclerosis lateral amiotrófica, distrofia miotónica, ataxia espinocerebelar y atrofia dentatorubro-pálidoluisiana).

Los datos de secuenciación fueron generados utilizando las muestras del repositorio Repositorio genético de células humanas NIGMS (del inglés, *National Institute of General Medical Sciences*), del Instituto Médico de Investigación Coriell: [EGAD00001003562]. Las muestras se secuenciaron mediante tecnología de secuencias cortas sin protocolo de PCR con el secuenciador HiSeq X de Illumina, obteniéndose lecturas de 150 pb *paired-end*. Las lecturas fueron alineadas frente al genoma de referencia GRCh37 con el alineador Isaac de Illumina. Previo consentimiento de la institución y tras solicitar acceso al repositorio al responsable del dataset (Michael A. Eberle), se obtuvieron los ficheros de alineamiento (BAM).

Una vez revisadas las muestras disponibles se seleccionaron tres cohortes de individuos con expansiones en los genes *FMR1*, *HTT* y *DMPK*. Se escogieron estos genes por considerarse representativos de distintos escenarios a evaluar: rangos de número de repeticiones variables, con puntos de corte de normalidad amplios y situados en distintas regiones del genoma.

Alteraciones en el gen *FMR1* se asocian fundamentalmente al desarrollo del Síndrome de X-frágil, un trastorno con un patrón de herencia dominante ligado al cromosoma X caracterizado por discapacidad intelectual moderada-severa y alteraciones del comportamiento, que afecta principalmente a los individuos varones. El diagnóstico es molecular, y se establece ante la presencia de un número de tripletes CGG en la región UTR 5' del gen superior a 200 repeticiones. Los individuos con un número de repeticiones comprendidas entre 55 y 200 se encuentran en rango de premutación, y están en riesgo de desarrollar el Síndrome de ataxia/tremor asociado a X-frágil independientemente delsexo [60].

Alteraciones en el gen *HTT* se asocian fundamentalmente al desarrollo de la Enfermedad de Huntington, un trastorno motor y cognitivo de carácter progresivo que debuta en la 4ª-5ª década de vida y presenta un patrón de herencia autosómico dominante. El diagnóstico es molecular y se establece ante la presencia de un número de tripletes CAG en el exón 1 del gen superior a 35 repeticiones. Los individuos con un rango de repeticiones comprendidas entre 27 y 35 se encuentran en rango de premutación, y si bien no están en riesgo de desarrollar sintomatología es importante su identificación dado que pueden transmitir el alelo expandido a su descendencia en rango de mutación [61].

Alteraciones en el gen *DMPK* se asocian fundamentalmente al desarrollo de distrofia miotónica tipo 1 con un patrón de herencia autosómico dominante. Se trata de un trastorno asociado a un espectro fenotípico de predominio óseo y muscular de expresividad variable, que engloba cuadros clínicos leves con cuadros congénitos severos que pueden conllevar la muerte a edades tempranas. El diagnóstico es molecular y se establece ante la presencia de un número de tripletes CTG en la región UTR 3' del gensuperior a 50 repeticiones. Los individuos con un rango de repeticiones comprendidas entre 35 y 49 se encuentran en rango de premutación, y si bien no están en riesgo de desarrollar sintomatología es importante su identificación dado que pueden trasmitir el alelo expandido a su descendencia en rango de mutación [(62)].

Para la constitución de las cohortes de trabajo, se seleccionaron muestras con expansiones en rango de mutación y premutación (en caso de disponibilidad) de los tres genes. Se seleccionaron también controles negativos con el objetivo de evaluar las herramientas en términos de especificidad. Los controles negativos corresponden a individuos incluidos en el dataset EGAS00001002462, que fueron escogidos por ser portadores de un alelo expandido en un gen diferente al de interés en cada una de las cohortes (ejemplo: como control negativo de la cohorte de Enfermedad de Huntington se seleccionaron individuos con un alelo expandido en el gen *DMPK*). Desde el repositorio, los datos disponibles de WGS son los ficheros de alineamiento (.bam) alineados frente al genoma de referencia GRCh37, que se descargan a través de la herramienta de descarga propia de EGA (pyega3). Tras la descarga de los ficheros de alineamiento y el genoma de referencia, se realizó su indexación mediante la herramienta SAMtools.

Adicionalmente, se recogió la información necesaria para la validación posterior de los resultados, que incluía el sexo de los individuos y el tamaño real de las expansiones. Esta información se ha obtenido a través del repositorio NIGMS, donde se ha depositado información fenotípica de los individuos y el genotipado de las expansiones, determinado experimentalmente mediante técnicas de referencia (*Southern-blot* o TP-PCR). En los casos en los que se informaba el tamaño de la expansión como un intervalo de confianza se ha utilizado la media como valor de la expansión, y en los casos en los que se informaba como ">" se ha tomado el punto de corte como valor de la expansión. La información referente a las muestras seleccionadas y su genotipo se recogen en la Tabla
3. En el caso de los controles negativos, no se dispone del genotipo de los individuos, habiéndose establecido de forma arbitraria su valor por debajo del límite de normalidad.

FMR1								
Control + Expansión (>200)			Control + Premutación (55-200)			Control Negativo (<55)		
Muestra	Sexo	CGGn	Muestra	Sexo	CGGn	Muestra	Sexo	CGGn
NA04025	M	645	NA06891	M	100-117	NA06895	M	23
NA06897	M	477	NA06894	F	30/78	NA13506	M	N
NA09237	M	931-940	NA06906	M	96	NA16227	F	N
NA07862	M	501-550				NA13509	F	15/70
NA07537	F	28/>200				NA13511	M	45-47

HTT					
Control + Expansión (>=36)			Control Negativo (<27)		
Muestra	Sexo	CAGn	Muestra	Sexo	CAGn
NA13509	F	15/70	NA03132	M	N
NA13506	M	17/48	NA03759	M	N
NA13503	F	17/45	NA03986	M	N
NA13511	M	45/47	NA03990	F	N
NA13514	F	15/52	NA04025	M	N
NA13512	F	16/44	NA04567	F	N
NA13515	M	16/66	NA07862	M	N
NA13508	M	22/58	NA09237	M	N

DMPK					
Control + Expansión (>=50)			Control Negativo (<35)		
Muestra	Sexo	CTGn	Muestra	Sexo	CTGn
NA03759	M	N/>2000	NA13503	F	N
NA03986	M	N/>500	NA13506	M	N
NA03990	F	N/50-80	NA13508	M	N
NA23378	M	N/80-90	NA13509	F	N
NA03132	M	N/>1700	NA13511	M	N
NA04567	F	N/700	NA13512	F	N
NA06075	M	N/66	NA13514	F	N
NA05164	F	N/340	NA13515	M	N

Tabla3: Muestras seleccionadas para el estudio de expansiones. N: alelo en rango de normalidad, en el que no se especificaba la longitud exacta

Selección de herramientas

A partir de los resultados del estudio EGAS00001002462 seleccionado, en 2019 Dolzhenko E. *et al* publicaron una evaluación de la herramienta bioinformática de desarrollo propio ExpansionHunter [63] para la detección de expansiones en genomas generados a partir de secuenciación de lecturas cortas. En dicho estudio, se compararon los resultados de ExpansionHunter con los generados a partir de las herramientas ya existentes GangSTR y TREDParse [67]. Se ha tomado como punto de partida las herramientas caracterizadas en dicha revisión para realizar la búsqueda de las herramientas candidatas a evaluar en este trabajo.

En primera instancia se han revisado su uso en la comunidad científica en la base de datos *Web of Science*. Ordenadas según el número de citaciones, están: TREDParse (44 citaciones), GangSTR (18 citaciones) y ExpansionHunter (10 citaciones). Es destacable el bajo número de citaciones observadas, si bien ha de tenerse en cuenta que las herramientas son de reciente desarrollo. Las tres herramientas presentan métodos de detección combinados basados en dos o más tipos de lecturas (*enclosing reads*, *spanning reads*, *fully repetitive reads* o *flanking reads*), a diferencia de otras herramientas que utilizan un único tipo de lecturas como lobSTR[64], HipSTR[65] o STRetch[66].

Con ello, finalmente se decide trabajar con las herramientas ExpansionHunter, GangSTR y TREDParse. Las tres se encuentran disponibles en el repositorio público github, desde donde se han descargado e instalado.

3.2 Análisis bioinformático

Una vez en disposición de las herramientas bioinformáticas a evaluar, los ficheros de alineamiento de partida y el genoma de referencia GRCh37, previamente indexados, se diseñó el flujo de trabajo a seguir.

Se planteó un *pipeline* de trabajo dividido en tres bloques, uno por cada cohorte de patologías. El flujo de trabajo es similar en cada bloque: procesamiento de los ficheros de partida (BAM) con cada una de las tres herramientas bioinformáticas seleccionadas para la generación de los ficheros iniciales VCF de resultados, que recogen la localización, tamaño, intervalo de confianza (IC) de la estimación y tipo de lecturas utilizadas en la determinación de las expansiones de interés. Es importante resaltar que el algoritmo de los tres detectores requiere de la existencia de un catálogo predefinido de expansiones frente a las que realizar el realineamiento, en el que se define la localización y estructura de cada expansión.

ExpansionHunter

El *script* utilizado se encuentra disponible en el enlace Link. El procesamiento de las muestras requiere como *input* los ficheros de alineamiento y el genoma de referencia previamente indexados, junto con el catálogo de expansiones proporcionado desde el repositorio por ExpansionHunter. Se ha usado como argumento adicional el sexo de los individuos, dada la localización del gen *FMR1* en el cromosoma X.

