



Facultad
de Medicina

Universidad Autónoma de Madrid
Facultad de Medicina

MÁSTER EN BIOINFORMÁTICA Y BILOGÍA COMPUTACIONAL

TRABAJO DE FIN DE MÁSTER

Optimización de la detección de CNVs en paneles y exomas mediante secuenciación masiva: implementación y validación clínica

Presentado por:
Sánchez Lara, Claudia

Tutor:
Rodríguez Antolín, Carlos
Unidad de Bioinformática, Servicio Genética-INGEMM, Hospital Universitario La Paz
de Dios Blázquez, Lucía
Unidad de Bioinformática, Servicio Genética-INGEMM, Hospital Universitario La Paz
Redrejo Rodríguez, Modesto
Departamento de Bioquímica, Facultad de Medicina, Universidad Autónoma de Madrid

Curso académico 2024-2025
10 de enero de 2025

Índice general

Índice de Abreviaturas	IV
Resumen	V
1 Introducción	1
1.1 Variantes estructurales en el genoma humano	1
1.2 La secuenciación de nueva generación (NGS)	3
1.3 Métodos de detección y evaluación de CNVs	4
2 Hipótesis y objetivo	7
3 Materiales y Métodos	8
3.1 Muestra de Estudio	8
3.1.1 Datos de Secuenciación de Exoma (WES) para NA12878	9
3.1.2 Datos de secuenciación de exoma (WES) para muestras control	9
3.2 Análisis bioinformático de los datos de secuenciación de exoma	10
3.2.1 Aplicación de tres análisis bioinformáticos para el procesamiento de los datos	10
3.2.2 Flujo teórico del análisis bioinformático	11
3.2.3 Calidad de las réplicas tras el análisis bionformático	11
3.3 Selección de estudios de variantes estructurales en NA12878 para la creación de un conjunto de validación de CNVs	11
3.3.1 <i>The 1000 Genomes Project</i>	12
3.3.2 <i>Genome in a Bottle Project (GIAB)</i>	13
3.3.3 <i>Benchmark</i> del algoritmo <i>svclassify</i>	13
3.4 Metodología para la construcción del <i>gold standard</i> de CNVs	14
3.5 Algoritmos de detección de CNVs en datos de exoma	15
3.5.1 cn.MOPS	15
3.5.2 CNVkit	15
3.5.3 CONTRA	15
3.5.4 ExomeDepth	15
3.5.5 Manta	16
3.5.6 LACONv	16
3.5.7 XHMM	16
3.6 Integración y optimización de herramientas bioinformáticas para la detección de CNVs	17
3.6.1 Singularity: contenedores científicos para la movilidad en computación	17
3.6.2 Snakemake: un motor de flujo de trabajo escalable para bioinformática	17
3.7 Métodos comparativos para la evaluación de los algoritmos	18
3.8 Disponibilidad del código	19
4 Resultados	20
4.1 Determinación del conjunto de referencia o <i>gold standard</i> para NA12878	20
4.1.1 Filtrado genérico de variantes y estandarización de los estudios de referencia	20
4.1.2 División de los conjuntos en función de calidad de variantes	21
4.1.3 Filtrado visual de variantes en función de su calidad	22
4.1.4 Análisis comparativo y fusión de los estudios con las variantes finales	25

Índice general

4.1.5	<i>Gold standard</i> de variantes para NA12878	28
4.2	Evaluación de las réplicas y los análisis bioinformáticos	29
4.2.1	Análisis de los kits de captura de exoma	30
4.2.2	Análisis de las réplicas de la muestra NA12878	30
4.2.3	Análisis de las <i>pipelines</i> bioinformáticas	31
4.3	Evaluación de los algoritmos	33
4.3.1	Análisis del tipo de variante reportada	35
5	Discusión	37
6	Conclusiones	40
Anexo 1		41
Anexo 2		43
Bibliografía		50

Índice de Abreviaturas

- aCGH** *Array Comparative Genomic Hybridization* (Arrays de Hibridación Genómica Comparada).
- ADN** Ácido desoxirribonucleico.
- AS** *Assembly de novo* (Ensamblaje de novo).
- BAM** *Binary Alignment Map*.
- BCL** *Binary Base Call* (Formato de archivo binario para secuenciación).
- BED** *Browser Extensible Definition*.
- Bp** *Base pairs* (Pares de bases).
- Chr** *Chromosome* (Cromosoma).
- CNP** *Copy Number Polymorphism* (Polimorfismos del número de copias).
- CNV** *Copy Number Variation* (Variación del número de copias).
- CPU** *Central Processing Unit* (Unidad Central de Procesamiento).
- dNTP** *Deoxyribonucleotide Triphosphate* (Desoxinucleótido).
- ddNTP** *Dideoxyribonucleotide Triphosphate* (Didesoxinucleótidos).
- DEL** *Deletion* (Deleción).
- DUP** *Duplication* (Duplicación).
- FDR** *False Discovery Rate* (Tasa de descubrimientos falsos).
- FN** *False Negative* (Falso negativo).
- FP** *False Positive* (Falso positivo).
- GATK** *Genome Analysis Toolkit*.
- GIAB** *Genome in a Bottle*.
- GRCh37** *Genome Reference Consortium Human Build 37*.
- HG** *Human Genome* (Genoma Humano).
- HMM** *Hidden Markov Model* (Modelo Oculto de Markov).
- HPC** *High-Performance Computing* (Computación de Alto Rendimiento).
- HT-NGS** *High-throughput next-generation sequencing* (Secuenciación de Nueva Generación de Alto Rendimiento).
- ID** Identificador.
- IDT** *Integrated DNA Technologies*.
- IGV** *Integrative Genomics Viewer* (Visualizador genómico).
- INDELS** *Small Insertions and Deletions* (Pequeñas Inserciones y Delecciones).
- INGEMM** Instituto de Genética Médica y Molecular.
- INS** *Insertion* (Inserción).
- INV** *Inversion* (Inversión).
- Kb** Kilobases.
- MAPQ** *Mapping Quality* (Calidad de mapeo).
- Mb** Megabases.
- NIST** *National Institute of Standards and Technology* (Instituto Nacional de Estándares y Tecnología).
- NGS** *Next Generation Sequencing* (Secuenciación masiva).
- NUMTs** *Nuclear Mitochondrial DNA Sequences* (Inserciones de ADN mitocondrial en el genoma nuclear).
- OCC** *One-Class Classification* (Clasificación de una sola clase).
- PCA** *Principal Component Analysis* (Análisis de Componentes Principales).
- PEM** *Paired-End Mapping* (Mapeo de lecturas emparejadas).
- PCR** *Polymerase Chain Reaction* (Reacción en cadena de la polimerasa).
- RD** *Read Depth* (Profundidad de lectura).
- SAM** *Sequence Alignment Map*.
- SBS** *Sequencing by Synthesis* (Secuenciación por Síntesis).
- SMS** *Single-Molecule Sequencing* (Secuenciación de Moléculas Únicas).
- SNP** *Single Nucleotide Polymorphisms* (Polimorfismos de un solo nucleótido).
- SNV** *Single Nucleotide Variants* (Variantes de un solo nucleótido).
- SR** *Split Reads* (Lecturas divididas).
- SV** *Structural Variants* (Variantes estructurales).
- TN** *True Negative* (Verdadero negativo).
- TP** *True Positive* (Verdadero positivo).
- VCF** *Variant Calling Format*.
- WES** *Whole-Exome Sequencing* (Secuenciación de todo el exoma).
- WGS** *Whole-Genome Sequencing* (Secuenciación de todo el genoma).

Resumen

Las variaciones en el número de copias (CNVs) son alteraciones genómicas que contribuyen significativamente a la diversidad genética humana y están estrechamente relacionadas con numerosas enfermedades. A pesar de su importancia, su detección continúa siendo un desafío debido a las restricciones de las tecnologías de secuenciación, la imprecisión en los puntos de corte y las dificultades para identificarlas en regiones complejas o repetitivas del genoma. Esto complica el desarrollo de conjuntos de variantes consistentes que permitan evaluar eficazmente el rendimiento de las herramientas de detección de variantes. Por ello, este trabajo propone dos objetivos principales: construir un *gold standard* de CNVs para la muestra NA12878 y evaluar el rendimiento de siete algoritmos de detección de CNVs.

Partiendo de cuatro estudios relevantes que reportan variantes estructurales, se generaron dos *gold standards*: uno robusto con variantes de alta confianza y otro más amplio, pero menos representativo. Para la evaluación, se utilizaron cinco réplicas de datos de exoma de la muestra NA12878, secuenciadas con diferentes kits de captura de exoma y analizadas con tres *pipelines* bioinformáticos. Se emplearon diferentes métricas de evaluación para validar el rendimiento de los algoritmos, además de calificar cómo los diferentes kits de captura y análisis bioinformáticos influyen en la detección de CNVs.

Los resultados mostraron un bajo rendimiento genérico de los algoritmos, como consecuencia de las limitaciones técnicas y analíticas de las herramientas utilizadas. A pesar de ello, este estudio proporciona metodologías útiles para mejorar la detección de CNVs en contextos clínicos.

1 Introducción

La variabilidad genética es la base de la diversidad entre individuos. Los genomas humanos varían entre sí de múltiples maneras, desde cambios a nivel de nucleótidos individuales hasta alteraciones estructurales más complejas [1]. Esta variabilidad sustenta la heredabilidad y constituye el núcleo de la relación entre la composición genética humana (genotipo) y sus rasgos asociados (fenotipo) [1, 2, 3]. Comprender cómo las variantes genéticas afectan a los rasgos observables y predisponen a enfermedades es uno de los objetivos centrales de la medicina actual [4].

Anteriormente, la relación entre el genotipo y el fenotipo ha sido difícil de comprender debido a las limitaciones tecnológicas [5]. Sin embargo, en las últimas décadas, la genética humana ha experimentado una revolución desde la publicación del borrador de la secuencia del genoma humano en 2001 por el Proyecto del Genoma Humano [6, 7]. Este hito proporcionó una base fundamental para el estudio de la genética humana y aceleró la comprensión de la diversidad genómica [4].

Posteriores avances en herramientas de secuenciación y análisis han permitido explorar el genoma humano en profundidad, posibilitando un mayor entendimiento de la variación genética hereditaria [3] y las mutaciones somáticas [6, 8].

1.1. Variantes estructurales en el genoma humano

Las diferencias más comunes entre los genomas pueden dividirse en dos tipos, la variantes de secuencia y las variantes estructurales [9, 10].

Las variantes de secuencia son las más comunes en el genoma humano, estando compuestas por variantes de un solo nucleótido (SNVs, *Single Nucleotide Variants*) y pequeñas inserciones y delecciones de menos de 50 pares de base (bp, *Base Pair*), denominados como INDELs (*small insertions and deletions*) [11, 9] (Figura 1.1). Cuando una variante SNV está presente en al menos el 1 % de la población se denomina polimorfismo de un solo nucleótido (SNP, *Single Nucleotide Polymorphisms*), siendo estos la variación genética más frecuente en el genoma humano [12].

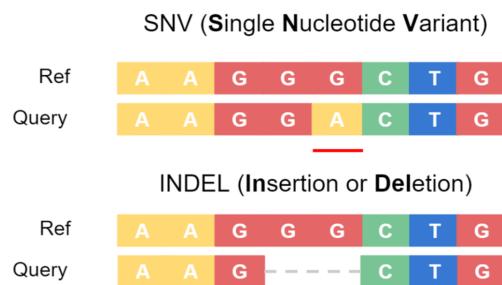


Figura 1.1: Variantes de secuencia, compuestas por SNVs e INDELs [10]

Por otro lado, las variantes estructurales (SV, *Structural Variant*) son extremadamente diversos en tipo y tamaño, incluyendo delecciones, inserciones, duplicaciones e inversiones, que van desde 50 bp hasta megabases (Mb) [9, 13] (Figura 1.2). Las SVs son la causa más significativa de variación genética humana, representan el 1,2 % de la variación entre los genomas humanos, mientras que los SNPs representan el 0,1 % [13, 14, 15].

A su vez, las SVs se dividen en dos categorías principales: variantes equilibradas y no equilibradas. Las variantes equilibradas no modifican el número de copias de un segmento de ADN (Ácido Desoxirribonucleico) en comparación con un genoma de referencia, como es el caso de las inversiones y translocaciones. Por otro lado, las variantes no equilibradas implican un cambio en el número de copias, como ocurre en las inserciones, duplicaciones o delecciones de material genético [13, 16]. En particular, las variantes no equilibradas representan el 99,8 % de las entradas reportadas en dbVar [14, 17]. dbVar es una base de datos del NCBI que recopila información sobre variantes estructurales genómicas, incluidas delecciones, duplicaciones, inserciones y otras alteraciones del genoma [17].

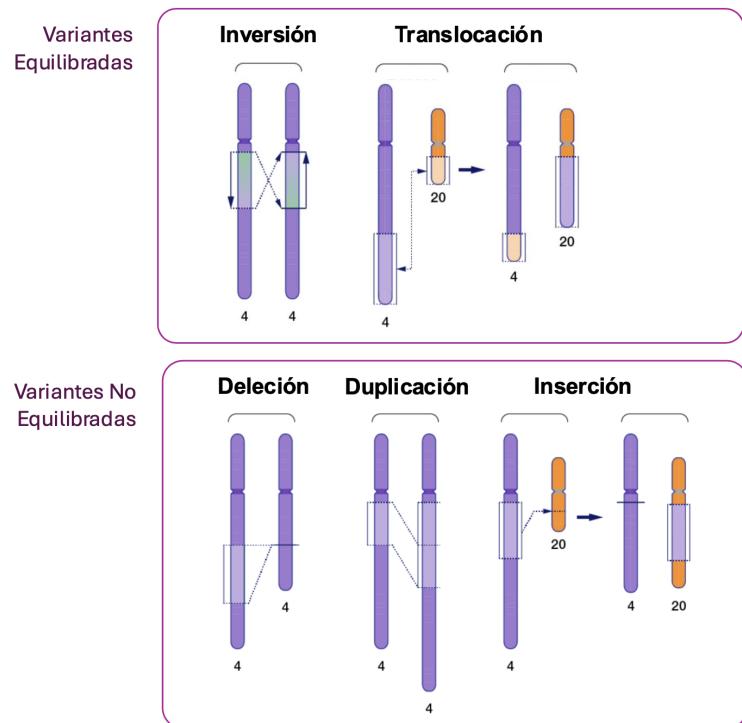


Figura 1.2: Representación de los tipos de variantes estructurales.

Las delecciones y duplicaciones no equilibradas también se conocen como variaciones del número de copias (CNVs, *Copy Number Variation*), con duplicaciones en tandem o intercaladas, en función de la distancia entre las copias duplicadas [11, 13]. Aquellas CNVs comunes en la población, con una frecuencia superior al 1 %, se denominan polimorfismos del número de copias (CNP, *Copy Number Polymorphism*). Aun siendo las CNVs la mayor causa de la variación genética en el genoma, han sido menos estudiadas que los SNVs. Esto se debe, en parte, a las dificultades en su detección, relacionadas con la diversidad de las SVs, su proximidad a regiones repetitivas y la falta de conjuntos de referencia confiables [9, 15, 18].

Al igual que las SNVs, las CNVs no necesariamente tienen un efecto negativo en la salud humana [18]. Sin embargo, entre el gran número de CNVs, algunas podrían estar directamente involucradas en enfermedades y fenotipos como el cáncer, la enfermedad de Parkinson, la pancreatitis, el lupus y trastornos neuropsiquiátricos [18, 19, 20]. Por ejemplo, en una gran cohorte del *UK Biobank* [21], una base de datos británica que contiene los datos genéticos y clínicos de medio millón de voluntarios, se identificaron 73 señales de asociación entre CNVs raras y 40 enfermedades, demostrando que estas CNVs incrementan el riesgo y adelantan la aparición de múltiples patologías en la población general [22].