El catálogo es un fichero en formato *.json* (Figura 17) que recoge para cada expansión: *(i)* el locus al que pertenece, *(ii)* su estructura, en forma de expresión regular, *(iii)* la región de referencia, definida en coordenadas genómicas en formato *0-based* (primera coordenada del cromosoma se establece en 0), *(iv)* el tipo de variante, pudiendo ser una repetición común (*repeat*, múltiples repeticiones de gran tamaño del motivo a lo largo del genoma) o una repetición rara (*rareRepeat*, no existen repeticiones largas del mismo motivo a lo largo del genoma) y *(v)* la existencia o no de regiones de alineamiento conflictivo (*offtargets*).

Figura 17: Ejemplo del catálogo de expansiones de ExpansionHunter, en formato json. Se muestra la información de Locus, motivo de la expansión, región de referencia, tipo de variante y regiones *offtarget*.

Tras el procesamiento de las muestras se obtuvo un archivo VCF individual para cada una de las muestras, en el que se detalla el genotipo de los 29 genes incluidos en el catálogo de expansiones. A continuación se extrajo la información relevante en relación al gen de interés (genotipo, metodología de la estimación, tamaño de la expansión, IC, tipo y número de lecturas usadas para su estimación y cobertura de la región) de cada fichero individual y se unificó en un archivo final de trabajo en formato .txt (Tabla 4). En función del tamaño de la expansión predicho por cada detector, se catalogaron las muestras como normales, premutadas o mutadas de acuerdo a la bibliografía.

Muestra	Genotipo	Metodología estimación	Tamaño	IC tamaño	Spanning Reads	Flanking reads	In-repeat reads	Cobertura	Clasificación
NA04025	1	INREPEAT	112	86-154	0	24	18	18.08108	Premutado
NA06897	1	INREPEAT	80	62-114	0	25	7	13.70270	Premutado
NA09237	1	INREPEAT	124	91-177	0	17	17	16.05405	Premutado
NA07862	1	INREPEAT	98	77-131	0	32	16	15.00000	Premutado
NA06891	1	INREPEAT	110	78-166	0	13	11	10.45946	Premutado
NA06906	1	INREPEAT	81	62-116	0	20	7	15.89189	Premutado
NA06895	1	SPANNING	23	23-23	7	21	0	18.48649	Normal
NA13506	1	SPANNING	30	30-30	3	21	0	15.16216	Normal
NA13511	1	SPANNING	30	30-30	8	21	0	17.83784	Normal
NA07537	1/2	SPANNING/INREPEAT	29/72	29-29/62-99	7/0	45/59	0/9	35.02703	Premutación
NA06894	1/2	SPANNING/INREPEAT	30/79	30-31/62-131	2/0	18/22	0/5	23.75676	Premutación
NA16227	1/2	SPANNING/SPANNING	25/34	25-25/34-34	4/3	12/14	0/0	24.56757	Normal
NA13509	1/1	SPANNING/SPANNING	30/30	30-30/30-30	11/11	28/28	0/0	30.72973	Normal

Tabla 4: Archivo final de trabajo con la información del procesamiento de las muestras seleccionadas para el estudio del gen *FMR1* mediante la herramienta ExpansionHunter.

GangSTR

El *script* utilizado se encuentra disponible en el enlace Link. Se ha usado como argumento adicional el sexo de los individuos, dada la localización del gen *FMR1* en el cromosoma X. El procesamiento de las muestras requiere como *input* los ficheros de alineamiento y el genoma de referencia previamente indexados, junto con el catálogo de expansiones proporcionado desde el repositorio por GangSTR.

El catálogo es un fichero en formato BED que recoge para cada expansión: *(i)* su localización en forma de coordenadas genómicas en formato *1-based* (primera coordenada del cromosoma se establece en 1), (ii) el tamaño del motivo, (iii) su secuencia nucleotídica, (iv) el gen al que pertenece y *(v)* la existencia o no de regiones de alineamiento conflictivo (*offtargets*).

Crom	Inicio	Fin	Extensión	Motivo	Gen	*Offtarget*
ChrX	146993569	146993628	3	CGG	FMR1	chr10:101295193-101295194, chr12:7781291-7781350, …
Chr4	3076604	3076660	3	CAG	HTT	
Chr19	46273463	46273522	3	CAG	DM1	

Tabla 5: Extracto del catálogo de expansiones de GangSTR en formato 1-*based*.

Conviene resaltar que si bien el catálogo se encuentra en formato 1-*based*, las especificaciones de GangSTR hacen al formato 0-*based*. Por ello, se generó unsegundo catálogo de expansiones al que se denominó catálogo ExpansionHunter (dadoque sus coordenadas coinciden con las del catálogo propio de ExpansionHunter).

Crom	Inicio	Fin	Extensión	Motivo	Gen	Offtarget
ChrX	146993568	146993628	3	CGG	*FMR1*	chr10:101295192-101295194, chr12:7781290-7781350, …
Chr4	3076603	3076660	3	CAG	*HTT*	
Chr19	46273462	46273522	3	CAG	*DM1*	

Tabla 6: Extracto del catálogo de expansiones de ExpansionHunter en formato 1-*based*.

Para facilitar su diferenciación, al catálogo 0-*based* se le asigna el nombre de catálogo ExpansionHunter. Con el objetivo de determinar si este hecho puede tener implicación en el rendimiento de la herramienta, se han procesado las muestras con el catálogo 1-*based* de GangSTR y el catálogo de ExpansionHunter en paralelo.

Por otra parte, GangSTR permite definir una serie de parámetros generales entre los que se incluye la opción de analizar el genoma entero en busca de expansiones o realizar un análisis dirigido a un loci específico (--*targeted*) y la opción de filtrar lecturas localizadas en regiones fuera de estos *loci*.

Con el objetivo de determinar la configuración más adecuada y la posible influencia del formato del catálogo en el rendimiento de la herramienta, se procesaron en primera instancia las muestras pertenecientes al bloque del gen *FMR1* de acuerdo a las siguientes aproximaciones:

-Catálogo 0-*based*
  -Catálogo de expansiones completo sin opción –*targeted*
  -Catálogo de expansiones completo con opción –*targeted*
  -Análisis de las muestras con un catálogo de expansiones dirigido al gen de interés y la opción –*targeted*
-Catálogo 1-*based*
  -Catálogo de expansiones completo sin opción –*targeted*
  -Catálogo de expansiones completo con opción –*targeted*
  -Análisis de las muestras con un catálogo de expansiones dirigido al gen de interés y la opción –*targeted*

Tras el procesamiento de las muestras bajo las diferentes combinaciones de parámetros, se obtuvieron los archivos VCF. Se extrajo la información relevante de cada VCF individual en relación al gen de interés (genotipo, tamaño de la expansión, IC, tipo y número de lecturas usadas para su estimación y cobertura de la región) y que se recoge en un archivo final de trabajo (Tabla 7), junto con la aproximación utilizada. En función del tamaño de la expansión predicho por cada configuración, se catalogaron las muestras como normales, premutadas o mutadas de acuerdo a la bibliografía.

Muestra	Aproximación	Genotipo	Tamaño	IC tamaño	Enclosing Reads	Spanning Reads	Flanking reads	In-repeat reads	Clasificación
NA04025	ALL_GangSTR	1	21	10-21	0	0	0	9	Normal
NA04025	ALL_EH	1	21	8-21	0	0	0	4	Normal
NA04025	ALL.target.GangSTR	1	85	85-85	0	0	6	10	Premutado
NA04025	ALL.target.EH	1	85	85-85	0	0	6	5	Premutado
NA04025	FMR1.GangSTR	1	118	80-118	0	0	6	10	Premutado
NA04025	FMR1.EH	1	118	118-118	0	0	6	5	Premutado
NA06897	ALL_GangSTR	1	18	17-20	0	0	0	3	Normal
NA06897	ALL_EH	1	17	20-20	0	0	0	0	Normal
NA06897	ALL.target.GangSTR	1	72	20-72	0	0	3	3	Premutado
NA06897	ALL.target.EH	1	72	20-72	0	0	3	1	Premutado
NA06897	FMR1.GangSTR	1	89	20-89	0	0	3	3	Premutado
NA06897	FMR1.EH	1	89	20-89	0	0	3	1	Premutado

Tabla 7: Muestra del fichero global con la información del procesamiento de las muestras seleccionadas para el estudio del gen *FMR1* mediante la herramienta GangSTR. Aproximaciones: ALL_GangSTR (catálogo de expansiones completo de GangSTR sin opción –targeted), ALL_EH (catálogo de expansiones completo de EH sin opción –targeted), ALL.target.GangSTR (catálogo de expansiones completo de GangSTR con opción –targeted) , ALL.target.EH (catálogo de expansiones completo de EH con opción –targeted), FMR1.GangSTR (catálogo de expansiones de GangSTR dirigido al gen *FMR1* y la opción –targeted) y FMR1.EH (catálogo de expansiones de EH dirigido al gen *FMR1* y la opción –targeted) . Genotipos posibles: 1 (homocigosis), 1/2 (heterocigosis)

Al comparar los resultados obtenidos con las diferentes aproximaciones (Figura 18), se observó que se obtenía un mejor grado de concordancia entre el tamaño estimado por GangSTR y el conocido experimentalmente al utilizar el catálogo de expansiones dirigido al gen de interés. A excepción de en la muestra NA07862, no se observaron diferencias entre la utilización del catálogo 0-*based* y el 1-*based*. En base a esto, se escogió la configuración dirigida al gen de interés con la opción –*targeted* y el catálogoen formato 1-*based* de GangSTR como configuración de referencia, haciéndose extensible su uso a los genes *HTT* y *DMPK*.