Por estas razones, se debe disponer de métodos que permitan descubrir y evaluar eficazmente las CNVs en la población humana, además de asociar estas variantes con funciones biológicas específicas y su implicación en enfermedades humanas comunes y complejas [19, 23].

En los últimos años, los *microarrays* de SNP y la hibridación genómica comparativa basada en *arrays CGH* (*aCGH, Array Comparative Genomic Hybridization*) se utilizaron comúnmente para la detección de CNVs [15, 23]. Sin embargo, estas metodologías presentan varios inconvenientes, incluyendo ruido de hibridación, cobertura limitada para el genoma, baja resolución y dificultad para detectar mutaciones nuevas y raras [18].

La reciente implementación de la tecnología NGS (Next-Generation Sequencing) ha logrado superar las limitaciones de los métodos basados en *arrays* y ha impulsado un aumento exponencial en el descubrimiento de eventos de variación genómica [11].

1.2. La secuenciación de nueva generación (NGS)

El concepto de la secuenciación del ADN fue introducido por primera vez en 1975 por Frederick Sanger, con la técnica de secuenciación enzimática de ADN mediante didesoxinucleótidos marcados (ddNTP, *Dideoxynucleotide Triphosphate*) [24]. La secuenciación Sanger, conocida también como **secuenciación de primera generación**, se basa en amplificar fragmentos de ADN mediante la reacción en cadena de la polimerasa (PCR, *Polymerase Chain Reaction*) y añadir una mezcla de desoxinucleótidos normales (dNTP, *Deoxynucleotide Triphosphate*) y ddNTPs [25, 26, 27]. Este método tenía un rendimiento de secuenciación limitado, con altos costes y largos tiempos de ejecución [25]. Estas limitaciones impulsaron el desarrollo de tecnologías más rápidas y económicas, dando lugar a nuevas plataformas de secuenciación.

A partir de 2005, las tecnologías de nueva generación de alto rendimiento (HT-NGS, *High-Throughput Next-Generation Sequencing*) transformaron el panorama de la secuenciación. Estas nuevas técnicas permitían la secuenciación masiva en paralelo de millones de moléculas únicas de ADN, a un alto rendimiento y con costes más accesibles [25, 28]. Esto ha mejorado significativamente la comprensión de las bases genéticas de las enfermedades, facilitando la identificación de variantes patogénicas que tienen un impacto directo en un fenotipo específico [11, 29].

La primera revolución tras la secuenciación de Sanger, dio lugar a la **secuenciación de segunda generación**, que introdujo las técnicas de lectura corta de NGS [25]. Entre estas tecnologías, destaca la secuenciación por síntesis (SBS, *Sequencing by Synthesis*), un método que amplifica fragmentos de ADN y los sintetiza de forma controlada mediante la incorporación de nucleótidos marcados con fluorescencia en cada ciclo [26]. Esta metodología se ha consolidado como una de las técnicas más utilizadas en NGS.

Entre las plataformas basadas en SBS, destacan las desarrolladas por la compañía Illumina [30]. Su tecnología permite la secuenciación masiva en paralelo de lecturas cortas de entre 50 y 300 pares de bases (bp) aproximadamente, generadas a partir de decenas de millones de fragmentos de ADN amplificados simultáneamente [26, 31].

A pesar de los buenos resultados obtenidos con esta técnica, la amplificación por PCR puede introducir errores en la secuencia de bases o sesgar el proceso al favorecer ciertas secuencias sobre otras [24]. Además, el uso de lecturas cortas requiere un proceso de alineación con el genoma de referencia para reconstruir fragmentos largos de ADN, lo que dificulta la detección de variantes estructurales y el análisis de regiones genómicas complejas o repetitivas, ya que estas no se pueden mapear sin ambigüedades [15, 25].

A diferencia de los métodos de secuenciación de lectura corta, las tecnologías de lectura larga, también conocidas como tecnologías de **secuenciación de tercera generación**, están liderando una nueva era en la secuenciación de moléculas únicas por síntesis (SMS, *Single-Molecule Sequencing*) [32]. Estas tecnologías eliminan la necesidad de amplificar el ADN mediante procesos como la PCR,

lo que reduce significativamente los sesgos y errores asociados a este paso [24]. Además, tienen la capacidad de generar secuencias de más de 10 kilobases (Kb) a partir del ADN nativo [25, 26].

En la actualidad, diversas compañías persiguen el desarrollo de plataformas de secuenciación por síntesis sin necesidad de amplificación previa [24]. Entre ellas destaca Pacific Biosciences, cuya metodología permite obtener lecturas de entre 30 y 50 kb [26]. Estas tecnologías de lectura larga abordan los desafíos asociados a las lecturas cortas, como las regiones repetitivas a lo largo del genoma y la detección precisa de SVs [25].

Dependiendo del objetivo del análisis clínico, estas tecnologías pueden aplicarse a diferentes niveles de resolución genómica.

- **Secuenciación del genoma completo (WGS, Whole-Genome Sequencing).** Permite abarcar la totalidad del genoma. A parte de las regiones codificantes, WGS permite la detección de variantes no codificantes y variaciones estructurales complejas implicadas en procesos patogénicos [33].
- **Secuenciación del exoma completo (WES, Whole-Exome Sequencing).** Se centra exclusivamente en las regiones codificantes del genoma, que representan solo el 1-2 % del total. Aun así, se estima que el 85 % de las mutaciones que causan la enfermedad se encuentran en regiones codificantes y funcionales del genoma [29]. Esta técnica se ha convertido en una herramienta esencial para el diagnóstico de trastornos hereditarios [34].
- **Paneles personalizados.** Es un método de secuenciación dirigido que se centra en un conjunto de genes seleccionados según el objetivo del estudio, en los que se han encontrado mutaciones asociadas con una determinada patología [35]. Este enfoque permite reducir los costes y aumentar la profundidad de secuenciación en las regiones de interés [36]. Además, estos paneles pueden derivarse de capturas de exoma, focalizando el análisis en regiones concretas, simulando una captura experimental dirigida a dichas zonas genómicas (conocidos como paneles virtuales).

Cada enfoque de la tecnología NGS (WGS, WES o paneles de genes) se elige según las características de la enfermedad y las necesidades del diagnóstico [29].

1.3. Métodos de detección y evaluación de CNVs

Las tecnologías NGS generan un gran volumen de lecturas que requieren un análisis bioinformático exhaustivo para procesar estos datos y transformarlos en información interpretable [34]. Este análisis incluye varias etapas fundamentales: el filtrado de lecturas en base a criterios de calidad, la alineación y mapeo de las lecturas contra un genoma de referencia, la identificación de variantes genómicas, y su posterior anotación [37], como se representa en la Figura 1.3.



Figura 1.3: Flujo de trabajo de un análisis bioinformático para datos NGS.

Uno de los pasos clave en este flujo de trabajo es la llamada de variantes, que consiste en identificar las alteraciones genéticas presentes en los datos de secuenciación [37, 38]. Los algoritmos desarrollados para la llamada de variantes están diseñados y optimizados para abordar diferentes tipos de alteraciones genómicas, desde pequeñas variantes, como SNVs e INDELS, hasta SVs más complejas [34, 35, 37].

Los algoritmos utilizados para la detección de SVs también pueden aplicarse a la identificación de CNVs [18], aunque también existen herramientas específicas para estas últimas. Estas herramientas

emplean las estrategias de detección basadas en datos NGS representadas en la figura 1.4, ya sea a través de WGS, WES o paneles personalizados [39].

- **Profundidad de lectura (RD, Read Depth):** Analiza el número de lecturas en una región de interés. Las delecciones se asocian a regiones con cobertura significativamente baja, mientras que las duplicaciones se asocian a regiones con cobertura significativamente alta. Este método no es preciso respecto a los punto de corte (Figura 1.4.A) [14, 15, 18, 39].
- **Lectura dividida (SR, Split Reads):** Mediante lecturas de extremo emparejado, SR busca las lecturas donde solo un par está alineada con el genoma de referencia, mientras que la otra se mapean a distancias u orientaciones inesperadas. Este método resuelve con precisión los puntos de corte (Figura 1.4.B) [14, 15, 18, 39].
- **Mapeo paired-end (PEM, Paired-end Mapping):** Compara el tamaño medio de inserción entre los pares de lectura secuenciados reales con el tamaño esperado del genoma de referencia. PEM solo es aplicable a lecturas de extremo emparejado (Figura 1.4.C) [14, 15, 18, 39].
- **Ensamblaje de novo (AS, Assembly):** Se basa en el ensamblaje *de novo* generando los *contigs* a partir de lecturas cortas y comparándolos contra el genoma de referencia para encontrar las variantes estructurales. El ensamblaje con lecturas cortas es un problema desafiante que requiere altos recursos computacionales (Figura 1.4.D) [14, 15, 18, 39].

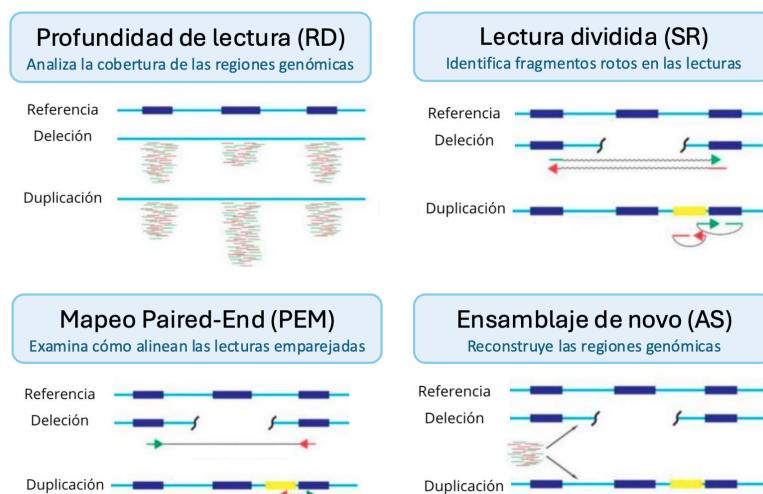


Figura 1.4: Visualización de los cuatro métodos para reportar SVs a partir de datos NGS [40]

Aunque los algoritmos de detección de CNVs pueden basarse en uno o varios de los métodos recién descritos, la profundidad de lectura es el método que predomina en las llamadas de CNVs basadas en datos NGS [40].

La secuenciación de lectura corta funciona razonablemente bien para detectar variantes pequeñas, como SNPs e INDELs. Sin embargo, analizar datos de NGS para identificar SVs y CNVs sigue siendo un desafío debido a problemas intrínsecos a la tecnología, como las longitudes de lectura cortas, la imprecisión de los puntos de corte, el sesgo de contenido de GC, la dificultad para detectar duplicaciones y las regiones genómicas complejas para secuenciar [15, 34, 36]. Además, todos los algoritmos actuales detectan delecciones con mayor precisión que duplicaciones, no pueden identificar CNVs en regiones repetitivas o áreas difíciles de mapear, y están limitados por la cobertura [18, 36].

Por esta razón, la secuenciación de tercera generación, con su capacidad para generar lecturas largas, está avanzando significativamente en la resolución de SVs, CNVs y secuencias complejas [34].

La evaluación del rendimiento de los algoritmos de detección de variantes requiere un conjunto de referencia confiable, conocido como *gold standard* [41]. Estos gold standards contienen variantes previamente caracterizadas que ya se conocen como verdaderas [34].

En la actualidad, los conjuntos de datos más ampliamente utilizados incluyen el consorcio *Genome in a Bottle* (GIAB) [42] y el Proyecto de los 1000 Genomas [43], que han generado amplios mapas de variantes utilizando tecnologías de secuenciación avanzadas [34]. Concretamente, GIAB ha caracterizado un genoma piloto (NA12878-HG001) del *International HapMap Project* [44], y dos tríos de hijo-padre-madre del proyecto *The Personal Genome Project* [45]. La muestra NA12878 se ha consolidado como el estándar de referencia para las SVs de línea germinal y se utiliza habitualmente como método de comparación [46].

Disponer muestras de ADN de estos individuos caracterizados y secuenciarlos de forma independiente permite evaluar el rendimiento de todo el flujo de trabajo bioinformático de un laboratorio, desde la preparación de la muestra hasta la llamada de variantes [34].

2 Hipótesis y objetivo

Las CNVs representan una proporción significativa de la variabilidad genética entre los genomas humanos, además están directamente relacionadas con diversas enfermedades y patologías. Por ello, es fundamental disponer de métodos precisos que permitan su identificación y caracterización en profundidad. Sin embargo, a pesar de los avances en las técnicas de secuenciación y el desarrollo de numerosas herramientas bioinformáticas para su detección y categorización, su identificación sigue siendo incompleta. Esto se debe, por un lado, a restricciones técnicas, como los errores generados durante la amplificación o la incapacidad de reconstruir fragmentos largos de ADN a partir de lecturas cortas, y, por otro, a las complejidades del genoma humano, como las regiones altamente repetitivas que dificultan el mapeo o el sesgo asociado a regiones con contenido extremo de CG.

Como consecuencia, la muestra *NA12878*, que se ha establecido como uno de los estándares de referencia en la comunidad científica para la validación de múltiples metodologías, carece de un conjunto de referencia o *gold standard* robusto y consistente para CNVs. Esto dificulta la evaluación del rendimiento y la identificación de las limitaciones de los algoritmos de detección de CNVs.

Ante estos desafíos asociados a la detección de CNVs, se plantea la siguiente hipótesis:

Partir de un conjunto de datos diverso, que incluya diferentes técnicas de captura de exoma y múltiples formas de procesar las lecturas, junto con la implementación de varios algoritmos bioinformáticos basados en distintas metodologías de detección y modelos matemáticos, permitirá identificar las herramientas y combinaciones más efectivas para la detección de CNVs. Además, la generación de un conjunto de validación interno para la muestra *NA12878*, basado en conjuntos de variantes ampliamente reconocidos y criterios de detección de alta confianza, facilitará una evaluación precisa del rendimiento de los algoritmos implementados y su capacidad para detectar delecciones y duplicaciones.

En este contexto, se plantea el siguiente objetivo general de este estudio:

Desarrollar un flujo de trabajo automatizado para la detección de CNVs en datos de exoma, integrando distintos algoritmos de detección de CNVs, técnicas de captura de exoma y herramientas bioinformáticas de procesamiento de lecturas. Además, se buscará generar un conjunto de validación interno basado en estudios de referencia para evaluar el rendimiento del flujo de trabajo. Para ello, se proponen los siguientes objetivos específicos:

1. Generar un conjunto de validación interno o *gold standard* robusto para la muestra *NA12878*, ampliamente estudiada y utilizada por la comunidad científica.
2. Implementar siete algoritmos bioinformáticos independientes para la detección de CNVs en datos de secuenciación de exoma.
3. Automatizar la ejecución de los algoritmos mediante un flujo de trabajo optimizado que utilice las herramientas Singularity y Snakemake, garantizando la reproducibilidad, escalabilidad y portabilidad de este análisis.
4. Evaluar el rendimiento de los algoritmos frente al *gold standard* generado, empleando métricas de validación.
5. Analizar cómo las distintas técnicas de captura de exoma y los diferentes procesamientos bioinformáticos de las muestras afectan los resultados obtenidos por los algoritmos.