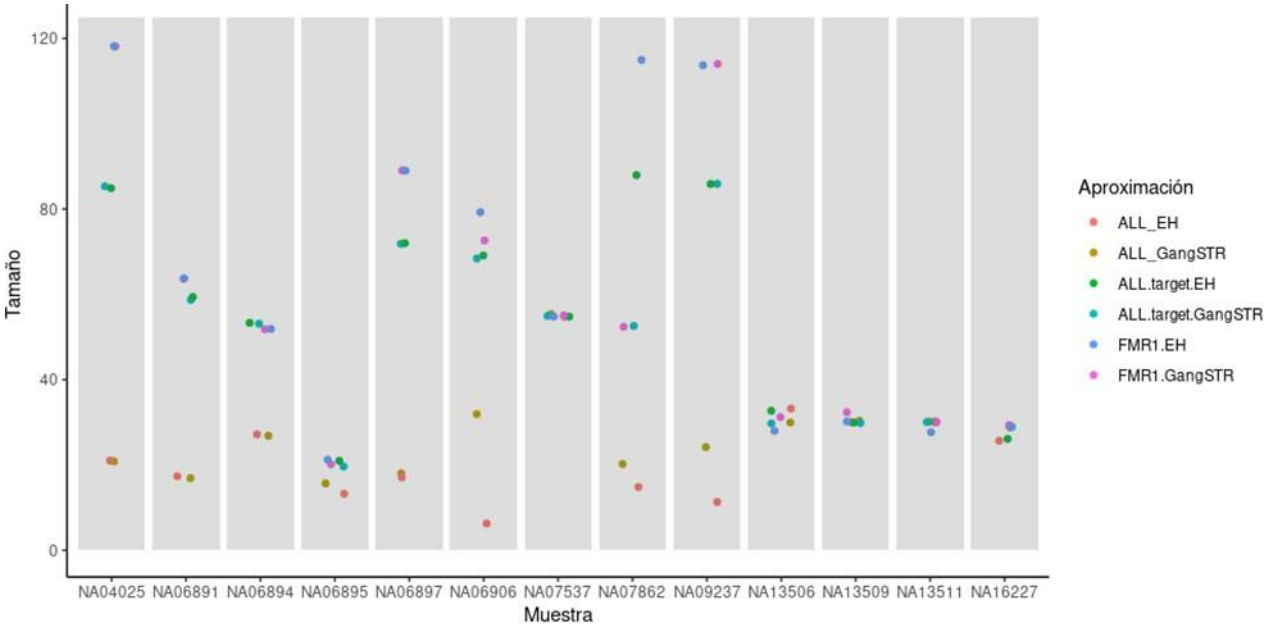


Figura 18: distribución de los tamaños estimados de la expansión del gen *FMR1* para cada una de las muestras en función de la aproximación utilizada. Muestras en rango de mutación: NA04025, NA06897, NA09237, NA07862, NA07537. Muestras en rango de mutación: NA06891, NA06894, NA06906.

TREDParse

El *script* utilizado se encuentra disponible en el enlace Link. Se ha usado como argumento adicional el sexo de los individuos, dado la localización de *FMR1* en el cromosoma X. El procesamiento de las muestras utiliza como *input* los ficheros de alineamiento y el genoma de referencia previamente indexados. El programa requiere de un catálogo de expansiones, que en este caso está incluido en el *pipeline* de trabajo del *software*.

TREDParse requiere definir las rutas de los ficheros de alineamiento en un fichero externo en formato csv. En dicho fichero puede especificarse si se desea realizar un estudio dirigido a un único gen de interés de los presentes en el catálogo, o si por el contrario se desea analizar todos. Una vez definido el csv de *input*, dirigido al gen de interés, se procesaron los archivos mediante el *script tred.py* proporcionado por la herramienta. A diferencia de los detectores anteriores, TREDParse no requiere de la especificación del sexo de la muestra, ya que infiere dicho parámetro a partir de la cobertura observada enel cromosoma Y.

Tras el procesamiento de las muestras se obtuvo un archivo VCF individual para cada una de las muestras, en el que se detalla el genotipo del gen seleccionado. A continuación se extrajo la información relevante (genotipo, tamaño de la expansión, IC, tipo y número de lecturas usadas para su estimación y cobertura de la región) de cada fichero individual y se unificó en un archivo final de trabajo en formato .txt (Tabla 8). En función del tamaño de la expansión predicho por cada detector, se catalogaron las muestras como normales, premutadas o mutadas de acuerdo a la bibliografía.

Muestra	Genotipo	Tamaño	IC tamaño	Spanning Reads	Flanking reads	In-repeat reads	Profundidad media del locus	Clasificación
NA04025	1/1	64/64	56-83|56-83	*NA*	4|1;5|2;8|3;9|1;11|1;12|2;15|1;17|1;22|1;40|2;44|1	50|5	17.57136	Premutado
NA06897	1/1	57/57	52-75|52-75	*NA*	4|1;5|1;8|1;12|1;19|1;23|2;35|1;43|1;45|1	49|1;50|1	14.60874	Premutado
NA09237	1/1	67/67	58-90|58-90	*NA*	7|2;9|2;12|1;23|1;24|1;25|3	49|1;50|4	14.68786	Premutado
NA07862	1/1	68/68	59-90|59-90	*NA*	6|1;7|2;8|2;11|2;12|1;15|2;16|2;20|1;26|1;32|1	49|2;50|4	16.49417	Premutado
NA06891	1/1	54/54	37-70|37-70	*NA*	6|1;7|1;10|1;12|1;20|1;37|1	49|1	12.91311	Normal
NA06906	1/1	57/57	35-74|35-74	*NA*	5|3;7|1;10|1;11|1;12|1;14|2;16|1;25|1;31|2;35|1	50|2	14.74078	Premutado
NA06895	1/1	23/23	23-23|23-23	23|2	5|2;7|2;8|2;9|1;10|3;13|3;16|1;21|1;23|5	49|1	19.04563	Normal
NA13506	1/1	30/30	30-30|30-30	30|2	5|1;10|2;13|2;16|1;17|1;20|4;22|1;23|1;29|1	*NA*	16.47864	Normal
NA13511	1/1	30/30	30-30|30-30	30|6	4|2;5|1;6|2;15|3;17|3;19|1;24|1;25|2;26|2;30|2	*NA*	18.58544	Normal
NA07537	1/2	29/41	29-29|41-60	29|4	4|1;5|2;6|1;8|4;9|1;12|2;14|4;15|4;19|1;20|3;22|2;2...	49|1	35.20388	Normal
NA06894	1/2	26/30	26-30|30-61	30|1	4|2;5|2;9|1;10|3;13|1;15|3;16|1;18|1;25|1;26|1	50|1	25.89175	Normal
NA16227	1/2	25/34	25-25|34-34	25|3;34|1	5|1;7|1;13|3;14|1;19|1;20|1;25|2;27|1;34|2	*NA*	28.78883	Normal
NA13509	1/1	30/30	14-30|30-30	14|1;30|6	5|1;6|1;7|1;9|1;10|1;12|1;13|1;16|1;17|1;19|1;20|4;...	*NA*	34.76262	Normal

Tabla 8: Muestra del fichero global con la información del procesamiento de las muestras seleccionadas para el estudio del gen *FMR1* mediante la herramienta TREDParse. Genotipos posibles: 1/1 (homocigosis), 1/2 (heterocigosis)

3.3 Preparación de los resultados

Los resultados obtenidos se evaluaron mediante la herramienta RStudio [57]. La sintaxis utilizada para la generación de las gráficas de resultados se encuentra disponible en el siguiente Link.

3.4 Resultados

Como resultado del análisis bioinformático, se generó un único archivo de resultados (.txt) para cada una de las tres cohortes de muestras en el que se incluye información sobre el genotipo, tamaño de la expansión, IC, tipo y número de lecturas usadas para su estimación, cobertura de la región del gen de interés y la clasificación del individuo en función del tamaño de expansión estimado.

En base a esta información, se realizó una primera aproximación global del rendimiento de las herramientas en términos de sensibilidad, especificidad, VPP y VPN. Posteriormente, se desglosaron los resultados por patología para evaluar el efecto del tamaño de las expansiones en el rendimiento del detector. Con el objetivo de simplificar la exposición de los resultados, se han unificado bajo la categoría de “mutado” tanto las expansiones en rango de premutación como en rango de mutación. De esta forma, se evaluó en primera instancia la capacidad de las herramientas de discriminar las expansiones en rango o no de normalidad (Tabla 9).

Resultado conocido **Global**	ExpansionHunter			GangSTR			TREDParse		
	Normal	Mutado	E / S	Normal	Mutado	E / S	Normal	Mutado	E / S
Normal	21	0	**E - 100**	21	0	**E - 100**	21	0	**E - 100**
Mutado	0	24	**S - 100**	4	20	**S - 83,3**	3	21	**S - 87,5**
VPN / VPP	**100**	**100**	**45**	**84**	**100**	**45**	**87,5**	**100**	**45**
Resultado conocido **FMR1**	ExpansionHunter			GangSTR			TREDParse		
	Normal	Mutado	E / S	Normal	Mutado	E / S	Normal	Mutado	E / S
Normal	5	0	**E - 100**	5	0	**E - 100**	5	0	**E - 100**
Mutado	0	8	**S - 100**	2	6	**S - 75**	3	5	**S - 62,5**
VPN / VPP	**100**	**100**	**13**	**71,4**	**100**	**13**	**62,5**	**100**	**13**
Resultado conocido **HTT**	ExpansionHunter			GangSTR			TREDParse		
	Normal	Mutado	E / S	Normal	Mutado	E / S	Normal	Mutado	E / S
Normal	8	0	**E - 100**	8	0	**E - 100**	8	0	**E - 100**
Mutado	0	8	**S - 100**	0	8	**S - 100**	0	8	**S - 100**
VPN / VPP	**100**	**100**	**16**	**100**	**100**	**16**	**100**	**100**	**16**
Resultado conocido **DMPK**	ExpansionHunter			GangSTR			TREDParse		
	Normal	Mutado	E / S	Normal	Mutado	E / S	Normal	Mutado	E / S
Normal	8	0	**E - 100**	8	0	**E - 100**	8	0	**E - 100**
Mutado	0	8	**S - 100**	2	6	**S - 75**	0	8	**S - 100**
VPN / VPP	**100**	**100**	**16**	**80**	**100**	**16**	**100**	**100**	**16**

Tabla 9: Rendimiento de las herramientas seleccionadas. Se presenta el rendimiento global de cada herramienta en términos de sensibilidad, especificidad, VPN y VPP (expresados en %) para el conjunto de los datos, y el desglose por patología.

En conjunto, la herramienta que mejor rendimiento presentó es ExpansionHunter, con una sensibilidad y especificidad del 100%. El resto de detectores presentaron un rendimiento menor y similar entre ellos, resultado de la existencia de falsos negativos. GangSTR posee una sensibilidad del 83.3% y un VPN del 84%, habiendo catalogado como dentro de la normalidad a 4 individuos con un alelo expandido. Por su parte, TREDParse posee una sensibilidad y un VPN del 87.5%, resultado de haber catalogado como dentro de la normalidad a 3 individuos con un alelo expandido.

Al desglosar los resultados por genes, se apreció un desempeño variable en función del gen estudiado.

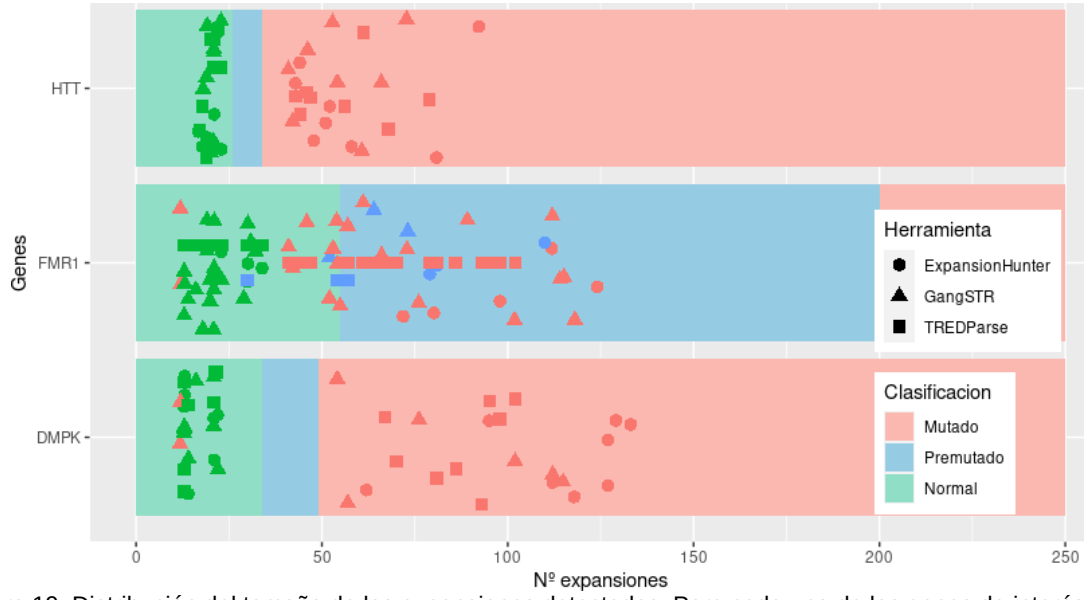


Figura 19: Distribución del tamaño de las expansiones detectadas. Para cada uno de los genes de interés, se resaltan los intervalos de clasificación establecidos en la interpretación de las expansiones (verde: normalidad, azul: premutación: rojo: mutación). Mediante un diagrama de puntos se representan las expansiones detectadas por ExpansionHunter, GangSTR y TREDParse, representados con un símbolo propio y colocadas a lo largo del eje x en el gráfico en función del tamaño de la expansión predicha. El color con el que se representa cada punto se basa en la categoría real de cada muestra (conocido experimentalmente).

En la figura 19 se representa el grado de acuerdo en la clasificación en la clasificación de los individuos entre las herramientas y el genotipo real de los individuos. Para cada gen, los puntos cuyo color es igual al color de fondo corresponden a muestras quehan sido clasificadas correctamente por el detector en su categoría de normal, premutado o mutado, mientras que los puntos cuyo color no es igual al color de fondo corresponden a muestras clasificadas erróneamente por el detector. Se observa un grado de concordancia mayor en los genes con un tamaño del rango de normalidad inferior.

-Resultados del gen *FMR1*

Ninguna de las expansiones en rango de mutación fue clasificada como tal por ninguno de los tres detectores, que estimaron valores significativamente inferiores a losconocidos experimentalmente. A la vista de este hecho, se planteó una evaluación de las herramientas enfocado a determinar su capacidad de discriminar las expansiones en rango o no de normalidad (Figura 18).

Bajo esta premisa, ExpansionHunter presenta una sensibilidad y una especificidad del 100% en las muestras estudiadas en la detección de individuos con un alelo expandido del gen. Por su parte, GangSTR presenta una sensibilidad del 75%, habiéndose detectado la presencia de dos falsos negativos (Tabla 10). Corresponden a los individuos NA06894 y NA07862, con un alelo expandido en rango de premutación y de mutación, respectivamente, en los que se estimó un tamaño de la expansión casi en el límite de la normalidad.

Muestra	Catálogo	Genotipo	Tamaño (repeticiones)	IC (tamaño)	Categoría
NA06894	1-*based*	1/2	32/52	28-55/30-55	Normal
	0-*based*		32/52	30-55/30-57	Normal
NA07862	1-*based*	1	52	20-115	Normal
	0-*based*		115	115-115	Premutación

Tabla 10. Resultado de las muestras NA06894 y NA07862 analizadas mediante la herramienta GangSTR. La herramienta ha detectado una expansión en el gen FMR1 de 52 tripletes, clasificándose como normal.

En la segunda muestra, si bien GangSTR ha detectado una expansión en el gen *FMR1* de 52 tripletes, el IC de la estimación engloba el punto de corte de riesgo de premutación (IC95%: 20-115 repeticiones). Así mismo, en este caso la utilización del catálogo en formato 0-*based* modifica significativamente el desempeño de la herramienta, ya que de haberse utilizado habría detectado adecuadamente la expansión.

Por último, la herramienta TREDParse presenta una sensibilidad del 62,5%, habiéndose detectado la presencia de tres falsos negativos: dos correspondientes a muestras con expansiones en rango de premutación y una correspondiente a la muestra NA07537 con expansión conocida en rango de mutación. En todos los casos el IC de la estimación englobaba el punto de corte de riesgo de la premutación.

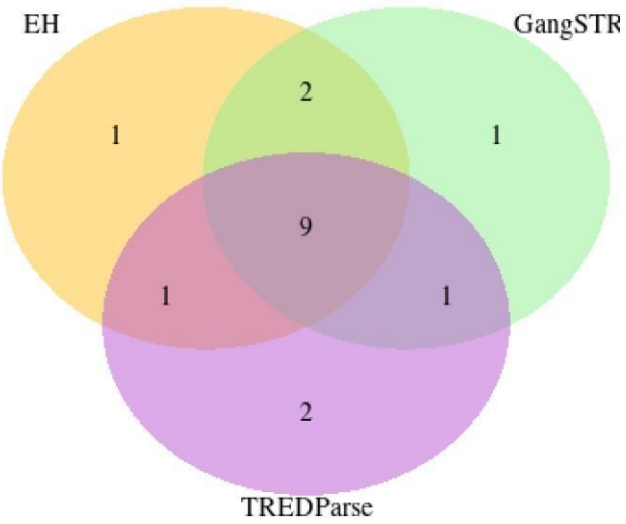


Figura 20: Resumen del grado de concordancia entre las 3 herramientas en la discriminación de los alelos en rango de normalidad. Los 3 detectores han incluido en la misma categoría 9 muestras.

Una vez evaluada la capacidad de las herramientas para identificar individuos con alelos expandidos del gen, se realizó una valoración del grado de concordancia entre detectores en la estimación del tamaño de la expansión (Figura 21).