3 Materiales y Métodos

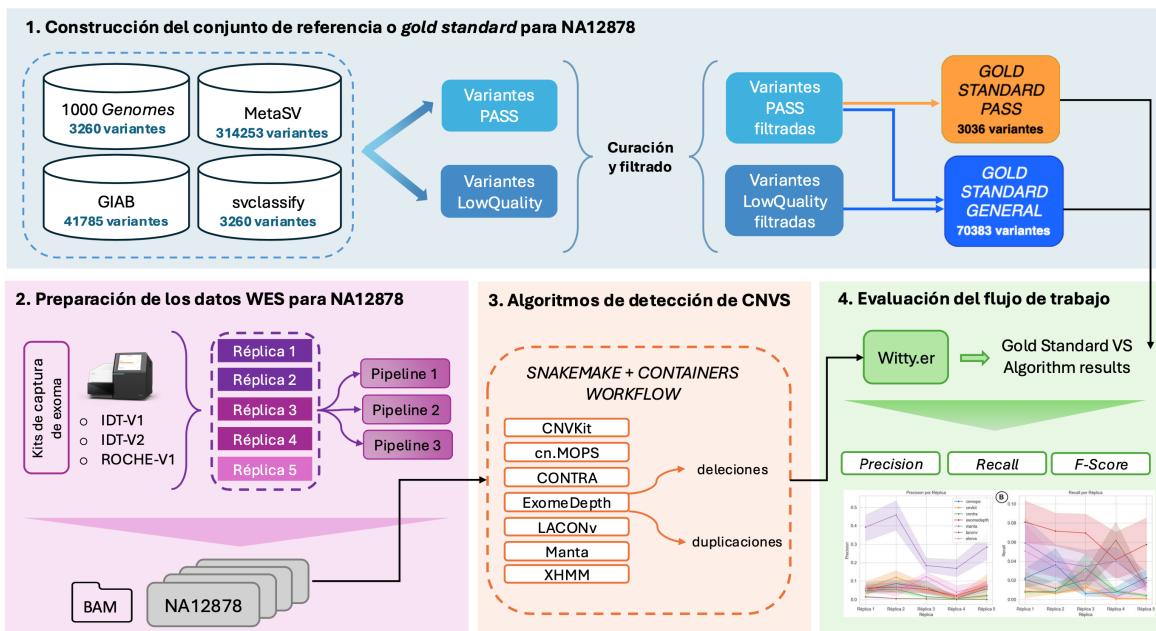


Figura 3.1: Descripción general del estudio. **a** Construcción del conjunto de CNVs de referencia para la muestra de interés NA12878. **b** Generación de múltiples conjuntos de datos de exoma para la muestra NA12878, utilizando cinco réplicas obtenidas con tres técnicas de captura de exoma y tres enfoques de análisis bioinformático. **c** Desarrollo e implementación de una pipeline automatizada, reproducible y escalable que integra siete algoritmos para la detección de CNVs en datos de exoma. **d** Validación integral del flujo de trabajo mediante métricas de rendimiento.

3.1. Muestra de Estudio

El consorcio *Genome in a Bottle* (GIAB) del Instituto Nacional de Estándares y Tecnología (NIST, *National Institute of Standards and Technology*) [42] desarrolla estándares y materiales de referencia para garantizar la precisión y calidad en la secuenciación del genoma humano en investigación y aplicaciones clínicas [42].

Sexo	Mujer
Población	Residentes de Utah (CEPH) con ascendencia europea del norte y oeste
Biosample ID	SAME123392
Padre	NA12891
Madre	NA12892

Tabla 3.1: Características demográficas y de parentesco de la muestra NA12878

Para este estudio se escogió la muestra NA12878 (NIST ID HG001; NIST RM 8398) (Tabla 3.1), siendo una de las muestras de referencia más utilizadas en genómica [34].

3.1.1. Datos de Secuenciación de Exoma (WES) para NA12878

La secuenciación de exoma completo (WES) [29] es una técnica fundamental para la identificación de variantes genéticas, enfocándose en las regiones codificantes del genoma humano.

Para evaluar la precisión y sensibilidad de los algoritmos de detección de variantes en el número de copias (CNVs), se utilizaron datos de secuenciación de exoma generados y procesados mediante análisis bioinformáticos en el Servicio de Genética-INGEMM (Instituto de Genética Médica y Molecular) del Hospital Universitario La Paz.

El servicio dispone de varias plataformas de secuenciación masiva, entre las que destacan los secuenciadores HiSeq 4000 y NovaSeq 6000 de Illumina [30], especializados en lecturas cortas de aproximadamente 150 bp. También cuenta con una gama de diversas tecnologías o kits de captura de distintos proveedores comerciales, que permiten la captura selectiva de regiones genómicas antes de la secuenciación, centrando el análisis en las regiones exónicas o paneles virtuales de interés.

Entre los kits de captura de exoma utilizados se incluyen IDT-V1, IDT-V2 y ROCHE-V1 especificados en la tabla 3.2. Cada kit contiene un archivo BED asociado que define las regiones exónicas especificadas por cada casa comercial que serán capturadas durante la secuenciación.

IDT-V1	xGen Exome Hybridization Panel IDT (Integrated DNA Technologies) [47]
IDT-V2	xGen Exome Hybridization Panel IDT versión 2 (Integrated DNA Technologies)
ROCHE-V1	KAPA HyperExome Roche [48]

Tabla 3.2: Kits de captura de exoma disponibles en el Servicio de Genética INGEMM

Con el objetivo de crear un conjunto de datos variado y robusto que evalúe eficazmente el rendimiento de los algoritmos, se utilizaron estos recursos para secuenciar cinco réplicas independientes de la muestra NA12878. Cada réplica se generó en distintas carreras de secuenciación, utilizando diferentes combinaciones de secuenciadores y kits de exoma, como se detalla en la tabla 3.3.

Plataforma	Kit de exoma	Réplica ID	Nombre de muestra
HiSeq 4000	IDT-V1	Réplica 1	NGS25603-HiSeq4000-exoma-Run200604-HG-0001
		Réplica 2	NGS25906-HiSeq4000-exoma-Run200604-HG-0001
NovaSeq 6000	IDT-V2	Réplica 3	Novaseq6000-Exome-Panel-v2-IDT-Run22070-HG001
		Réplica 4	Novaseq6000-Exome-Panel-v2-IDT-Run230711-HG001
	Roche-V1	Réplica 5	NGS38792-Novaseq6000-Exoma-Run230512-HG001

Tabla 3.3: Condiciones experimentales de las réplicas de la muestra NA12878.

Para facilitar la organización de este trabajo, las réplicas se nombrarán según su ID numérico (1, 2, 3, 4, 5), que corresponde al orden cronológico en el que fueron secuenciadas.

Disponer de estas cinco réplicas WES permitió evaluar no solo el rendimiento de los algoritmos de detección de CNVs, sino también el impacto de las condiciones experimentales. Esto incluye los distintos kits de exoma empleados, la profundidad de cobertura obtenida para cada muestra, y los posibles artefactos generados propios de la secuenciación.

3.1.2. Datos de secuenciación de exoma (WES) para muestras control

La mayor parte de los algoritmos de detección de CNVs implementados en este estudio requieren un conjunto de muestras control de referencia. Para ello se seleccionaron muestras de individuos

sanos y libres de patologías conocidas que pudieran interferir en el análisis de CNVs. Además, se incluyeron las muestras de referencia HG002, HG003 y HG004 del NIST.

Para garantizar la consistencia de la estructura de las réplicas en función del kit de exoma utilizado, se generaron tres conjuntos de muestras control: uno de 13 individuos secuenciados con el kit de exoma IDT-V1, otro de 12 individuos con el kit IDT-V2 y un tercer grupo de 13 individuos secuenciados con el kit Roche-V1.

Estos subconjuntos control se relacionan con cada una de las cinco réplicas de la muestra NA12878 en función del kit de exoma utilizado, generando tres grupos de muestras control finales que servirán como datos de entrada para el análisis bioinformático previo que requieren los datos y para los algoritmos de detección de CNVs. La información detallada de estos grupos de muestras se presenta en las tablas 1, 2 y 3 del Apéndice 1.

3.2. Análisis bioinformático de los datos de secuenciación de exoma

En un flujo de trabajo bioinformático, no solo se contempla la variabilidad experimental asociada al uso de diferentes secuenciadores y kits de captura de exoma. Las distintas herramientas bioinformáticas empleadas para convertir las lecturas crudas generadas por el secuenciador en datos limpios y listos para el análisis de variantes también introducen sesgos y variaciones en los resultados.

Disponer de un conjunto de datos procesados con diversas herramientas bioinformáticas permite analizar cómo las distintas combinaciones de herramientas influyen en los resultados finales.

Es por ello que se optó por implementar tres flujos de trabajo o *pipelines* bioinformáticas para el procesado de las cinco réplicas de la muestra NA12878 y de sus controles asociados.

3.2.1. Aplicación de tres análisis bioinformáticos para el procesamiento de los datos

Con el objetivo de maximizar la diversidad de los datos y capturar la variabilidad de las distintas herramientas bionformáticas usadas para el procesado de las lecturas, se implementaron tres flujos de trabajo de análisis bioinformático o *pipelines* disponibles en el Servicio de Genética-INGEMM que combinan diferentes herramientas para el procesado de las muestras. En la tabla 3.4 se detalla la combinación de herramientas en cada *pipeline*.

Pipeline	Genoma de referencia	Demultiplexado	Trimming	Alineamiento	Eliminar duplicados	Realineamiento y recalibrado
Pipeline 1	hg19	bcl2fastq	Trimmomatic	Bowtie2	Picard-tools	GATK3
Pipeline 2	b37	bcl2fastq	Trimmomatic	Bowtie2	SAMTools	GATK4
Pipeline 3	b37	bcl2fastq	Trimmomatic	Minimap2	SAMTools	GATK4

Tabla 3.4: Distribución de las tres *pipelines* bioinformáticas y herramientas empleadas en cada una

Para facilitar la organización de este trabajo, el ID de cada *pipeline* se va a denotar con sus herramientas distintivas, como se especifica en la tabla 3.5.

Pipeline 1	Bowtie2-Picard-GATK3
Pipeline 2	Bowtie2-SAMTools-GATK4
Pipeline 3	Minimap2-SAMTools-GATK4

Tabla 3.5: Denotación de las *pipelines*

Esta diversidad en los análisis bioinformáticos permite analizar cómo las distintas herramientas implementadas en cada *pipeline* afectan a los resultados finales obtenidos con los algoritmos de

llamada de CNVs, además de comparar diferentes métodos de alineamiento y procesado para identificar la mejor combinación de herramientas para la detección de CNVs.

3.2.2. Flujo teórico del análisis bioinformático

El análisis bionformático de las lecturas comienza con la herramienta `bc12fastq v2.20.0.442` [49], que convierte las lecturas obtenidas por el secuenciador Illumina del formato BCL (*Binary Base Call*) al formato FASTQ [50]. Este proceso de demultiplexado permite separar las lecturas de cada muestra en función de los índices y descartar aquellas lecturas de baja calidad. Posteriormente, `trimomatic v0.36` [51] realiza el recorte (*trimming*) para descartar lecturas de baja calidad y eliminar los adaptadores.

Las lecturas ya procesadas se alinean contra un genoma de referencia para determinar su posición genómica, generando archivos de alineamiento en formato BAM (*Binary Alignment Map*). En este estudio el alineamiento se realizó con las herramientas `bowtie2 v2.0.6` [52] y `minimap2 v2.22` [53]. Los genomas de referencia utilizados son:

- hg19: deriva de la versión GRCh37 (*Genome Reference Consortium Human Build 37* [54]) del genoma humano publicada en febrero de 2009. Contiene una secuencia mitocondrial diferente denotada con chrM y 9 ensamblajes de haplotipos alternativos.
- b37: el *Broad Institute* desarrolló este genoma derivado de GRCh37, incluyendo la secuencia mitocondrial rCRS y la secuencia del herpesvirus humano tipo 4. Tiene diferencias en la nomenclatura frente a hg19, omitiendo el prefijo *chr* y en la anotación de las regiones pseudo-autosómicas (PAR) de los cromosomas sexuales. Este genoma está optimizado para mejorar la compatibilidad con las herramientas del *Genome Analysis Toolkit* (GATK) [55, 56].

Tras el alineamiento, tanto Picard v1.141 como SAMtools v1.9 [57] se encargan de marcar y eliminar los duplicados, aquellas lecturas que parecen más de una vez debido a los ciclos de PCR llevados a cabo en el procesamiento de la muestra. Finalmente, GATK [56] realiza el realineamiento y recalibrado los datos ya alineados para mejorar aún más la calidad de las lecturas. La versión 3 de GATK utiliza el genoma hg19, mientras que la versión 4 está optimizada para trabajar con b37.

3.2.3. Calidad de las réplicas tras el análisis bioinformático

Cada una de las cinco réplicas definidas en la tabla 3.3, junto con sus controles descritos en el Anexo 1, fueron analizadas utilizando las tres *pipelines* bioinformáticas previamente descritas. En la tabla 3.6 se muestran las métricas de calidad de cada réplica en función del análisis bioinformático aplicado, obtenidas con la herramienta *SAMTools* [57]. También se reportan las coberturas medias, calculadas usando `mosdepth` [58] con un umbral de calidad de mapeo (MAPQ, *mapping quality*) de 20, un valor típico para asegurar que solo se cuentan las lecturas correctamente alineadas [59].

La cobertura es una medida clave para comparar los resultados finales, ya que refleja la profundidad de secuenciación. Una cobertura más baja podría limitar la precisión del análisis.

3.3. Selección de estudios de variantes estructurales en NA12878 para la creación de un conjunto de validación de CNVs

La muestra NA12878 ha sido seleccionada debido a su extenso uso como estándar para el estudio de la variabilidad del genoma humano y a la caracterización previa de sus variantes genómicas en diversos estudios [46]. Sin embargo, a pesar de ser un referente en la comunidad científica, no se dispone de un conjunto de validación de CNVs suficientemente robusto que permita una evaluación precisa y consensuada de los algoritmos utilizados para su detección [60].

Pipeline	Kit de exoma	Réplica ID	Cobertura media	Mismatch rate en lectura pareada	Lecturas mapeadas (%)	Lecturas alineadas sin duplicados (%)	Lecturas alineadas sin duplicados on target (%)	Eficiencia total (%)
<i>Pipeline 1</i>	IDT-V1	Réplica 1	72.02	0.0036	99.76 %	93.35 %	78.83 %	67.02 %
	IDT-V1	Réplica 2	99.65	0.0034	99.94 %	91.04 %	75.64 %	62.40 %
	IDT-V2	Réplica 3	107.82	0.0023	99.77 %	85.00 %	81.26 %	64.38 %
	IDT-V2	Réplica 4	312.14	0.0033	99.47 %	88.31 %	78.85 %	61.52 %
	ROCHE-V1	Réplica 5	40.77	0.0037	99.39 %	94.05 %	81.83 %	71.00 %
<i>Pipeline 2</i>	IDT-V1	Réplica 1	72.09	0.0037	99.77 %	93.34 %	78.83 %	67.06 %
	IDT-V1	Réplica 2	99.74	0.0035	99.95 %	91.03 %	75.64 %	62.44 %
	IDT-V2	Réplica 3	106.32	0.0026	99.79 %	84.98 %	81.09 %	63.04 %
	IDT-V2	Réplica 4	311.69	0.0033	99.48 %	88.39 %	78.48 %	60.84 %
	ROCHE-V1	Réplica 5	40.76	0.0038	99.39 %	94.05 %	81.81 %	70.85 %
<i>Pipeline 3</i>	IDT-V1	Réplica 1	73.25	0.0026	98.72 %	92.84 %	78.84 %	67.93 %
	IDT-V1	Réplica 2	102.21	0.0030	99.32 %	90.76 %	75.66 %	63.81 %
	IDT-V2	Réplica 3	109.73	0.0020	99.09 %	84.61 %	81.02 %	63.94 %
	IDT-V2	Réplica 4	321.42	0.0024	98.28 %	87.68 %	78.34 %	60.87 %
	ROCHE-V1	Réplica 5	41.71	0.0029	98.34 %	93.83 %	81.79 %	70.76 %

Tabla 3.6: Métricas de calidad para las réplicas de NA12878 en función del análisis bionformático aplicado.

Este trabajo tiene como objetivo comparar el rendimiento de siete algoritmos de detección de CNVs, por lo que se procede a generar una referencia interna para la muestra NA12878 que será usada como conjunto de validación.

Para ello se realizó una revisión exhaustiva de la literatura científica, recopilando los estudios más relevantes que han caracterizado SVs en profundidad para esta muestra. Tras un análisis detallado, se seleccionaron tres estudios que ofrecen cuatro conjuntos robustos de SVs de la muestra NA12878 para construir un conjunto de validación o *gold standard* robusto y consistente.

3.3.1. The 1000 Genomes Project

El grupo de Variaciones Estructurales del Proyecto 1000 Genomas generó un conjunto de datos ampliado de variantes estructurales para los 2.504 genomas de los participantes en la fase 3 del Proyecto 1000 Genomas [4, 43]. Para ello, se utilizaron los datos de secuenciación WGS de Illumina de los individuos de la fase 3 [61], junto con otras técnicas complementarias (como secuenciación de molécula única de lectura larga) para una mejor caracterización de SVs.

Se mapearon las lecturas (~ 100 bp, cobertura media de 7,4x) de 2.504 individuos a una versión modificada del genoma de referencia GRCh37 utilizando dos algoritmos de mapeo: BWA (*Burrows-Wheeler Aligner*) y mrsFAST. Después se realizó la llamada y el genotipado de SVs utilizando nueve algoritmos de detección de variantes diferentes: BreakDancer, Delly, VariationHunter, CNVnator, Read-depth, Genome STRiP, Pindel, MELT y Dinumt. El uso de secuenciación de lecturas largas permitió incorporar SVs adicionales que no se habían detectado en el proyecto original de los 1000 genomas.

Este conjunto de SVs está disponible en formato VCF (*Variant Calling Format*) en el repositorio fase 3, conteniendo 68.818 variantes estructurales detectadas entre los 2.504 individuos.

En cuanto al conjunto específico de variantes para la muestra NA12878, incluye 3.260 SVs, especificando los tipos en la siguiente tabla:

De las 3.260 SVs para NA12878, 2.192 están filtradas como variantes de alta calidad o *PASS*, siendo las 1.068 restantes consideradas como variantes de baja calidad *LowQual*.

En la Figura 3.2.1), se presentan la distribución general del tipo y la calidad de las variantes identificadas tanto por 1000 Genomes, como por el resto estudios explicados a continuación utilizados para generar el *gold standard*. Tanto esta Figura como todas las reportadas a lo largo del trabajo se obtuvieron utilizando los paquetes Matplotlib [62] y Seaborn [63] de Python v3.12 [64].

Tipo de variante	Descripción	Cantidad
Delecciones bialélicas	Delección en uno o ambos alelos	1982
Duplicaciones bialélicas	Duplicación en uno o ambos alelos	8
CNVs	Variantes en el número de copias	146
Inversiones bialélicas	Inversión en uno o ambos alelos	28
Inserciones NUMTs	Inserciones de ADN mitocondrial en el genoma nuclear	4
Inserciones de elementos móviles	Inserciones de elementos que se desplazan en el genoma	1092

Tabla 3.7: Variantes estructurales de la muestra NA12878 descritas por 1000 Genome Project.

3.3.2. Genome in a Bottle Project (GIAB)

En este trabajo [65], se secuenció el ADN genómico de NA12878 mediante tecnología de lectura larga de PacBio [66] para generar una cobertura de 23x con longitudes de lectura medias alineadas de 2.500bp y 5.000bp. La corrección de errores de las lecturas se realizó con Falcon [67], siguiendo los principios descritos en Chin et al.[68]. Las lecturas se alinearon con el genoma de referencia hg19 usando el alineador Blasr v1.3.1 [69] con los parámetros predeterminados. Para la detección de SVs se emplearon dos enfoques: PBHoney con su configuración por defecto y con una *pipeline* personalizada. Ambos métodos se aplicaron tanto a las lecturas en crudo como a las lecturas con corrección de errores. Adicionalmente, la *pipeline* personalizada se ejecutó con las lecturas alineadas mediante Blasr v1.3.2 [69], lo que incrementó la especificidad en la detección de variantes.

Para el ensamblaje genómico, las lecturas corregidas se ensamblaron en *contigs* utilizando Celera [70], y luego se aplicó la metodología de Chaisson et al. [71] para la llamada de variantes.

La combinación de estas herramientas dio lugar a siete metodologías de detección de SVs. Una variante se clasificaba como *PASS* si era reportada por al menos tres de las siete metodologías de detección de variantes. De lo contrario, se reportaba como baja calidad o *lt3* (*less than 3*).

El conjunto final está disponible en formato VCF en el repositorio de GIAB e incluye un total de 43.156 SVs para NA12878, representadas en la Figura 3.2.2), de las cuales:

- 20.957 variantes se reportan como delecciones y 22.199 como inserciones.
- 10.594 variantes están filtradas como *PASS* y 32.562 restantes son de calidad baja (*lt3*).
- 1.371 variantes están localizadas en cromosomas alternativos.

3.3.3. Benchmark del algoritmo svclassify

En el estudio de Parikh et al. [72], se presentó el método de clasificación svclassify, que calcula anotaciones de archivos BAM alineados y construye un modelo de clasificación de una sola clase (OCC, *One-Class Classification*) para identificar SV candidatos, clasificándolos como verdaderos positivos o falsos positivos con base en un criterio de puntuación.

Para validar este método, se utilizó la muestra NA12878 con cuatro conjuntos de datos de secuenciación (dos de lectura corta y dos de lectura larga). Se realizó un análisis con miembros de la familia de NA12878 y se seleccionaron delecciones detectadas en esta muestra y al menos dos muestras familiares. Además, se incorporaron dos conjuntos de variantes de los proyectos Personalis y 1000 Genomes, con 5.035 delecciones y 70 inserciones resueltas.

Con el objetivo de validar las SVs de alta confianza, se realizó una validación por PCR, junto con un análisis de trío madre-padre-hija, utilizando muestras de Illumina *Platinum Genomes* con una profundidad de secuenciación de 50x. En el análisis de trío se utilizó el algoritmo de consenso ortogonal MetaSV [73], el cual permite fusionar las variantes reportadas por diferentes métodos de detección de SVs, mejorando su confianza y precisión. Las SV candidatas con puntuaciones altas mostraron una concordancia del 99,7% con la validación de PCR y el método MetaSV.

Finalmente, los autores publicaron dos archivos BED en el repositorio de material suplementario, que contienen 2.676 delecciones y 68 inserciones de alta confianza. Estos archivos forman el tercer conjunto de referencia de CNVs para la muestra NA12878, representado en la Figura 3.2.4). Sin

embargo, los BED no contienen información adicional como genotipado o calidad de la variante, lo que impide obtener estadísticas adicionales.

Por otro lado, la herramienta MetaSV usada en este estudio generó un VCF de variantes estructurales que almacenaron en el repositorio [metasv_trio_validation](#). Este conjunto de variantes representa el cuarto y último conjunto de referencia utilizado en este trabajo para la creación de conjunto final de validación. Este conjunto, representado en la Figura 3.2.3), contiene 314.253 SVs, de las cuales:

- 4.718 variantes están filtradas como *PASS* y 309.535 como *LowQual*.
- 16.987 variantes son delecciones, 101.777 son inserciones y 195.489 son inversiones.

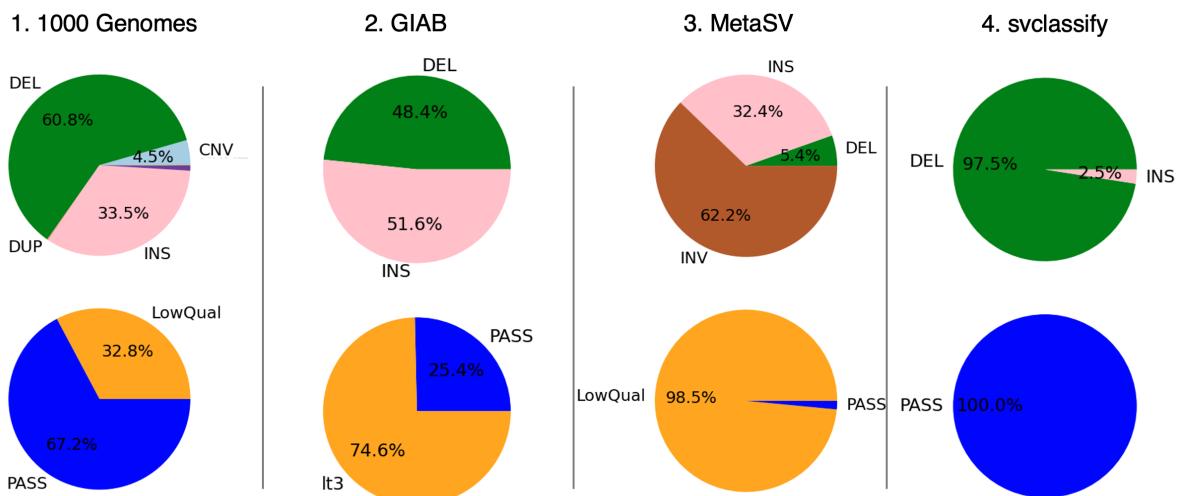


Figura 3.2: Distribución de las SVs por tipo y calidad, reportadas en los repositorios de: 1) 1000 Genomes, 2) GIAB, 3) MetaSV, 4) svclassify. En verde se representan las DEL, en rosa las INS y en marrón las INV. En azul se representan las variantes clasificadas como *PASS* y en naranja las *LowQual*.

La Tabla 4 del Apéndice 1 presenta un resumen del número total de variantes reportadas en los cuatro estudios de referencia, junto con su clasificación por tipo y filtrado de calidad.

3.4. Metodología para la construcción del gold standard de CNVs

A partir de las SVs reportadas en los cuatro estudios de referencia, el objetivo es integrar estos conjuntos de datos para construir un conjunto de validación o *gold standard* de CNVs para la muestra NA12878.

Para esta integración, se utilizó el paquete BEDTools v2.31.2, instalado en un entorno de Conda [74], con la versión 24.3.0 de conda. Debido a la dificultad para definir con precisión los puntos de corte de las variantes, se implementaron dos herramientas complementarias:

1. BEDTools Merge [75]: Genera puntos de corte en las máximas coordenadas de solapamiento, permitiendo así una mayor inclusión de variantes (Figura 3.3).
2. BEDTools Multiinter [76]: Define los puntos de corte en las mínimas coordenadas de solapamiento, garantizando una mayor especificidad de variantes (Figura 3.4). Para maximizar la precisión, esta herramienta se utilizó con los siguientes parámetros:
 - -f 0,5: exige que al menos el 50 % de una región esté cubierta por una región en los otros archivos para que se considere un solapamiento .
 - -r: asegura que el solapamiento sea recíproco en todas las regiones.

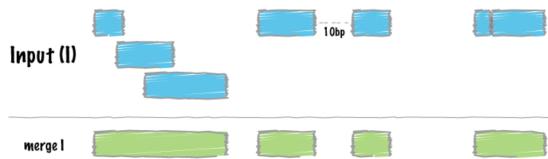


Figura 3.3: Representación gráfica del funcionamiento de BEDTools Merge [75].

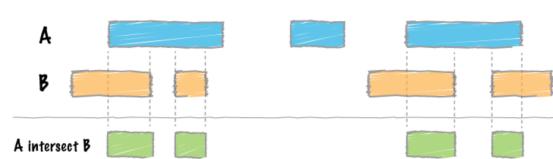


Figura 3.4: Representación gráfica del funcionamiento de BEDTools Multiinter [76].

3.5. Algoritmos de detección de CNVs en datos de exoma

En este trabajo se implementaron siete algoritmos de detección de CNVs en exomas, desarrollando los análisis en línea de comandos o *bash*, R [77] o Python [64], dependiendo de las características de cada algoritmo y siguiendo los manuales proporcionados por sus autores. En la mayoría de los casos, se utilizaron los parámetros predeterminados.

3.5.1. cn.MOPS

El paquete cn.MOPS [78] (*Copy Number estimation by a Mixture Of PoissonS*) implementado en R incorpora un modelo probabilístico de profundidad de cobertura en múltiples muestras para cada posición del genoma. A través de un enfoque Bayesiano, es capaz de separar el recuento de lecturas en CNVs y el ruido mediante componentes de mezcla y distribuciones de Poisson. Esta estimación del ruido permite reducir la tasa de falsos positivos (FDR, *False Discovery Rate*) al filtrar variantes con alto nivel de ruido.

Esta herramienta requiere al menos seis muestras, ya que se basa en comparar variaciones en el número de copia y ruido entre múltiples muestras.

3.5.2. CNVkit

CNVkit [79] es una herramienta de línea de comandos y una biblioteca de Python creada para detectar CNVs tanto en exomas completos como en paneles personalizados. CNVkit utiliza las lecturas *target* y las lecturas *off-target* capturadas para detectar CNVs en todo el genoma. Con ello logra una resolución a nivel de exón en las regiones objetivo y una resolución suficiente en las regiones intrónicas e intergénicas más grandes para identificar cambios en el número de copias. Su diseño flexible permite que se adapte con facilidad a diferentes plataformas de secuenciación, como Illumina o Ion Torrent.

3.5.3. CONTRA

El paquete CONTRA [80] (*COpy Number Targeted Resequencing Analysis*) detecta variantes tanto en exoma completo como en panel customizado de regiones pequeñas, para las cuales presenta muy buenos resultados. CONTRA infiere el número de ganancias y pérdidas en cada región de interés en función de la profundidad de lectura normalizada. Para ello utiliza diversas estrategias, como el uso de *log-ratios* a nivel base para corregir el sesgo de contenido de GC, la compensación de efectos de tamaño de biblioteca desequilibrado en los *log-ratios* y la estimación de las variaciones de *log-ratio* mediante *binning* e interpolación.

3.5.4. ExomeDepth

ExomeDepth [81] es un paquete de R que utiliza la profundidad de lectura para detectar CNVs. Compara el exoma de interés con un conjunto de referencia añadido, el cual se construye a partir

de exomas provenientes del mismo lote de secuenciación. Este algoritmo asume que las CNVs de interés no están presentes en el conjunto de referencia, por lo que los individuos emparentados deben ser excluidos. Esto supone una limitación al poder pasar por alto aquellas CNVs comunes entre la muestra de interés y la referencia. ExomeDepth es más adecuado para la detección CNVs raras y tiene un alto rendimiento tanto en exomas como en paneles customizados más pequeños, siempre que tengan suficientes exones (al menos 20 genes) .

3.5.5. Manta

El algoritmo Manta [82] es capaz de detectar SVs e INDELs combinando las metodologías de lecturas de mapeo paired-end (PEM) y lectura dividida (SR). Para ello, utiliza únicamente un genoma de referencia y los archivos BAM obtenidos a partir de un alineador. Está optimizado para el análisis de la variación de la línea germinal en pequeños conjuntos de individuos y la variación somática en pares de muestras tumor/normal. Además, destaca por su rapidez de análisis en condiciones estándar: analiza el genoma NA12878 con una cobertura de 50x en menos de 20 minutos en un servidor de 20 cores.

3.5.6. LACONv

LACONv (*Large Copy Number Variation Detection Algorithm*) es un algoritmo de desarrollo interno no publicado para CNVs en muestras secuenciadas con kits de lectura corta. Identifica CNVs en exones usando la profundidad de lectura (RD) normalizada, sin necesidad de muestras de referencia externas, ya que cada muestra actúa como control para las demás. Para cada exón se calcula un valor de dosificación y su valor z correspondiente en relación con la distribución de dosificación del conjunto de referencia.

3.5.7. XHMM

El algoritmo XHMM [83] (*eXome-Hidden Markov Model*) recoge información de CNVs a partir de datos de secuencias de exomas. Para ello, calcula la profundidad de lectura para cada exón y muestra, lo que proporciona una medida inicial de cobertura en las regiones de interés. Luego, realiza una normalización para corregir los sesgos técnicos y aplica un análisis de componentes principales (PCA, *Principal Component Analysis*) para identificar y eliminar las principales fuentes de variabilidad no biológica, como el contenido GC, el sexo y efectos del lote de secuenciación. Finalmente aplica un modelo oculto de Markov (HMM, *Hidden Markov Model*) para detectar CNVs mediante el análisis de variaciones en la cobertura a lo largo de las regiones de interés.