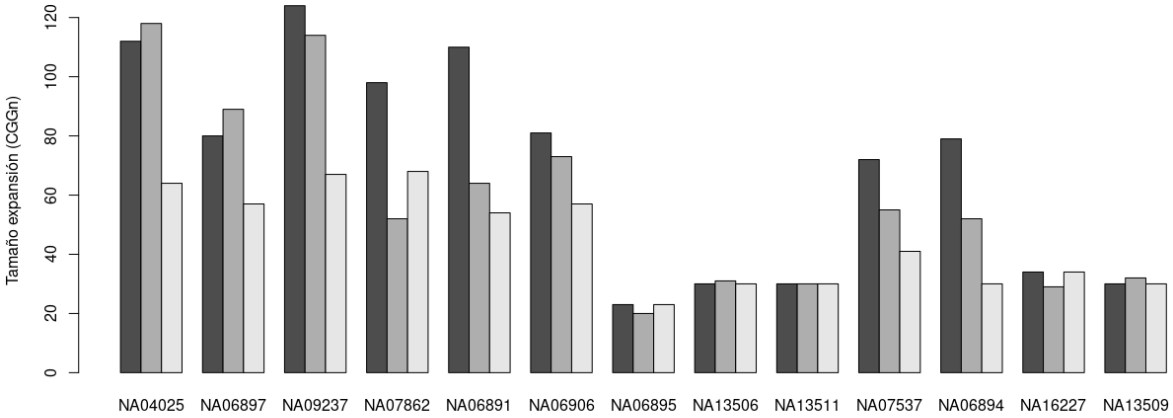


Figura 21: Distribución del tamaño de las expansiones detectadas para el gen *FMR1*

Como puede observarse, en muestras con expansiones en rango de normalidad las tres herramientas presentan un comportamiento similar y no generan estimaciones con diferencias significativas. Por el contrario, las diferencias se acentúan ante expansiones de mayor tamaño: en conjunto ExpansionHunter estima un mayor número de repeticiones, mientras que TREDParse presenta una tendencia a estimar valores significativamente más bajos.

-Resultados del gen *HTT*

El rendimiento observado en esta cohorte de muestras fue superior al presenciado en el escenario anterior en la detección de individuos con un alelo expandido del gen. Las tres herramientas presentan una sensibilidad y una especificidad del 100%. La relación entre el tamaño estimado y el tamaño real del alelo tiende a la linealidad y es independiente del tamaño. (Figura 22).

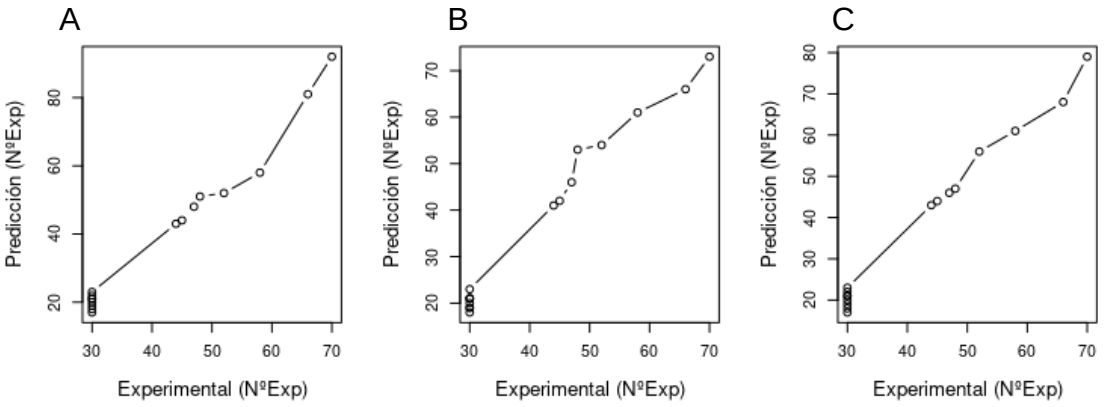


Figura 22: Relación entre los tamaños de las expansiones detectadas y las observadas para el gen *HTT*. A: ExpansionHunter, B: GangSTR, C: TREDParse.

En relación al grado de concordancia entre detectores en la estimación del tamaño de la expansión, las tres herramientas presentan de forma general un comportamiento similar (Figura 23). Cabe destacar el desempeño de ExpansionHunter, que estima valores superiores al resto de detectores en las muestras con expansiones de más de 60 repeticiones, por encima del valor real de la expansión.

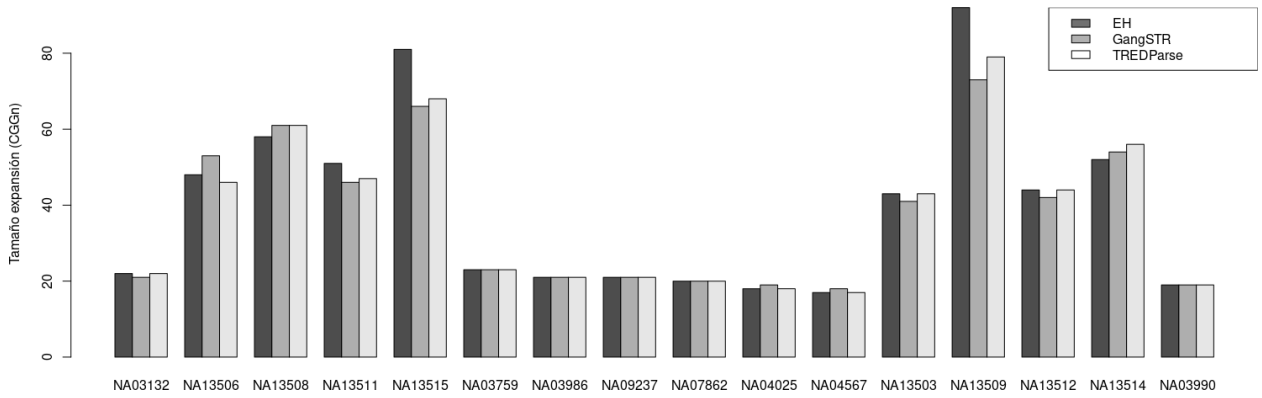


Figura 23: Distribución del tamaño de las expansiones detectadas para el gen *HTT*

-Resultados del gen *DMPK*

ExpansionHunter y TREDParse presentan una sensibilidad y una especificidad del 100% en las muestras evaluadas, mientras que GangSTR presenta una sensibilidad disminuida del 75%, consecuencia de la presencia de dos falsos negativos (Tabla 11). Corresponden a las muestras NA03986 y NA06075, pertenecientes a individuos con un alelo expandido del gen en rango de premutación.

Muestra	Catálogo	Genotipo	Tamaño (repeticiones)	IC (tamaño)	Categoría
NA03986	1-*based*	1	12	12-18	Normal
	0-*based*		12	12-18	Normal
NA06075	1-*based*	1	12	12-13	Normal
	0-*based*		12	12-66	Normal

Tabla 11. Resultado de las muestras NA03986 y NA06075 analizadas mediante la herramienta GangSTR. Se muestran los resultados del análisis utilizando catálogo de coordenadas de GangSTR y de ExpansionHunter.

En ambas muestras GangSTR ha detectado una expansión en el gen DMPK de 12 tripletes, sin contener el IC de la estimación el punto de corte de riesgo de premutación (IC95%: 12-66 repeticiones). Sin embargo, en el segundo caso la utilización del catálogo en formato 0-*based* modifica significativamente el desempeño de la herramienta, ya que de haberse usado el IC habría incluido la expansión.

En relación al grado de concordancia del tamaño de la estimación entre detectores, se observó un comportamiento similar en muestras con expansiones en rango de normalidad (Figura 24). Por el contrario, al aumentar el tamaño de la expansión se aprecian diferencias significativas, resaltando de forma particular la herramienta GangSTR que de forma ocasional estima valores significativamente más bajos.

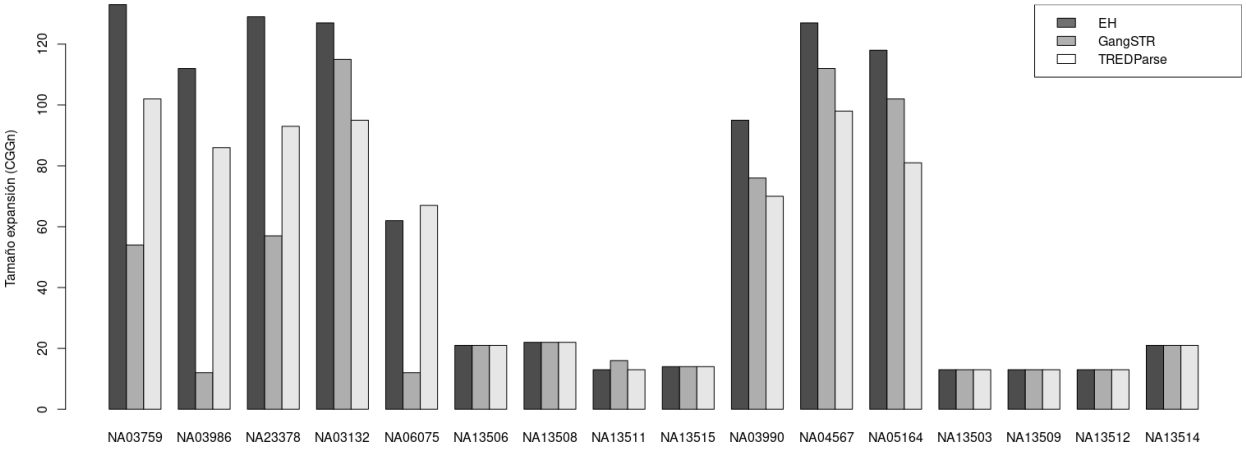


Figura 24: Distribución del tamaño de las expansiones detectadas para el gen *DMPK*

Respecto al grado de correlación entre el tamaño real de la expansión y el estimado por los detectores (Figura 25), se observa correlación menor que en el gen *HTT*, perdiéndose la linealidad a partir de valores experimentales superiores a 100 repeticiones. Al restringir la evaluación a expansiones con un número de repeticiones inferior a este punto de corte, se observa un aumento de la tendencia a la linealidad general para las tres herramientas, destacando el desempeño de la herramienta TREDParse.