Los algoritmos utilizados y sus características principales se resumen en la tabla 3.8.

Algoritmo	Versión	Metodología	Tipo de secuenciación	Lenguaje de implementación	Necesidad de muestras control	Citas
cn.MOPS	v0.1	RD	WGS/WES	R	Sí	499
CNVkit	v0.9.11	RD	WGS/WES	Bash	Sí	1644
CONTRA	v2.0.8	RD	WES	Bash	Sí	385
ExomeDepth	v1.1.16	RD	WES	R	Sí	728
Manta	v1.6.0	RP/SR	WGS/WES	Bash	No	1748
LACONv	-	RD	WES	Python	Sí	-
XHMM	v1.0	RD	WES	Python/Bash	Sí	207

Tabla 3.8: Algoritmos implementados para la detección de CNVs en el exoma NA12878.

3.6. Integración y optimización de herramientas bioinformáticas para la detección de CNVs

Los análisis bioinformáticos implican flujos de trabajo complejos que encadenan diversos procesos, normalmente ejecutados por línea de comandos. Como se detalla en el apartado 3.2, para realizar la llamada de variantes no solo son necesarios los algoritmos de detección de variantes, sino también un análisis bioinformático previo completo que transformen las lecturas crudas obtenidas por el secuenciador en datos optimizados y de alta calidad adecuados para estos algoritmos.

Estos flujos de trabajo, cuando se implementan en la práctica clínica, deben estar optimizados y ser reproducibles para garantizar su ejecución eficiente en diferentes entornos. Es por esto que en este trabajo se emplearon herramientas bioinformáticas avanzadas capaces de automatizar y estandarizar el análisis. Por un lado, se utilizó Singularity [84], una plataforma de contenedores, que permite crear entornos reproducibles y portables. Por otro lado, se implementó Snakemake [85], un motor de flujos de trabajo que automatiza y escala procesos complejos mediante la integración de herramientas.

3.6.1. Singularity: contenedores científicos para la movilidad en computación

Un contenedor es una unidad estándar de *software* que agrupa el código y todas sus dependencias, permitiendo que la aplicación se ejecute en distintos entornos. A diferencia de las máquinas virtuales, que virtualizan el *hardware* físico e incluyen un sistema operativo completo, los contenedores son más eficientes y portátiles al compartir el núcleo del sistema operativo y optimizar sus recursos[86].

Aun siendo Docker [87] la herramienta de contenedores más utilizada en la industria tecnológica, Singularity [84] destaca en entornos científicos y en computación de alto rendimiento (HPC, *High-Performance Computing*), ambos de código abierto. Singularity ofrece entornos de computación totalmente portables mediante un único archivo de imagen en formato sif. Al usar imágenes, los usuarios trabajan en entornos reproducibles diseñados según sus necesidades, y estos pueden ejecutarse en cualquier sistema operativo que tenga instalado Singularity.

En este trabajo, se utilizó Singularity v3.7.0. La mayoría de las imágenes utilizadas para implementar los algoritmos de CNVs se obtuvieron del repositorio de contenedores [Docker Hub](#). Las imágenes restantes se crearon desde cero, empaquetando el algoritmo de CNVs y las herramientas necesarias para ejecutarlo en un archivo de definición para la imagen en formato def.

3.6.2. Snakemake: un motor de flujo de trabajo escalable para bioinformática

Snakemake [85] es un motor de automatización de flujos de trabajo comúnmente utilizado en bioinformática. Se basa en el sistema GNU Make, diseñado originalmente para gestionar la compilación de proyectos de *software* [88].

Un flujo de trabajo en Snakemake se define mediante un archivo Snakefile en formato smk integrado con el intérprete de Python, donde se especifican reglas que descomponen el proceso en pasos más pequeños. Cada regla indica cómo generar archivos de salida a partir de archivos de entrada, y Snakemake determina automáticamente las dependencias entre las reglas al hacer coincidir los nombres de archivos, optimizando así el orden de ejecución y evitando pasos redundantes.

Snakemake proporciona escalabilidad al optimizar el número de procesos paralelos, asignando los núcleos o *threads* de CPU (*Central Processing Unit*) que se deseen a cada regla, pudiendo hacer uso tanto a nivel usuario como en HPC sin modificar el flujo de trabajo. Además, se integra con el gestor de paquetes Conda y con el motor de contenedores Singularity, integrando el entorno en el propio flujo de trabajo. La integración con Singularity permite la ejecución de cada regla en una imagen, lo que garantiza portabilidad y consistencia en distintos sistemas operativos.

```

1 rule manta:
2     input:
3         bam = os.path.join(config["sample_dir"], config["sonda"], config["alignment"], "{sample}.bam"),
4         ref = config["human_reference"]
5     output:
6         vcf_diploid = os.path.join(config["results_dir"], "manta",
7             config["sonda"], config["alignment"], "{sample}", "diploidSV.vcf.gz"),
8         vcf_cand = os.path.join(config["results_dir"], "manta",
9             config["sonda"], config["alignment"], "{sample}", "candidateSV.vcf.gz"),
10    params:
11        outdir = os.path.join(config["results_dir"], "manta",
12            config["sonda"], config["alignment"], "{sample}")
13    threads: 10
14    singularity:
15        os.path.join(config["singularity_dir"], "manta.sif")
16    shell:
17        "/usr/bin/manta/bin/configManta.py --bam {input.bam} "
18        "--referenceFasta {input.ref} --runDir {params.outdir} --exome; "
19        "{params.outdir}/runWorkflow.py --mode local --jobs {threads}"

```

Listing 3.1: Ejemplo de regla definida en Snakefile para ejecutar el algoritmo Manta.

El flujo de trabajo se lanza con un archivo de configuración en formato yaml, en el que se especifican todas las variables necesarias que luego son referenciadas en el Snakefile.

```

1 snakemake -j 20 -s /ingemm/scratch/TFM/CNV/TFM/algorithms/pipeline/rule_manta.smk
2     --profile /ingemm/scratch/TFM/CNV/TFM/algorithms/profile
3     --configfile /ingemm/scratch/TFM/CNV/TFM/algorithms/pipeline/configuraciones/IDT-V2/
4     config_b37_minimap.yaml

```

Listing 3.2: Ejemplo de comando para lanzar un Snakefile.

En este trabajo se utilizó Snakemake v8.14.0 instalado en un entorno Conda [74].

3.7. Métodos comparativos para la evaluación de los algoritmos

Una vez obtenidos los resultados de los algoritmos para cada una de las cinco réplicas analizadas bioinformáticamente con las tres *pipelines* desarrolladas en el apartado 3.2, se requiere aplicar un método de validación eficaz que permita evaluar la concordancia entre las llamadas de variantes obtenidas con los algoritmos y los *gold standards* generados.

Para ello, se implementó el algoritmo *witty.er* v0.5.2 de Illumina [89], una herramienta diseñada para comparar SVs entre dos archivos VCF, evaluando tanto el solapamiento como el tipo de variante. Esta herramienta se puede ejecutar mediante el *framework* dotnet v6.0[90] o docker [87], aunque en este trabajo se utilizó una imagen disponible en el repositorio Docker Hub implementada con Singularity, que contiene el algoritmo preconfigurado.

Los archivos de entrada incluyeron los dos *gold standards* generados y los resultados individuales de cada algoritmo. Tanto los *gold standards* como los resultados fueron estandarizados al formato VCF para su uso.

La herramienta *witty.er* fue parametrizada con los siguientes valores para adaptarse a los datos generados en este estudio:

- **-percentDistance 1:** distancia proporcional máxima del tamaño de la variante para considerarlas equivalentes.
- **-bpDistance 10000:** distancia máxima entre los límites de variantes para considerarlas equivalentes, fijada en 10 kb.
- **-em SimpleCounting:** modo de evaluación simple por tipo de variante.

Esta parametrización permite realizar una validación permisiva, adecuada para este trabajo, dado que el *gold standard* no tiene definidos puntos de corte estrictos.

El conjunto de métricas reportadas por esta herramienta permiten evaluar el funcionamiento de los algoritmos de detección de CNVs en términos de precisión y sensibilidad. A continuación, se describen las métricas principales:

- ***QuerySample*** y ***TruthSample***: Identificadores del conjunto de variantes detectadas por un algoritmo (*Query*) y del *gold standard* (*Truth*) utilizados en la comparación.
- ***QueryTP (true positive)***: Número de variantes reportadas por el algoritmo que coinciden con variantes presentes en el *gold standard*.
- ***QueryFP (false positive)***: Número de variantes reportadas por el algoritmo que no coinciden con variantes presentes en el *gold standard*.
- ***Precision***: Proporción de variantes correctamente detectadas respecto al total de variantes reportadas por el algoritmo, calculada como:

$$\text{Precision} = \frac{\text{QueryTP}}{\text{QueryTP} + \text{QueryFP}}$$

- ***TruthTotal***: Número total de variantes presentes en el *gold standard*.
- ***TruthTP (true positive)***: Número de variantes presentes en el *gold standard* que coinciden con variantes reportadas por el algoritmo.
- ***TruthFN (false negative)***: Número de variantes presentes en el *gold standard* que no coinciden con variantes reportadas por el algoritmo.
- ***Recall***: Proporción de variantes correctamente detectadas respecto al total de variantes reales, definida como:

$$\text{Recall} = \frac{\text{TruthTP}}{\text{TruthTP} + \text{TruthFN}}$$

- ***F-score***: Métrica combinada que representa el equilibrio entre precisión y sensibilidad, calculada como:

$$\text{F-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.8. Disponibilidad del código

En el repositorio de GitHub [CNV-detection-validation-pipeline](#) están disponibles todos los datos de los estudios de referencia, junto con su clasificación, filtrado y fusión, así como los códigos utilizados para la estandarización, filtrado, fusión y generación del *gold standard*. También se incluyen los códigos desarrollados para el lanzamiento de cada algoritmo, la *pipeline* automatizada en Snakemake con sus correspondientes reglas y archivos de configuración, el *Snakefile* final, los resultados normalizados por algoritmo, la evaluación de los resultados frente al *gold standard* utilizando *witty.er*, y los scripts generados para generar las gráficas de representación de los resultados.

4 Resultados

4.1. Determinación del conjunto de referencia o *gold standard* para NA12878

En el apartado 3.3 se describieron los cuatro conjuntos de SVs utilizados en este estudio para generar un *gold standard* de la muestra NA12878. El objetivo es generar un conjunto de referencia de CNVs compuesto por aquellas variantes identificadas consistentemente en múltiples estudios, garantizando así una referencia sólida para la validación de los algoritmos de detección de CNVs.

Las variantes de los estudios se clasificaron en dos categorías según la categoría de filtrado de calidad asignada por los autores originales: *PASS* y *LowQuality*.

Mediante Python 3.12 [64] y la herramienta IGV [91], se realizó un filtrado de las variantes de cada categoría, eliminando aquellas irrelevantes para este estudio, así como aquellas potencialmente artefactuales o que pudieran introducir ruido en los resultados.

Finalmente, se generaron dos *gold standards* mediante la fusión de las variantes ya filtradas: un *gold standard pass* de alta confianza, que incluye solo las variantes *PASS*, y un *gold standard general* más permisivo, que incorpora tanto variantes *PASS* como *LowQuality*. En ambos casos, se incluyeron únicamente aquellas variantes presentes en al menos dos estudios tras la fusión.

El flujo de trabajo seguido en este apartado se ilustra en la Figura 4.1. En dicha imagen se presenta un diagrama de flujo que detalla cada etapa del proceso de filtrado y fusión, acompañada del número de variantes resultantes en cada apartado. En cada bloque del diagrama se indica cuántas variantes permanecen tras completar cada paso.

4.1.1. Filtrado genérico de variantes y estandarización de los estudios de referencia

Como primer filtrado genérico se realizaron las siguientes tareas:

- El estudio de los 1000 genomas incluye las SVs de los 2.504 individuos pertenecientes a la fase 3 del proyecto. Se generó un nuevo VCF que contiene únicamente la muestra NA12878 con sus correspondientes 3.260 variantes estructurales.
Este estudio no especifica la posición final de las 1.068 variantes filtradas como baja calidad o *LowQual*. Sin embargo, la variable *SVLEN* del archivo VCF registra la longitud de estas variantes, luego la posición final se calculó sumando la posición inicial con la longitud.
- El estudio de GIAB contiene 1.371 variantes localizadas en cromosomas alternativos, todas filtradas como calidad baja. Al ser el único estudio que contiene variantes cromosomas alternativos, y ser estas de baja confianza, se descartaron de los estudios posteriores, quedando 43.156 variantes en cromosomas principales.

Una vez eliminadas las variantes irrelevantes para este estudio, se procede a normalizar los cuatro estudios a un formato BED que incluye las siguientes columnas especificadas en la tabla 4.1.

Cromosoma	Inicio	Final	Tipo de SV	Longitud	Filtrado de calidad	Metodología de detección	Genotipo
-----------	--------	-------	------------	----------	---------------------	--------------------------	----------

Tabla 4.1: Columnas incluidas en el formato BED generado para estandarizar los estudios.

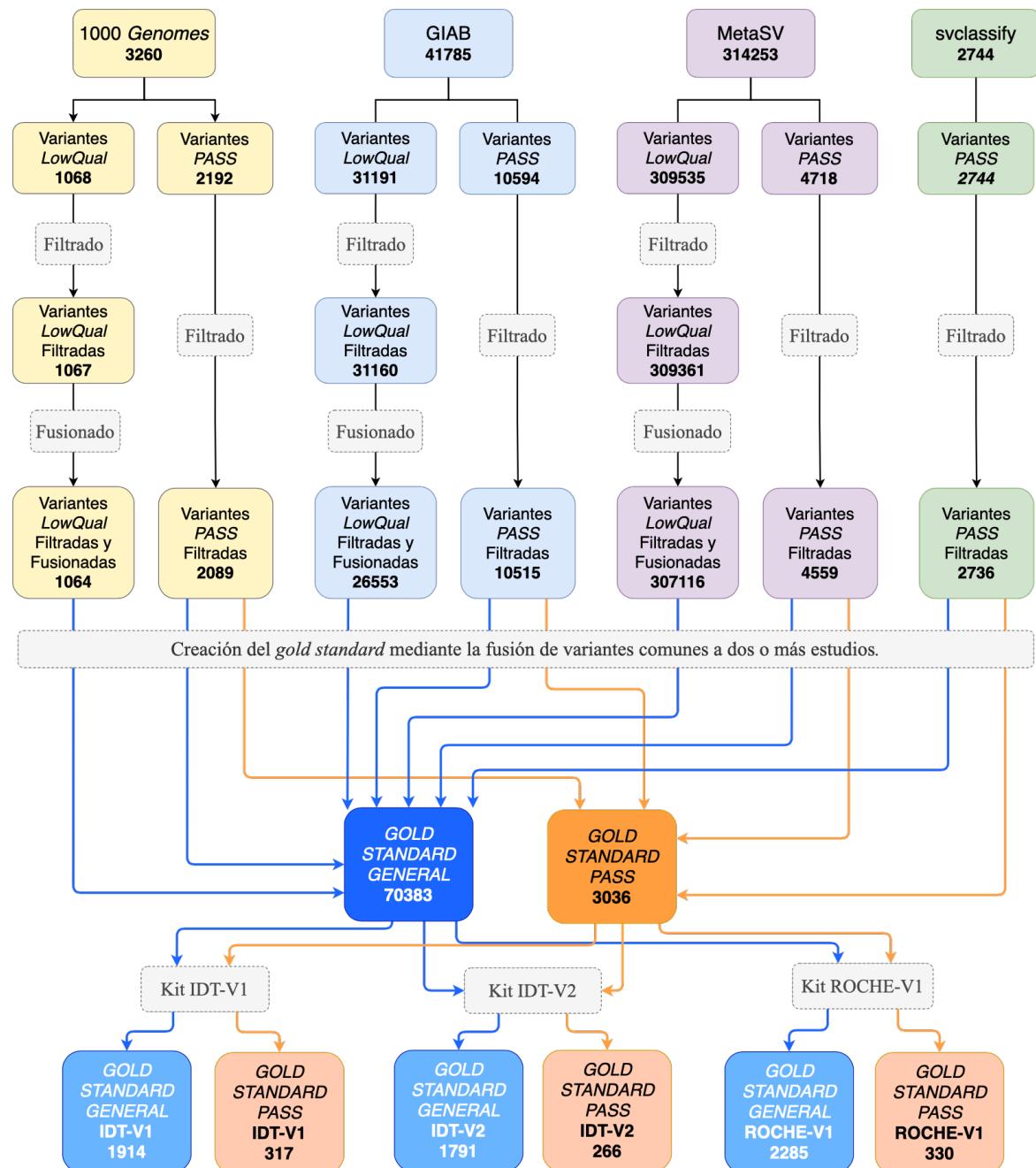


Figura 4.1: Diagrama de flujo que resume los pasos de filtrado y fusión de variantes desde los conjuntos de SVs iniciales hasta la generación de los *gold standard pass* y *gold standard general* restringidos a los kits de exoma. Cada bloque indica la cantidad de variantes de cada etapa.