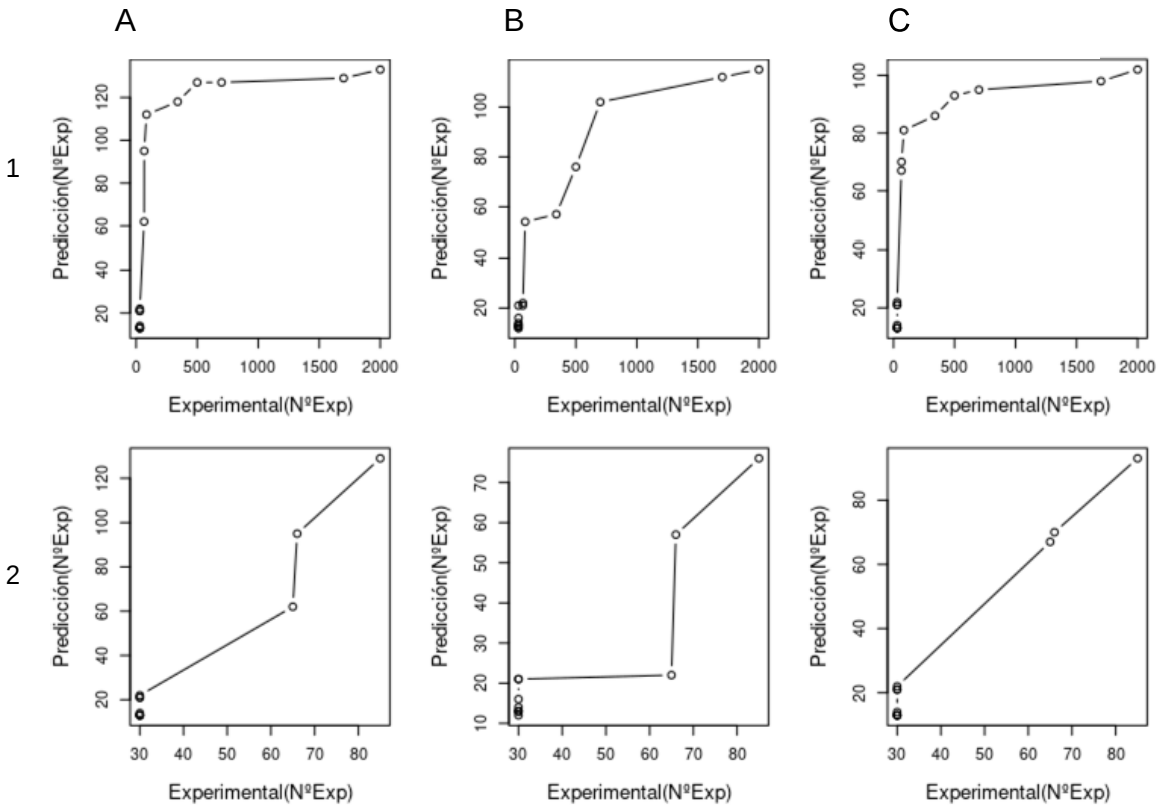


Figura 25: Relación entre los tamaños de las expansiones estimadas y las caracterizadas para el gen *DMPK (A*, ExpansionHunter, B GangSTR, C TREDParse). Cohortes: (1) En la fila se recogen los resultados de la comparación para el conjunto de muestras de la cohorte (2) En la fila se recogen los resultados de la comparación para las muestras con expansiones inferiores a 100 repeticiones de la cohorte.

3.5 Discusión

En este capítulo se ha llevado a cabo una evaluación del rendimiento de tres herramientas bioinformáticas dirigidas a la detección de expansiones a partir de datos de WGS. La evaluación surge con el objetivo de valorar el potencial de la WGS en el abordaje de este tipo de alteraciones genéticas, habitualmente estudiadas mediante técnicas alternativas.

Se realizó un análisis comparativo del grado de concordancia en la detección de expansiones mediante las herramientas ExpansionHunter, GangSTR y TREDParse en una cohorte de individuos afectos de enfermedades genéticas causadas por expansiones de tripletes en los genes *FMR1*, *HTT* y *DMPK*. La evaluación se ha centrado en dos puntos: la capacidad de las herramientas para distinguir los individuos sanos de los que están en riesgo de presentar un alelo expandido, y la precisión de las herramientas en la estimación del tamaño de la expansión.

A raíz de los resultados, la herramienta que mejor rendimiento en conjunto presenta es ExpansionHunter, que ha clasificado correctamente todas las muestras evaluadas. El resto de detectores presentan un rendimiento menor y similar entre ellos, resultado de la existencia de falsos negativos (ver Tabla 9). Al desglosar los resultados por genes, las diferencias entre GangSTR y TREDarse se acentúan, observándose un desempeño variable en función del gen estudiado. Este hecho podría tener relación con el tamaño del rango de normalidad de la expansión y el contexto genómico de la región que la contiene, como se infiere en la Figura 16. El rango de normalidad de los genes *HTT* y *DPMK* es inferior al tamaño de la *read* (hasta 27 expansiones para el gen *HTT* y hasta 35 repeticiones para el gen *DMPK*, correspondientes a 81 y 105 nucleótidos respectivamente), mientras que el rango de normalidad del gen *FMR1* se define hasta 55 repeticiones (165 pb) De acuerdo a lo esperado, la precisión de las herramientas es superior en genes con estimaciones de menor tamaño y/o cuando el tamaño de la expansión es inferior al tamaño de las lecturas de la secuenciación.

-Gen FMR1

En primer lugar se ha evaluado el desempeño de los detectores en el abordaje de la detección de expansiones en el gen *FMR1*. Se trata de una región compleja de abordar mediante *short-read sequencing*: por un lado presenta un elevado contenido en guanina/citosina (GC), que dificulta su secuenciación y disminuye la precisión de las estimaciones de los detectores [(68)], y por otro lado el tamaño de los alelos deletéreos (punto de corte de 200 repeticiones) es cercano al tamaño de la lectura. Acorde con ésto, el rendimiento observado en esta cohorte de muestras está en línea con lo descrito: ninguna de las expansiones en rango de mutación fueron clasificadas como tal por ninguno de los tres detectores, que estiman valores significativamente inferiores a los conocidos experimentalmente (ver Figura 19).

En relación a la capacidad de los detectores de discriminar los individuos con alelos en rango de normalidad de los que no, los tres presentan una especificidad y un VPN del 100%, sin diferencias significativas en el grado de concordancia del tamaño estimado de la expansión (ver Tabla 9). En términos de sensibilidad, todos los detectores presentan una sensibilidad y un VPP superior al 80%. ExpansionHunter fue la única herramienta que catalogó correctamente a todos los individuos en riesgo de presentar un alelo expandido, presentando una precisión superior al resto de detectores en la estimación del tamaño de los alelos en rango de premutación. La concordancia en el tamaño estimado es variable, y puede oscilar hasta en más de 40 repeticiones entre ExpansionHunter y TREDParse (ver figura 21).

GangSTR y TREDParse presentan una sensibilidad del 75% y 62,5%, respectivamente. La evaluación de los falsos negativos puso de manifiesto la importancia de considerar el intervalo de confianza de la estimación del tamaño de la expansión, que puede contener el punto de corte de riesgo de la premutación, como se observa en la Tabla 10.

Por otro lado, ninguna de las expansiones en rango de mutación fueron clasificadas como tal: los tamaños estimados para los alelos expandidos en rango de mutación son muy inferiores a lo observado experimentalmente, y no difieren significativamente de los de premutación. La incapacidad de las herramientas de estimar de forma precisa expansiones con un número de repeticiones superior al rango de normalidad limita su rendimiento a la hora de diferenciar entre individuos con expansiones en rango de premutación y de mutación.

-Gen HTT

En segundo lugar se ha evaluado el desempeño de las herramientas en el abordaje del gen *HTT.* Se trata de un gen con un umbral de normalidad menor (punto de corte inferior a 36 repeticiones) y un rango de alelos deletéreos de menor tamaño. Por ende, el tamaño de la STR es inferior al tamaño de la lectura y existe una menor diferencia de tamaño entre los dos alelos del gen que a *prior*i favorecen su abordaje. En relación a la capacidad de los detectores de discriminar los individuos con alelos en rango de normalidad de los que no, los resultados obtenidos están en línea con este planteamiento, y las tres herramientas poseen una sensibilidad y especificidad del 100% (ver Tabla 9). Por otro lado, el grado de concordancia en el tamaño estimado no presenta diferencias significativas entre detectores excepto en expansiones con una longitud superior a 60 repeticiones, en las que ExpansionHunter estima en promedio 10 tripletes más, como se aprecia en la Figura 23. La relación entre las estimaciones y el valor real de las expansiones presenta una correlación prácticamente lineal, con una ligera desviación atamaños inferiores a 40 repeticiones en las muestras correspondientes a los controles negativos (ver Figura 22). Se debe tener en cuenta que el tamaño de los controles negativos no ha sido determinado de forma experimental, sino que se ha establecido de forma arbitraria en 30 repeticiones, por lo que es probable que la desviación observadano corresponda con la realidad.

-Gen *DMPK*

En tercer lugar se ha evaluado el desempeño de las herramientas en el abordaje del gen *DMPK*, con un perfil intermedio entre los dos genes anteriores: presenta un umbral de normalidad similar a *HTT* y un rango de alelos deletéreos que tiende a lo observado en *FMR1*.

En relación a la capacidad de los detectores de discriminar los individuos con alelos en rango de normalidad de los que no, los resultados obtenidos son coherentes con este hecho y las tres herramientas presentan una especifidad del 100% sin diferencias significativas en la estimación del tamaño de la expansión. En términos de sensibilidad todos los detectores poseen un desempeño superior al 80%, siendo GangSTR el único detector que no ha catalogado correctamente a todos los individuos en riesgo de presentar un alelo expandido (ver Tabla 9).

Por otro lado, el grado de concordancia en el tamaño estimado para expansiones de mayor tamaño es variable en relación al tamaño de la expansión. En expansiones con un tamaño inferior a 100 repeticiones se observa un grado de concordancia elevado entre detectores y una correlación entre el tamaño real de la expansión y el estimado por los detectores que tiende a la linealidad, como se observa en la Figura 25. En expansiones con un número de repeticiones superior se acentúan las diferencias, que pueden oscilar hasta en más de 50 repeticiones entre ExpansionHunter y TREDParse. Cabe resaltar el desempeño de GangSTR, que presenta un grado de concordancia muy variable que no parece estar influenciado exclusivamente por el tamaño de la expansión, (ver Figura 24). Las muestras NA03759 y NA03132 presentan ambas un genotipo con un tamaño de expansión superior a 1700 repeticiones, por lo que a *priori* cabría esperar un comportamiento similar del detector en ambos casos. Sin embargo, en el primer caso el tamaño estimado de la expansión es significativamente inferior. Se plantea que esta diferencia esté relacionada con la cigosidad estimada por GangSTR, dado que todas las muestras en las que se reduce el tamaño de la estimación corresponden a individuos con un genotipo estimado homocigoto. Conviene señalar que la cigosidad estimada no corresponde con la estimada mediante métodos experimentales, como en el caso del individuo NA06075 (Tabla 11).

En conjunto, la herramienta que mejor rendimiento presenta es ExpansionHunter, que ha clasificado correctamente todas las muestras evaluadas. El resto de detectores presentan un rendimiento menor y similar entre ellos, resultado de la existencia de falsos negativos. GangSTR presenta una sensibilidad del 83.3% y un VPN del 84%, habiendo catalogado como dentro de la normalidad a 4 individuos con un alelo expandido. Por su parte, TREDParse presenta una sensibilidad y un VPN del 87.5%, resultado de haber catalogado como dentro de la normalidad a 3 individuos con un alelo expandido.

Al desglosar los resultados por genes, las diferencias entre GangSTR y TREDParse se acentúan, observándose un desempeño variable en función del gen estudiado. Este hecho podría tener relación con el tamaño del rango de normalidad de la expansión, como se infiere en la Figura 19. El rango de normalidad de los genes *HTT* y *DPMK* es inferior al tamaño de la *read* (hasta 27 expansiones para el gen *HTT* y hasta 35 repeticiones para el gen *DMPK*, correspondientes a 81 y 105 nucleótidos respectivamente), mientras que el rango de normalidad del gen *FMR1* se define hasta 55 repeticiones (165 pb)

En líneas generales, las tres herramientas han presentado un rendimiento adecuado en el abordaje de expansiones con un tamaño inferior al tamaño de la lectura, ejemplificado en el gen *HTT*. En el abordaje de expansiones de mayor tamaño, ExpansionHunter presenta un rendimiento global superior al resto de detectores, especialmente notorio en el abordaje de expansiones en el gen *FMR1*, mientras que GangSTR ha demostrado un desempeño inferior al resto.

Durante la evaluación de GangSTR se ha utilizado la configuración básica del detector, modificando exclusivamente los parámetros mencionados previamente. Por ello, no puede descartarse que una configuración distinta pudiera afectar positivamente a su desempeño. Independiente, cabe señalar la implicación de la utilización de una aproximación dirigida a un gen de interés y de la utilización del catálogo en formato 1-*based* en lugar del formato 0-*based* especificado en el rendimiento del detector.

4 Conclusiones

La WGS resulta una metodología de gran interés al permitir secuenciar el genoma completo del individuo y posibilitar la detección de una mayor variedad de alteraciones genómicas en un único estudio. A día de hoy la tecnología más extendida por su antigüedad y coste es la secuenciación de lecturas cortas*, que* presenta un elevado rendimiento en la detección de eventos SNVs e INDELs. En este trabajo se ha evaluado el potencial de distintas herramientas bioinformáticas en la detección de SVs y expansiones a partir de datos de WGS de lecturas cortas.

Tras la evaluación de los resultados se realizan las siguientes conclusiones sobre el abordaje del estudio de SVs y expansiones:

En relación al estudio de SVs:

-El rendimiento de las herramientas es dependiente del tipo de alteración. Se observa un mejor desempeño en la detección de deleciones e inversiones, y un desempeño inferior en la detección de inserciones.

-El rendimiento de las herramientas es dependiente del tamaño de la alteración. El tamaño repercute fundamentalmente en la precisión de los detectores, observándose un desempeño superior en la detección de eventos inferiores a 3000 pb y un desempeño muy inferior en la detección de inserciones mayores de 200pb.

-El rendimiento de las herramientas se ve influenciado por parámetros como la calidad de la llamada y la precisión de la estimación de sus puntos de ruptura. La restricción de las llamadas a aquellas con calidad PASS disminuye el desempeño general de los detectores en términos de sensibilidad y precisión. La restricción de las llamadas a aquellas con calidad PASS y precisión PRECISE limita aún más la sensibilidad de los detectores, aunque mejora su desempeño en términos de precisión con respecto al anterior.

-En la evaluación de los detectores, el grado de solapamiento mínimo requerido es un factor importante a tener en cuenta, dado que el desempeño de las herramientas varía considerablemente entre el corte más permisivo y el más restrictivo.

-En cuanto a la comparación entre detectores, en conjunto **Manta** presenta un rendimiento superior al resto e independiente del tipo de alteración, especialmente al trabajar con SVs con calidad PASS y puntos de ruptura PRECISE. Es la herramienta de elección en el estudio de eventos del tipo inserción. Por su parte, **Delly** presenta en primera instancia un desempeño superior al resto en términos de sensibilidad. Este hecho está en relación con el mayor número de llamadas que realiza, disminuyendo considerablemente su sensibilidad al restringir la calidad y precisión de las mismas. En último lugar, **Lumpy** no detecta inserciones.

-La utilización de varios detectores para la obtención de SVs de mayor confianza supone un aumento de la precisión diagnóstica aunque en la mayoría de los casos va acompañada de un importante descenso de la sensibilidad.

En relación al estudio de las expansiones:

-La capacidad de las herramientas de discriminar entre individuos sin riesgo de enfermedad es dependiente del límite superior del rango de normalidad de la expansión. Se observa un mejor desempeño en la detección de eventos con un rango de normalidad inferior a la longitud de la lectura.

-La capacidad de las herramientas de discriminar a los individuos en riesgo de presentar un alelo en rango de expansión es dependiente del límite inferior del rango de mutación de la expansión. Se observa un mejor desempeño en ladetección de eventos con un rango de mutación inferior a la longitud de la *read*.

-El rendimiento de las herramientas es dependiente del contexto genómico de la región que contiene la expansión. Se observa un desempeño inferior en el abordaje del estudio del gen *FMR1*, en relación al elevado contenido en citosina/guanosina de la región.

-El grado de precisión en la estimación del tamaño de la expansión está en relación con el tamaño real de la misma. Se observa un mejor desempeño en la estimación de la longitud de expansiones con un número bajo de repeticiones.

-En patologías causadas por expansiones en las que el umbral del rango de mutación sea inferior al tamaño de la lectura, el abordaje mediante las herramientas evaluadas presenta valor diagnóstico.

-En patologías causadas por expansiones en las que el umbral del rango de mutación sea superior al tamaño de la lectura, el abordaje mediante las herramientas evaluadas no presenta valor diagnóstico. Sin embargo, puede contribuir como método de cribado en la discriminación entre individuos sanos e individuos en riesgo, que requieran estudios adicionales de caracterización mediante técnicas de referencia.

-En cuanto a la comparación entre detectores, en conjunto **ExpansionHunter** presenta un rendimiento superior al resto e independiente del tamaño de la expansión y su localización. GangSTR presenta un rendimiento inferior relacionado principalmente con el tamaño de la expansión y TREDParse presenta un rendimiento intermedio dependiente sobre todo de la región genómica. Con todo, ninguna de las herramientas realiza una adecuada gestión de las expansiones de gran tamaño, ya que estiman longitudes muy inferiores a las reales.

Limitaciones del estudio

En relación al abordaje de SVs, la principal limitación radica en no haberse evaluado todos los eventos posibles incluidos en la definición de SVs: no se ha podido valorar el desempeño de las herramientas en la detección de traslocaciones ni de CNVs.

En el primer caso, el archivo *gold standard* no contiene traslocaciones, por lo que no se disponía de una referencia frente a la que evaluar el rendimiento de las herramientas. En el segundo caso, el estudio de las CNVs no fue posible dado el tiempo disponible para la realización de este trabajo. Aunque durante la planificación se contempló la dedicación de tiempo al abordaje de este tipo de eventos, el retraso en la consecución de los hitos limitó el tiempo disponible para su análisis. El retraso se debió fundamentalmente a la inclusión de una herramienta bioinformática adicional no contemplada inicialmente y a un cambio en la metodología de estudio de las SVs. En un

primer momento se planteó un abordaje mediante la herramienta vcftools, que permite comparar alteraciones con puntos de ruptura idénticos. Tras un primer análisis, y dado que la herramienta no contempla el grado de solapamiento entre alteraciones, se decidió cambiar de metodología y realizar un abordaje mediante la herramienta *bedtools*.