4.1.2. División de los conjuntos en función de calidad de variantes

Con los cuatro estudios estandarizados a un mismo formato, es posible realizar comparaciones consistentes y fiables entre ellos.

Como primer análisis, las variantes se clasificaron según la categoría de filtrado de calidad asignada en los estudios originales. Esta clasificación divide cada estudio en dos grupos de variantes: variantes *PASS* y variantes *LowQual*.

La categoría *PASS* agrupa las variantes que superaron los filtros de calidad definidos por los autores de cada estudio, mientras que las *LowQual* fueron identificadas por las técnicas de detección de variantes de cada estudio pero no cumplieron con los criterios establecidos, lo que podría deberse a insuficiente profundidad de lectura, inconsistencias en la alineación o artefactos técnicos. En la tabla 4.2 se presentan las métricas descriptivas y la distribución de las longitudes para las variantes de cada categoría y estudio.

Categoría	Estudio	Cantidad	Longitud media	Desv. Est.	Mínimo	Mediana	Máximo
PASS	1000 Genomes	2.192	4.533,13	15.215,95	3	346	255.315
	GIAB	10.594	577,55	3.769,93	2	2	141.122
	MetaSV	4.718	9.008,19	101.506,40	1	271	2.378.892
	svclassify	2.744	1.322,55	4.801,9	12	323	139.619
LowQual	1000 Genomes	1.068	550,62	1.066,53	98	280	6.018
	GIAB	31.191	1.520,01	58.937,47	1	44	5.686.936
	MetaSV	309.535	1.292,76	171.952,5	1	35	73.872.940
	svclassify	-	-	-	-	-	-

Tabla 4.2: Métricas de longitud para variantes clasificadas como *PASS* y *LowQual* en pares de bases.

Los resultados detallados que nos aportan estas métricas son los siguientes:

- 1000 *Genomes* es el estudio más equilibrado entre el número de variantes *PASS* y *LowQual*. Las variantes *PASS* presentan mayor longitud media (4.533,13 bp), con una alta dispersión reflejada en la desviación estándar de 15.215,95 bp. Las variantes *LowQual* son significativamente más cortas, con la menor longitud máxima (6.018 bp) entre todos los estudios.
- GIAB representa el mayor conjunto de variantes *PASS* con 10.594 variantes, pero cuya mediana es extremadamente baja (2 bp), lo que indica que la mayoría de las variantes son de corta longitud. Las variantes *LowQual* son más largas (media de 1.520,01 bp) y presentan una mayor variabilidad. Además, su longitud máxima incluye valores extremos, con un máximo de 5.686.936 bp.
- En *MetaSV* destaca el gran número de variantes *LowQual*, con 309.535 variantes, y la extrema longitud de estas, con un máximo inusualmente grande de 73.872.940 bp. Representa la mayor longitud media, siendo de 9.008,19 bp en variantes *PASS*. Además, la desviación estándar de ambas categorías es extremadamente alta debido a las elevadas longitudes máximas de algunas variantes.
- Como se detalló en el apartado 3.3.3, *svclassify* reporta sus variantes sin información adicional y todas son variantes de alta confianza, luego no se dispone de variantes *LowQual*. Este conjunto representa unas métricas equilibradas con un rango de longitudes controlado.

4.1.3. Filtrado visual de variantes en función de su calidad

Una vez divididos los estudios según su categoría de calidad, se realizó un filtrado de los conjuntos *PASS* y *LowQual* por separado, con el objetivo de comparar ambos grupos, evaluar la consistencia entre estudios y detectar posibles limitaciones en las variantes de baja calidad.

En primer lugar se realizó el filtrado de las variantes *PASS*. Estas muestran una mayor uniformidad en sus métricas de longitud, con desviaciones estándar más bajas y rangos de longitudes más restringidos. Aun así, para minimizar la inclusión de variantes artefactualas en el *gold standard*, se cargaron los datos de los cuatro conjuntos de variantes *PASS* en IGV [91] y se llevó a cabo un filtrado visual.

Durante el visionado, se identificaron variantes solapantes dentro de cada estudio, es decir, un mismo estudio reportaba dos o más variantes en la misma región genómica. Estas variantes solapantes se visualizaron una a una y se observaron dos escenarios:

4 Resultados

- Las variantes solapantes presentan longitudes similares y corresponden al mismo tipo de variante (Figura 4.2). En estos casos, se seleccionó la variante de mayor longitud que abarcaba al resto, eliminando las redundantes. Este patrón es el más común en los estudios de 1000 Genomes, GIAB y svclassify.

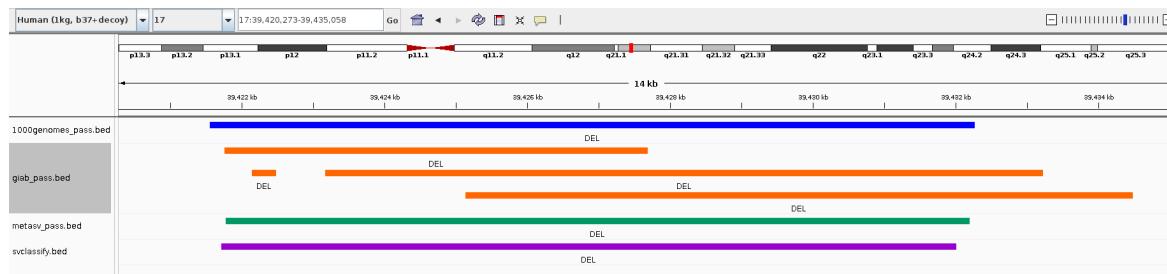


Figura 4.2: Ejemplo de variantes PASS solapantes identificadas en el estudio GIAB representado en IGV. Color azul para variantes de 1000 Genomes, naranja para GIAB, verde para MetaSV y morado para svclassify.

- Las variantes solapantes difieren significativamente en longitud y no coinciden en el tipo de variantes (Figura 4.3). Este escenario fue especialmente frecuente en las variantes reportadas por MetaSV. En estos casos, se eliminaron las variantes inconsistentes que no contaban con la evidencia de su existencia en los otros tres estudios.

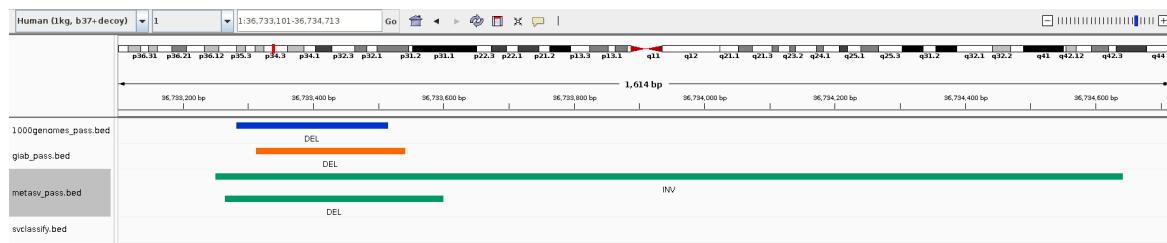


Figura 4.3: Ejemplo de variantes PASS solapantes inconsistentes en el estudio MetaSV representado en IGV. Color azul para 1000 Genomes, naranja para GIAB y verde para MetaSV.

Como resultado, se eliminaron 103 variantes de 1000 Genomes, 93 de GIAB, 119 de MetaSV y 8 de svclassify. Este filtrado redujo el ruido en los datos sin alterar significativamente las métricas de longitud en los estudios 1000 Genomes, GIAB y svclassify. Los cambios más significativos se observaron en MetaSV, donde la longitud media se reduce de 9.008,19 bp a 2.847,96 bp y la desviación estándar baja considerablemente.

Las métricas descriptivas de las variantes PASS filtradas se actualizan en la tabla 4.3.

Categoría	Estudio	Cantidad	Longitud media	Desv. Est.	Mínimo	Mediana	Máximo
PASS	1000 Genomes	2.192	4.533,13	15.215,95	3	346	255.315
	GIAB	10.594	577,55	3.769,93	2	2	141.122
	MetaSV	4.718	9.008,19	101.506,40	1	271	2.378.892
	svclassify	2.744	1.322,55	4.801,90	12	323	139.619
Filtradas	1000 Genomes	2.089	4.144,14	14.799,18	3	339	255.315
	GIAB	10.501	565,50	3.649,39	2	2	141.122
	MetaSV	4.599	2.847,96	36.656,23	0	277	1.935.684
	svclassify	2.736	1.322,11	4.807,28	12	323	139.619

Tabla 4.3: Métricas de longitud para variantes PASS antes y después del filtrado en pares de bases.

4 Resultados

A continuación se procedió con el filtrado de las variantes *LowQual*. Aunque estas variantes no son necesariamente falsos positivos, la presencia de variantes extremadamente grandes, la abundancia de variantes pequeñas y las altas desviaciones estándar sugieren que estas variantes son menos consistentes y podrían incluir falsos positivos o artefactos.

El gran número de variantes *LowQual* en GIAB y, especialmente, en MetaSV con 309.535 variantes, dificulta el filtrado visual y manual para identificar y eliminar posibles artefactos. Ante esta situación, y debido a la presencia de variantes extremadamente grandes en estos estudios, se optó por un filtrado visual genérico en IGV [91] únicamente de aquellas variantes mayores de 1 kb. Durante este proceso, se descartaron las variantes que aparecían únicamente en un estudio y no estaban respaldadas por variantes similares en los otros estudios, como se ilustra en la Figura 4.4.

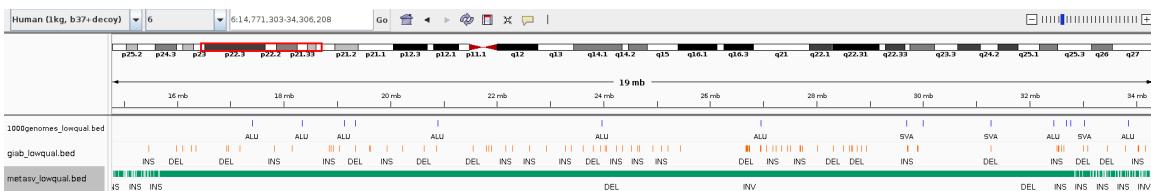


Figura 4.4: Ejemplo de variante *LowQual* de tamaño inusual (17 Mb) en el estudio MetaSV. Color azul para variantes de 1000 Genomes, naranja para GIAB y verde para MetaSV.

El filtrado visual también confirmó la presencia de múltiples variantes solapantes en los conjuntos *LowQual*. A diferencia de las variantes *PASS*, que mostraban solapamientos coherentes, las variantes *LowQual* se caracterizaron por una abundancia de solapamientos discordantes, lo que complicó aún más el análisis. En particular, el estudio GIAB destacó por su alto nivel de redundancia entre variantes, como se ilustra en la Figura 4.5.

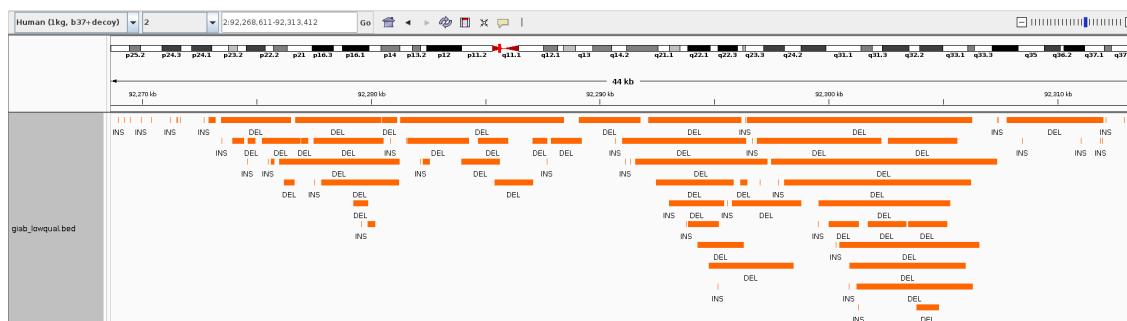


Figura 4.5: Ejemplo de múltiples variantes *LowQual* solapantes redundantes reportadas por GIAB.

Dado el elevado número de variantes *LowQual*, se optó por realizar un paso intermedio utilizando la herramienta BEDTools Merge, descrita en el apartado 3.4, para fusionar las variantes solapantes dentro de cada estudio. Aunque este enfoque elimina variantes redundantes y evita la contabilización de duplicados, no es una práctica ideal, ya que puede introducir un alto nivel de ruido o extender de manera imprecisa los puntos de corte de las variantes fusionadas.

Tras el filtrado manual y la fusión de las variantes solapantes de cada estudio mediante BEDTools Merge, se actualizan las métricas de longitud en la tabla 4.4.

Las métricas resultantes tras el filtrado y la fusión presentan longitudes y desviaciones estándar más consistentes, acercándose a las variantes *PASS*.

Se observa una reducción significativa en la cantidad de variantes de MetaSV tras el filtrado manual, pero la fusión tiene poco impacto en el número de variantes finales. En cambio, la fusión en el estudio de GIAB reduce considerablemente el número de variantes, lo que confirma la presencia de muchas variantes solapadas redundantes en la misma región, como se observó previamente.

4 Resultados

Categoría	Estudio	Cantidad	Longitud media	Desv. Est.	Mínimo	Mediana	Máximo
LowQual	1000 Genomes	1.068	550,62	1.066,53	98	280	6.018
	GIAB	31.191	1.520,01	58.937,47	1	44	5.686.936
	MetaSV	309.535	1.292,76	171.952,50	1	35	73.872.940
	svclassify	-	-	-	-	-	-
LowQual Filtradas	1000 Genomes	1.067	550,62	1.066,54	98	278	6.018
	GIAB	31.160	572,53	5.714,88	1	43	330.470
	MetaSV	309.361	115,88	1.912,32	1	35	360.609
	svclassify	-	-	-	-	-	-
LowQual filtradas y fusionadas	1000 Genomes	1.064	551,12	1.067,97	98	278	6.018
	GIAB	26.253	597,25	6.325,27	1	40	330.470
	MetaSV	307.116	115,44	1.911,65	1	36	360.609
	svclassify	-	-	-	-	-	-

Tabla 4.4: Métricas de longitud para variantes *LowQuality*, filtradas y fusionadas en pares de bases.

Las variantes de 1000 *Genomes* se mantienen estables a lo largo del filtrado y la fusión. Sin embargo, como se señala en la figura 5 del Anexo 2, este conjunto solo incluye inserciones, lo cual supone una limitación en la detección de delecciones.

La distribución de los tipos de variantes, clasificados por rango de tamaño, para las variantes *PASS* y *LowQual* filtradas, se representa en los gráficos de barras 1-7 del Anexo 2.

4.1.4. Análisis comparativo y fusión de los estudios con las variantes finales

Tras la categorización de las variantes *PASS* y *LowQual* y su correspondientes filtrados, se procedió a fusionar los estudios con el objetivo de identificar aquellas variantes reportadas en varios estudios, ya que su recurrencia aumenta su fiabilidad y contribuye a construir un *gold standard* más robusto.

En este proceso, se generaron dos tipos de fusiones, según la calidad de las variantes:

- **Fusión *PASS*:** Incluye únicamente las variantes *PASS* ya filtradas.
- **Fusión general:** Combina variantes *PASS* y *LowQual*, previamente filtradas.

De este modo, se generan dos versiones del *gold standard*: una versión basada en las variantes *PASS* garantizando un conjunto de alta calidad, y una versión general más amplia e inclusiva con variantes *PASS* y *LowQual*. Aunque las variantes *LowQual* tienen limitaciones en calidad, su inclusión permite considerar variantes que podrían ser relevantes.

Para llevar a cabo estas fusiones, se realizó un análisis en paralelo utilizando las herramientas BEDTools Multiinter y BEDTools Merge explicadas en el apartado 3.4 de forma complementaria, ya que, tras su uso, se comprobó que ambas tienen fortalezas y limitaciones en el análisis:

- **BEDTools Multiinter:** Identifica a qué estudios pertenece cada variante fusionada y aplica criterios rígidos de intersección (mínimo 50 % de solapamiento) con puntos de corte estrictos. Su principal limitación es que no conserva la información sobre el tipo de variante.
- **BEDTools Merge:** Genera variantes fusionadas más amplias, maximizando puntos de corte. Además, conserva el tipo de variante, lo que es crucial para la validación de algoritmos.

Mediante estas herramientas es posible determinar si una variante está presente en dos, tres o en los cuatro estudios analizados. Estas situaciones se representan con IGV[91] en la figura 4.6.

Esto permite clasificar las variantes en los siguientes tres conjuntos, los cuales facilitarán la generación del *gold standard*:

1. Variantes comunes en dos o más estudios: aquellas variantes que cumplen con el criterio de solapamiento mínimo entre al menos dos estudios.
2. Variantes comunes en tres o más estudios: variantes compartidas por al menos tres estudios.
3. Variantes comunes en los cuatro estudios: variantes presentes en todos los estudios.

4 Resultados

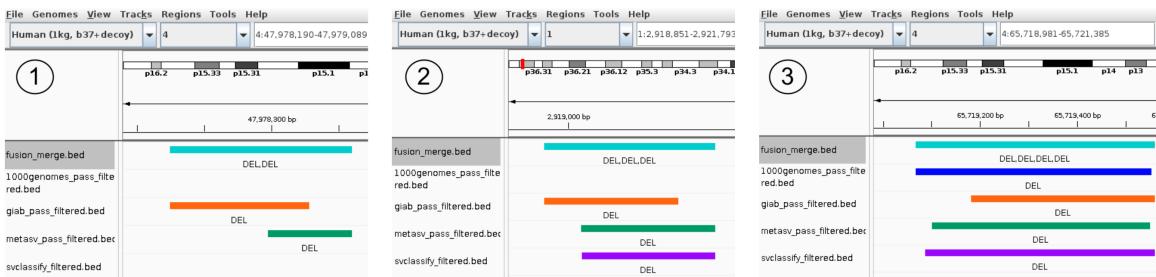


Figura 4.6: Ejemplo de variantes comunes en: 1. dos estudios, 2. tres estudios, 3. cuatro estudios.
Color azul claro para la fusión realizada con BEDTools Merge. Color azul oscuro para el estudio 1000 Genomes, naranja para GIAB y verde para MetaSV.

Fusión de las variantes PASS

Primero, se fusionaron las variantes PASS ya filtradas de los cuatro estudios. Utilizando BEDTools Multiinter, se identificó a qué estudios pertenece cada variante fusionada. Con ello, se calcularon todas las combinaciones posibles de presencia de variantes en los estudios. La distribución de estas combinaciones se presenta en la figura 4.7.

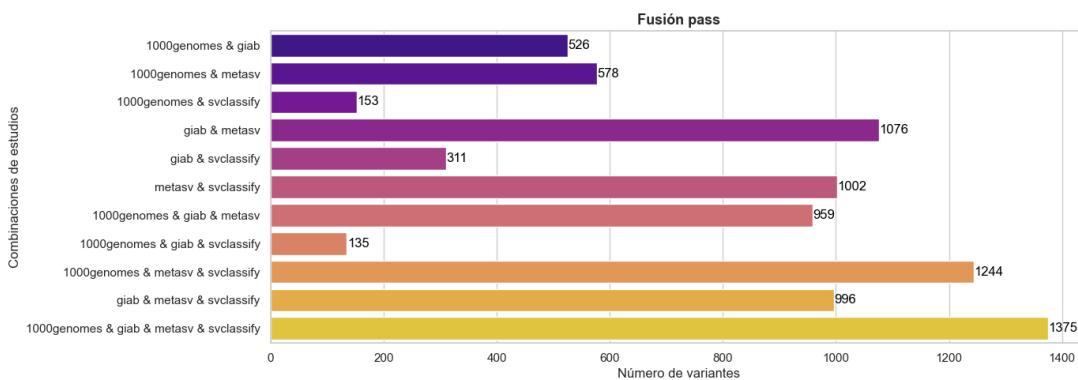


Figura 4.7: Distribución del número de variantes PASS para cada una de las combinaciones posibles entre los estudios: variantes solapantes en dos estudios, tres, o en los cuatro.

Se observa que svclassify es el conjunto que presenta la menor cantidad de variantes comunes con los demás, excepto con MetaSV, ya que ambos provienen del mismo estudio, como se explicó en el apartado 3.3.3. Esto se refleja aún más en la combinación de los tres estudios: 1000 Genomes, GIAB y svclassify, con solo 135 variantes comunes.

Por otro lado, MetaSV y GIAB son los estudios que comparten más variantes (1.076 SVs), lo cual es consistente al ser los estudios que más variantes PASS reportan. La mejor combinación de tres estudios la componen 1000 Genomes, MetaSV y svclassify, con 1.244 variantes comunes.

En paralelo, se utilizó la herramienta BEDTools Merge, que, aunque genera variantes más generales y menos precisas, permite conservar la información sobre el tipo de variante fusionada.

La figura 4.8 muestra la distribución de los tipos de variantes comunes en dos o más estudios, tres o más y las variantes que aparecen en los cuatro estudios. La categoría *indeterminado* corresponde a variantes cuya clasificación no coincide entre los estudios, como aquellas que resultan de la fusión de una delección y una inversión. Estas variantes serán analizadas con cautela.

Los resultados de las fusiones son muy positivos, al apenas haber variantes clasificadas como indeterminadas. Sin embargo la mayoría son delecciones, con una presencia muy baja de inserciones e inversiones, e incluso nula en las variantes reportadas por los cuatro estudios.

4 Resultados



Figura 4.8: Distribución del número de variantes *PASS* según su tipo y su presencia en combinaciones de dos o más estudios, tres o más estudios, o en los cuatro.

Fusión general de las variantes *PASS* y *LowQual*

De la misma forma, se repitió el análisis generando la fusión general, que incluye variantes *PASS* y *LowQual* previamente filtradas en cada estudio. Se utilizó BEDTools Multiinter para determinar a qué estudios pertenece cada variante fusionada, obteniendo todas las combinaciones posibles entre los estudios, representadas en la figura 4.9.

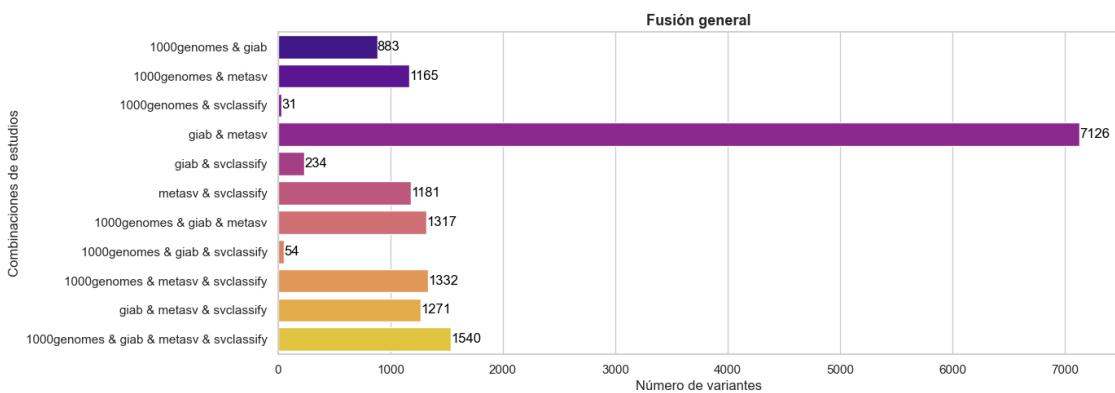


Figura 4.9: Distribución del número de variantes *PASS* y *LowQual* para cada una de las combinaciones posibles entre los estudios: variantes solapantes en dos estudios, tres, o en los cuatro.

Se observa de nuevo, aunque más diferenciado, que los dos estudios con más variantes comunes son GIAB y MetaSV. Esto tiene sentido al ser los estudios que más variantes reportan. El número de variantes reportadas en las combinaciones de tres estudios es ligeramente mayor en la fusión general que en la fusión *pass*, aunque sin cambios muy significativos. En cuanto a las variantes comunes en los cuatro estudios, se observan 1.540 variantes, con 165 adicionales en comparación con las variantes *PASS* comunes a los cuatro estudios.

A su vez, se generaron intersecciones con BEDTools Merge para conocer como se distribuyen los tipos de variantes presentes en dos o más estudios, en tres y cuatro estudios, y en los cuatro estudios. Los resultados se presentan en la figura 4.10.

En este análisis se aprecia un notable sesgo debido al funcionamiento de BEDTools Merge. Las fusiones generales, que incluyen variantes *LowQual*, generan un número significativamente mayor de variantes indeterminadas, como se observa al comparar las figuras 4.8 y 4.10:

4 Resultados

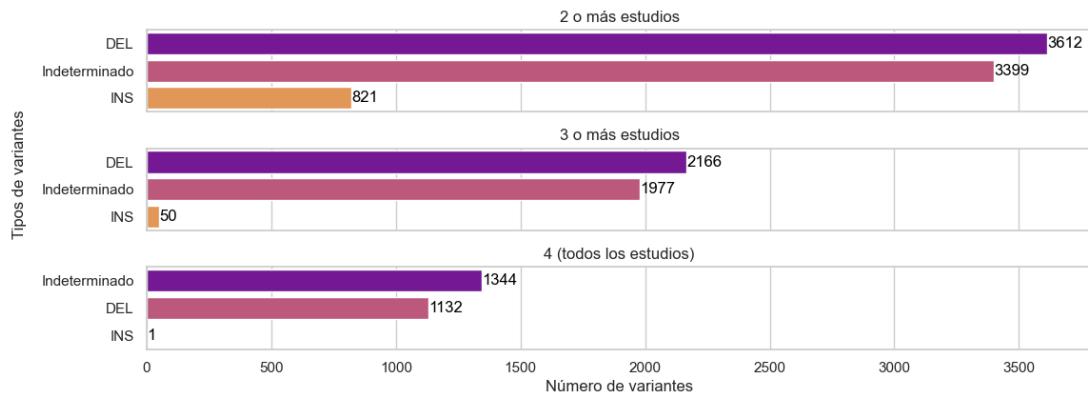


Figura 4.10: Distribución del número de variantes *PASS+LowQual* según su tipo y su presencia en combinaciones de dos o más estudios, tres o más estudios, o en los cuatro.

- En las variantes presentes en dos estudios, se pasa de 112 variantes indeterminadas en la fusión *pass* a 3.399 en la fusión general.
- En las variantes de tres estudios, el número aumenta de 66 a 1.977.
- En las variantes comunes a los cuatro estudios, se reportan 1.344 variantes indeterminadas frente a las 33 de la fusión *pass*. En este último caso, incluso hay más variantes indeterminadas que determinadas.

Con respecto a las variantes determinadas, se consiguió aumentar el número de delecciones e inserciones óptimamente solapadas en dos o más estudios. Sin embargo, para las fusiones de tres o más estudios y cuatro, se disminuye el número de delecciones reportadas.

4.1.5. Gold standard de variantes para NA12878

Una vez realizadas las fusiones generales con las variantes *PASS* y *LowQual*, y las fusiones *pass* únicamente con las variantes *PASS*, se procedió a la creación de los *gold standards*. Para ello se escogieron aquellas variantes presentes en dos o más estudios con longitud mayor a 50 bp obtenidas con la herramienta BEDTools Merge, ya que para la validación de los algoritmos es estrictamente necesario el conocimiento del tipo de variante. Con ello, se crearon dos *gold standards*:

- **Gold standard pass:** Contiene 3.036 variantes *PASS* presentes en dos o más estudios. Durante un último filtrado manual, se consiguió especificar el tipo de 97 variantes clasificadas como indeterminadas, 15 variantes fueron eliminadas debido a la imposibilidad de deducir su tipo y 25 inversiones se eliminaron, al solo querer generar un *gold standard* de CNVs.
- **Gold standard general:** Contiene las 7.383 variantes *PASS + LowQual* presentes en dos o más estudios obtenidas con BEDTools Merge. Al haber 3.399 variantes reportadas como indeterminadas, se imposibilita hacer un filtrado manual, por lo que el tipo de variante se le asignó en función al tipo de variante más común en cada fusión. 16 variantes fueron descartadas debido al alto nivel de ruido que presentaban y 20 inversiones se eliminaron.

Dado que la mayoría de los estudios no incluyen duplicaciones como categoría explícita, las inserciones se considerarán equivalentes a duplicaciones en los *gold standards*. Esto se basa en la posibilidad de que algunas inserciones detectadas en los estudios correspondan a duplicaciones [92].

Ambos conjuntos contienen variantes tanto en regiones exónicas como intrónicas, ya que los estudios iniciales generaron sus conjuntos de variantes mediante técnicas WGS.

En este trabajo se utilizan datos WES de la muestra NA12878, generados con tres kits comerciales de captura de exoma (IDT-V1, IDT-V2 y ROCHE-V1). Para adaptar los *gold standards* a estas técnicas,

se utilizó la herramienta BEDTools Intersect para seleccionar únicamente las variantes de los *gold standards* presentes en las regiones exónicas capturadas por cada kit. Estas regiones están especificadas en los archivos BED proporcionados por las casas comerciales.

De esta forma, se generaron tres subconjuntos específicos para cada *gold standard* (*pass* y general), dado que los kits de exoma de las distintas casas comerciales cubren regiones genómicas diferentes. Las variantes y sus tipos obtenidas en cada intersección se especifican en la tabla 4.5.

<i>Gold Standard</i>	Total	IDT-V1		IDT-V2		ROCHE-V1	
General	7.383	DEL - 6.558	1.914	DEL - 1.451	1.791	DEL - 1.366	2.285
		DUP - 825		DUP - 463		DUP - 425	DEL - 1.760 DUP - 525
Pass	3.036	DEL - 2.980	317	DEL - 302	266	DEL - 253	330
		DUP - 56		DUP - 15		DUP - 13	DEL - 316 DUP - 14

Tabla 4.5: Distribución de los *gold standards* finales tras su intersección con las regiones exónicas capturadas por cada kit de captura. Se presentan las variantes totales y la separación por tipo DEL y DUP.

Al realizar la intersección con los BEDs, se pierde un número significativo de variantes en ambos casos (general y *pass*). Además, los nuevos *gold standards* resultantes difieren considerablemente entre sí. Se observa que el kit de exoma que captura menos variantes es IDT-V2, mientras que ROCHE-V1 es el que reporta un mayor número. Las métricas de longitud de las variantes resultantes en cada *gold standard* adaptado a las técnicas WES se especifican en la tabla 1 del Anexo 2.