Asimismo, y como consecuencia del *dataset* de trabajo elegido para el estudio, no se ha determinado el rendimiento de las herramientas en términos de especificidad ni de valor predictivo negativo. Por otro lado, de forma general se ha trabajado utilizando la configuración predefinida de las herramientas, por lo que es probable que el uso de una configuración distinta o un proceso de post-procesamiento de los datos impacte el su rendimiento.

En relación al abordaje de expansiones, la principal limitación consiste en el limitado número de patologías estudiadas, que no resultan representativas de la diversidad de escenarios a considerar en la evaluación de las herramientas. Asimismo, no ha podido evaluarse correctamente el grado de correlación entre el tamaño estimado y el valor real de las expansiones en rango de normalidad. Ésto se debe a que el tamaño de las expansiones presentes en los controles negativos no ha sido determinado de forma experimental, sino que se ha establecido de forma arbitraria en 30 repeticiones.

Para concluir, el presente trabajo constituye una aproximación al potencial de las herramientas bioinformáticas disponibles en el abordaje de este tipo de alteraciones. Se escogió este tema por su utilidad clínica y traslacional, ya que la evaluación resultaba de interés con vistas a implementarlas en un futuro para su uso en proyectos de investigación. Por ello, todas las cuestiones mencionadas como limitaciones del trabajo resultan de interés y abren una línea de trabajo en el futuro para tratar de solventarlas.

5 Glosario

SV	Variaciones estructurales
STR	Short-tandem Repeats
BAM/CRAM	Archivos comprimidos que contienen las secuencias alineadas.
VCF	Archivo de llamada de variantes
BED	Formato de texto utilizado para almacenar coordenadas genómicas y anotaciones.

6 Bibliografía

1. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. Essays Biochem. 2018;62(5):643–723.
2. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(DATABASE ISS.):514–7.
3. Watson M. Clinical utility of genetic and genomic services: A position statement of the American College of Medical Genetics and Genomics. Genet Med. 2015;17(6):505–7.
4. Sanger F. NS and CR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74:5463–7.
5. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. XThe next-generation sequencing revolution and its impact on genomics. Cell [Internet]. 2013;155(1):27. Available from: http://dx.doi.org/10.1016/j.cell.2013.09.006
6. Access O. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. Genome Biol. 2015;16(1).
7. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090–6.
8. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A. 2007;104(49):19428–33.
9. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare mendelian disorders. JAMA - J Am Med Assoc. 2014;312(18):1880–7.
10. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–17.
11. Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016;18(7):696–704.
12. Yska HAF, Elsink K, Kuijpers TW, Frederix GWJ, van Gijn ME, van Montfrans JM. Diagnostic Yield of Next Generation Sequencing in Genetically Undiagnosed Patients with Primary Immunodeficiencies: a Systematic Review. J Clin Immunol. 2019;39(6):577–91.
13. Stenton SL, Prokisch H. Genetics of mitochondrial diseases: Identifying mutations to help diagnosis. EBioMedicine. 2020;56.
14. Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med [Internet]. 2019;21(11):2413–21. Available from: http://dx.doi.org/10.1038/s41436-019-0554-6
15. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: A practical guide to its clinical application. Brief Funct Genomics. 2016;15(5):374–84.
16. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. Hum Mutat. 2014;35(7):899–907.
17. Burdick KJ, Cogan JD, Rives LC, Robertson AK, Koziura ME, Brokamp E, et al. Limitations of exome sequencing in detecting rare and undiagnosed diseases. Am J Med Genet Part A. 2020;182(6):1400–6.
18. Verdura E, Schlüter A, Fernández-Eulate G, Ramos-Martín R, Zulaica M, Planas-Serra L, et al. A deep intronic splice variant advises reexamination of presumably dominant SPG7 Cases. Ann Clin Transl Neurol. 2020;7(1):105–11.

19. Barbitoff YA, Polev DE, Glotov AS, Serebryakova EA, Shcherbakova I V., Kiselev AM, et al. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. Sci Rep. 2020;10(1):1–13.
20. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112(17):5473–8.
21. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum Mutat. 2015;36(8):815–22.
22. An integrated encyclopedia of DNA Elements in the Human Genome. Nature. 2012;489:57–74.
23. Spielmann M, Mundlos S. Looking beyond the genes: The role of non-coding variants in human disease. Hum Mol Genet. 2016;25(R2):R157–65.
24. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102–10.
25. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):1–16.
26. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. Front Genet. 2019;10(MAY):1–14.
27. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med [Internet]. 2018;20(4):435–43. Available from: http://dx.doi.org/10.1038/gim.2017.119
28. Mattick JS, Dinger M, Schonrock N, Cowley M. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing the integration of genome sequencing with clinical records and data from the internet of things will transform health care. Med J Aust. 2018;209(5):197.
29. Marshall CR, Bick D, Belmont JW, Taylor SL, Ashley E, Dimmock D, et al. The Medical Genome Initiative: Moving whole-genome sequencing for rare disease diagnosis to the clinic. Genome Med. 2020;12(1):10–3.
30. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: Description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14(5):305–14.
31. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47(15):1–13.
32. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, Van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27(11):1895–903.
33. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online [Internet]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;00(00):1–3. Available from: http://arxiv.org/abs/1303.3997
35. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
36. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1–16.
37. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020;48(D1):D941–7.
38. Santpere G, Darre F, Blanco S, Alcami A, Villoslada P, Albà MM, et al. Genome-wide analysis of wild-type epstein-barr virus genomes derived from healthy individuals of the 1000 genomes project. Genome Biol Evol. 2014;6(4):846–60.

39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
40. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):333–9.
41. Michaelson, J.J. & Sebat J. ForestSV: structural variant discovery through statistical learning. Nat Methods. 2012;9:819– 821.
42. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. Nat Genet. 2015;47(3):296–303.
43. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):1–19.
44. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220–2.
45. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Stephen Pittard W, et al. The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. Genome Res. 2017;27(11):1916–29.
46. Chong Z. CK. Copy Number Variants. Methods in Molecular Biology. Structural Variant Breakpoint Detection with novoBreak. In: Bickhart D. (eds) , vol 1833. 2018. 129–141 p.
47. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865–71.
48. Zhao X, Emery SB, Myers B, Kidd JM, Mills RE. Resolving complex structural genomic rearrangements using a randomized approach. Genome Biol [Internet]. 2016;17(1):1–13. Available from: http://dx.doi.org/10.1186/s13059-016-0993-1
49. Arda Soylev, Can Kockan, Fereydoun Hormozdiari CA. Toolkit for automated and rapid discovery of structural variants. Methods. 2017;129:3–7.
50. Hormozdiari, F, Hajirasouliha, I , Dao, P, Hach, F, Yorukoglu, D, Alkan, C, Eichler, EE & Sahinalp S. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics. 2010;26:i350–7.
51. Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, et al. Wham: Identifying Structural Variants of Biological Consequence. PLoS Comput Biol. 2015;11(12):1–19.
52. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun [Internet]. 2019;10(1):1–11. Available from: http://dx.doi.org/10.1038/s41467-019-11146-4
53. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. Brief Bioinform. 2021;22(3):1–15.
54. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
55. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):1–4.
56. Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC. Simultaneous structural variation discovery in multiple paired-end sequenced genomes. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2011;6577 LNBI:104–5.
57. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA [Internet]. Available from: https://rstudio.com
58. Liu Y, Zhang M, Sun J, Chang W, Sun M, Zhang S, et al. Comparison of multiple algorithms to reliably detect structural variants in pears. BMC Genomics. 2020;21(1):1–15.
59. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical

research. Nat Genet [Internet]. 2015;47(7):692–5. Available from: http://dx.doi.org/10.1038/ng.3312
60. Hunter JE, Berry-Kravis E, Hipp H, et al. FMR1 Disorders. 1998 Jun 16 [Updated 2019 Nov 21]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2021. Available from: https://www.ncbi.nlm.nih.gov/books/
61. Caron NS, Wright GEB, Hayden MR. Huntington Disease. 1998 Oct 23 [Updated 2020 Jun 11]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2021. Bookshelf URL: https://www.nc.
62. Bird TD. Myotonic Dystrophy Type 1 Summary Genetic counseling Suggestive Findings. GeneReviews® [Internet]. 2019;1–27. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1165/pdf/Bookshelf_NBK1165.pdf
63. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35(22):4754–6.
64. Gymrek M, Golan D, Rosset S, Erlich Y. LobSTR: A short tandem repeat profiler for personal genomes. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2012;7262 LNBI:62–3.
65. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y, et al. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;14(6):590–2.
66. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: Detecting and discovering pathogenic short tandem repeat expansions. bioRxiv. 2017;
67. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, et al. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet [Internet]. 2017;101(5):700–15. Available from: https://doi.org/10.1016/j.ajhg.2017.09.013
68. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with short-read next-generation sequencing [version 1; referees: 3 approved]. F1000Research. 2018;7:1–11.

7 Anexos

Se incluyen en el anexo los siguientes archivos, disponibles en el enlace Link :

-Fichero global de rendimiento del procesamiento de SVs.

-Fichero *gold-standard Illumina Integrate* de SVs.

-Listado de programas con la versión utilizada.