Cada uno de estos nuevos *gold standards* se utilizó para la validación de los resultados de los algoritmos, asignando los *gold standard* IDT-V1 para aquellos resultados obtenidos de las réplicas 1 y 2, los *gold standard* IDT-V2 para las réplicas 3 y 4 y los *gold standard* ROCHE-V1 para la réplica 5. Esta asignación asegura que cada réplica valide únicamente las variantes que caen dentro de las regiones cubiertas por el kit de captura utilizado.

4.2. Evaluación de las réplicas y los análisis bioinformáticos

Para capturar la variabilidad técnica y metodológica se analizaron cinco réplicas de la muestra, secuenciadas con diferentes kits de captura y procesadas mediante tres *pipelines* bioinformáticos. Esto generó un total de 15 conjuntos de datos de entrada por algoritmo, obteniéndose en consecuencia 15 resultados independientes para cada uno.

Como primer paso, se estandarizó el formato de salida de cada algoritmo a un formato común VCF v4.2 mediante el uso de Python 3.12, ya que cada algoritmo generaba sus resultados en un formato distinto. La mayoría de los algoritmos reportaron variantes de tipo delecciones (DEL) y duplicaciones (DUP), menos el algoritmo Manta [82], el cual reportaba delecciones e inserciones (INS). Para evaluar su funcionamiento, se optó por interpretar las inserciones como duplicaciones, manteniendo así la coherencia entre los resultados y permitiendo su comparación [92].

Para validar los resultados ya estandarizados, se utilizaron los dos *gold standards* (general y *pass*) previamente generados en el apartado 4.1.5. La comparación entre los resultados de los algoritmos y estos *gold standards* se llevó a cabo mediante la herramienta witty.er v0.5.2 desarrollado en el apartado 3.7, el cual reporta las métricas de *precision*, *recall* y *F-score* tanto para el conjunto total de variantes (*overall*) como de forma desglosada por tipo de evento (DEL o DUP en este caso).

Los resultados se representaron mediante gráficos tipo *boxplots* y gráficos de líneas para facilitar la comparación entre las distintas casuísticas evaluadas. Los *boxplots* permiten visualizar la distribución de las métricas de *precision*, *recall* y *F-score*, destacando la variabilidad y los valores extremos. Además, permiten representar los resultados según kits de captura, réplicas, algoritmos o *gold standards*. Los *lineplots* muestran como varían métricas promedio de *precision* y *recall* para los resultados que se estén representando, obteniendo una visión general pero representativa.

4.2.1. Análisis de los kits de captura de exoma

En primer lugar, se evaluaron las métricas de *precision*, *recall* y *F-score* en función del kit de exoma.

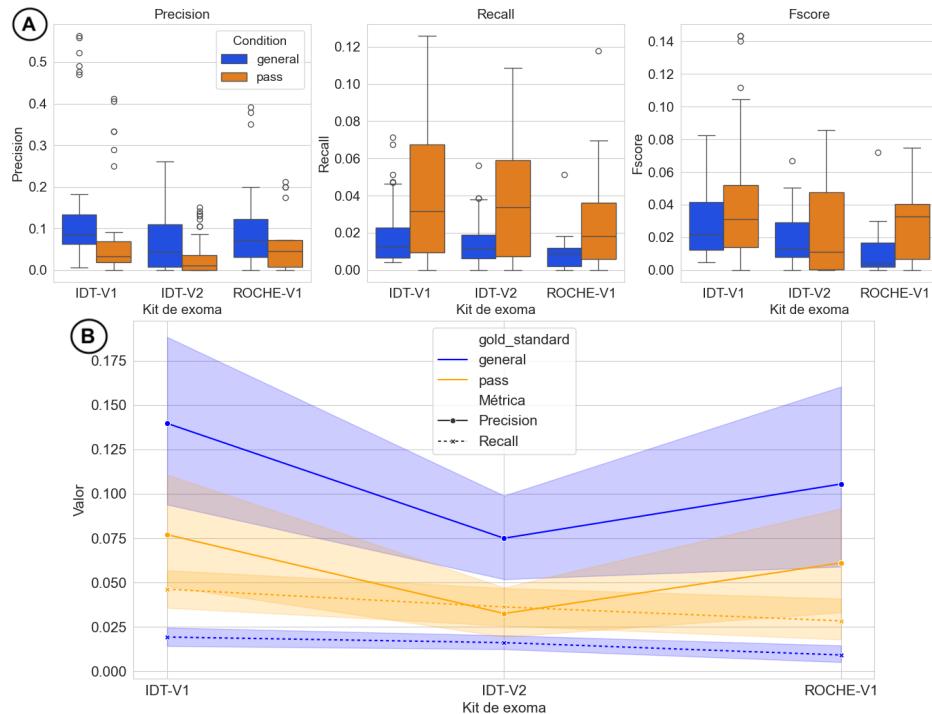


Figura 4.11: A. Boxplots que representan la distribución de las tres métricas para cada kit de exoma. B. Lineplots que comparan las métricas de *precisión* (líneas rectas) y *recall* (líneas punteadas) para cada kit. En azul el *gold standard general*, en naranja el *gold standard pass*.

En la Figura 4.11 se observa que, en general, el rendimiento de los kits de captura es limitado. IDT-V1 destaca con los mayores valores de precisión y sensibilidad para ambos *gold standards*. Por otro lado, IDT-V2 muestra las precisiones más bajas, especialmente con el *gold standard pass*, aunque su sensibilidad relativamente alta compensa parcialmente su rendimiento.

Mientras que ROCHE-V1 alcanza buenos valores de precisión, su sensibilidad es la más baja entre los tres kits, especialmente con el *gold standard general*, lo que afecta negativamente su rendimiento global.

Al comparar entre los dos *gold standards*, se visualiza un claro patrón en todos los kits de captura: la precisión es ligeramente mayor con el *gold standard general*, mientras que la sensibilidad aumenta significativamente con el *gold standard pass*.

4.2.2. Análisis de las réplicas de la muestra NA12878

La cobertura de secuenciación influye directamente en la precisión de la detección de variantes. Coberturas bajas pueden limitar la identificación de variantes, mientras que coberturas altas generan ruido y aumentan los costos [93]. En la tabla 4.2 se reportaron las métricas de calidad y cobertura media de cada réplica en función del análisis bioinformático implementado. En ella, se observa una cobertura excesivamente alta en la réplica 4. Esto desemboca en la peor métrica de eficiencia total en las tres *pipelines*. También destaca la baja cobertura en la réplica 5, lo que podría limitar la precisión en la detección de variantes. Pese a su baja cobertura, esta réplica presenta buenas métricas de calidad y la mayor eficiencia total.

4 Resultados

Posteriormente, se evaluaron las métricas de *precisión*, *recall* y *F-score* de cada réplica individual, con el objetivo de identificar posibles variaciones debidas a las diferencias en las condiciones de secuenciación y los kits de captura.

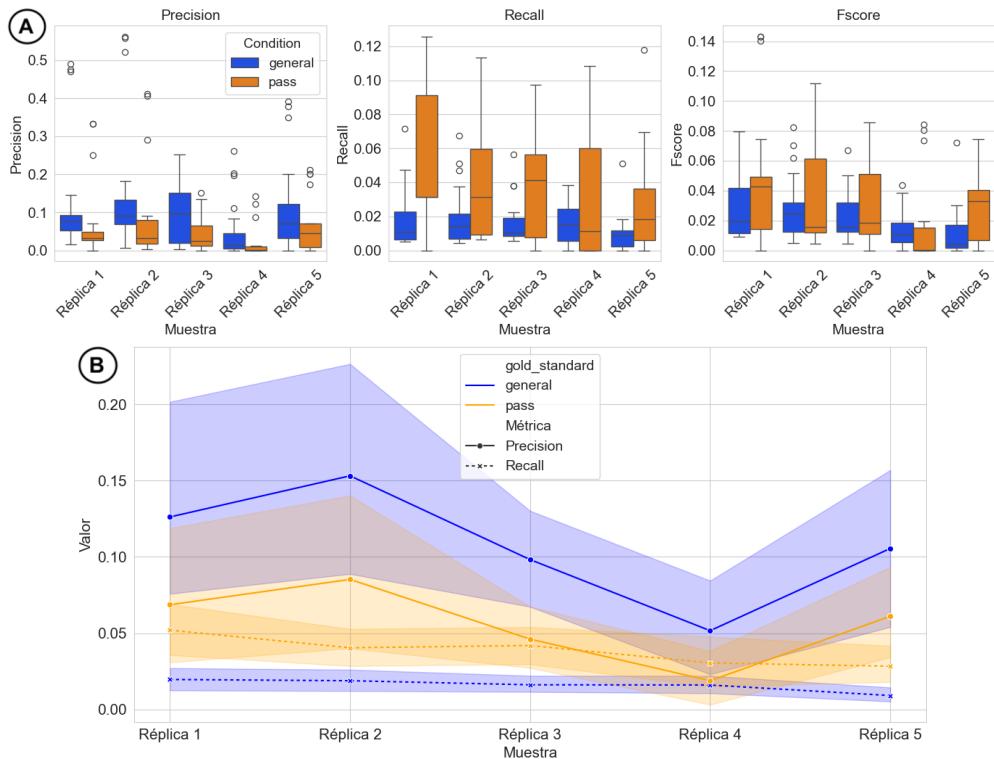


Figura 4.12: A. Boxplots que representan la distribución de las tres métricas para cada réplica. B. Lineplots que comparan las métricas de *precisión* (líneas rectas) y *recall* (líneas punteadas) para cada réplica. En azul el *gold standard general*, en naranja el *gold standard pass*.

En la Figura 4.12 se observa nuevamente las bajas métricas de calidad obtenidas en la evaluación de los resultados. Las réplicas 1 y 2, secuenciadas con el kit IDT-V1, presentan las mayores precisiones entre todas las muestras, especialmente la réplica 2. Sin embargo, la sensibilidad, aun siendo ligeramente superior para la réplica 1, es generalmente muy baja en todas las situaciones.

La réplica 3, secuenciada con IDT-V2 no reporta métricas tan desfavorables. Sin embargo, la réplica 4, presenta una precisión significativamente baja, e incluso nula en el *gold standard pass*. Esto puede estar directamente relacionado con su alta cobertura.

Finalmente, la réplica 5, secuenciada con ROCHE-V1 y con una cobertura especialmente baja, muestra valores de *precisión* relativamente buenos, pero sus bajos valores en sensibilidad afectan negativamente a su eficiencia global.

De nuevo, el *gold standard general* tiende a generar una *precisión* mayor en las cinco réplicas, pero la sensibilidad reportada en todas ellas es especialmente baja. Con el *gold standard pass* se genera un incremento en la sensibilidad, lo que contribuye a mayores valores de *F-score*, especialmente en las réplicas 1.

4.2.3. Análisis de las pipelines bioinformáticas

A su vez, se evaluaron las métricas en función de los tres análisis bioinformáticos utilizados en cada grupo de muestras, desarrollados en el apartado 3.2.

4 Resultados

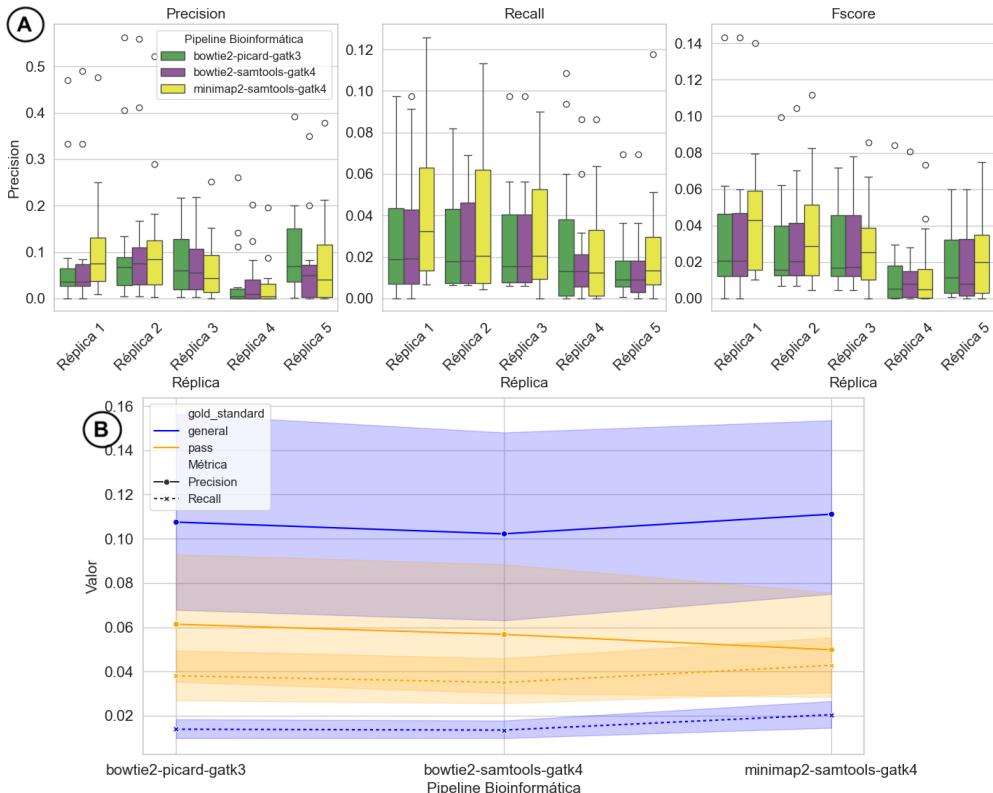


Figura 4.13: A. Boxplots que representan la distribución de las tres métricas para cada pipeline. B. Lineplots que comparan las métricas de *precisión* (líneas rectas) y *recall* (líneas punteadas) para cada pipeline. En azul el *gold standard general*, en naranja el *gold standard pass*.

Las gráficas de la figura 4.13 muestran que, aunque las *pipelines* reportan métricas muy similares, se pueden observar pequeñas diferencias entre ellas.

Bowtie2-Picard-GATK3 presenta métricas de precisión ligeramente más bajas en las réplicas 1 y 2. Sin embargo, destaca positivamente en precisión en la réplica 5, y ligeramente en la réplica 3 respecto a las otras *pipelines*. Bowtie2-SAMTools-GATK4 presenta una ligera mejora en precisión para las réplicas 1, 2 y 4 con respecto a la *pipeline* anterior. Estas dos primeras *pipelines* reportan sensibilidades muy parecidas en la mayoría de los casos.

En contraste, la *pipeline* Minimap2-SAMTools-GATK4 ofrece métricas superiores tanto en precisión como en sensibilidad en la mayoría de las réplicas. Esta *pipeline* destaca particularmente en las réplicas 1 y 2, donde mejora significativamente los resultados obtenidos.

A nivel genérico, destacan los resultados de precisión significativamente desfavorables obtenidos en la réplica 4 en las tres *pipelines*. También destaca la réplica 5 con los valores más bajos en sensibilidad para las tres *pipelines*.

En cuanto al análisis por *gold standard*, las *pipelines* Bowtie2-Picard-GATK3 y Bowtie2-SAMTools-GATK4 muestran un rendimiento similar en precisión y sensibilidad en ambos *gold standards*. Por otro lado, Minimap2-SAMTools-GATK4, aunque mejora notablemente en sensibilidad en ambos *gold standards*, muestra una mejora en precisión con el *gold standard general*, pero una precisión ligeramente inferior en el *gold standard pass* respecto a las otras dos *pipelines*.

4.3. Evaluación de los algoritmos

Una vez analizadas todas las variables que se han usado para obtener un conjunto de datos más variado y completo, capaz de abarcar diversas condiciones experimentales y técnicas, se procede a presentar los resultados obtenidos por los algoritmos. Para ello, se evalúan las métricas de *precisión*, *recall* y *F-score* de cada algoritmo de detección de CNVs implementados en este trabajo.

Estos algoritmos están diseñados para detectar variantes de tipo delección o duplicación.

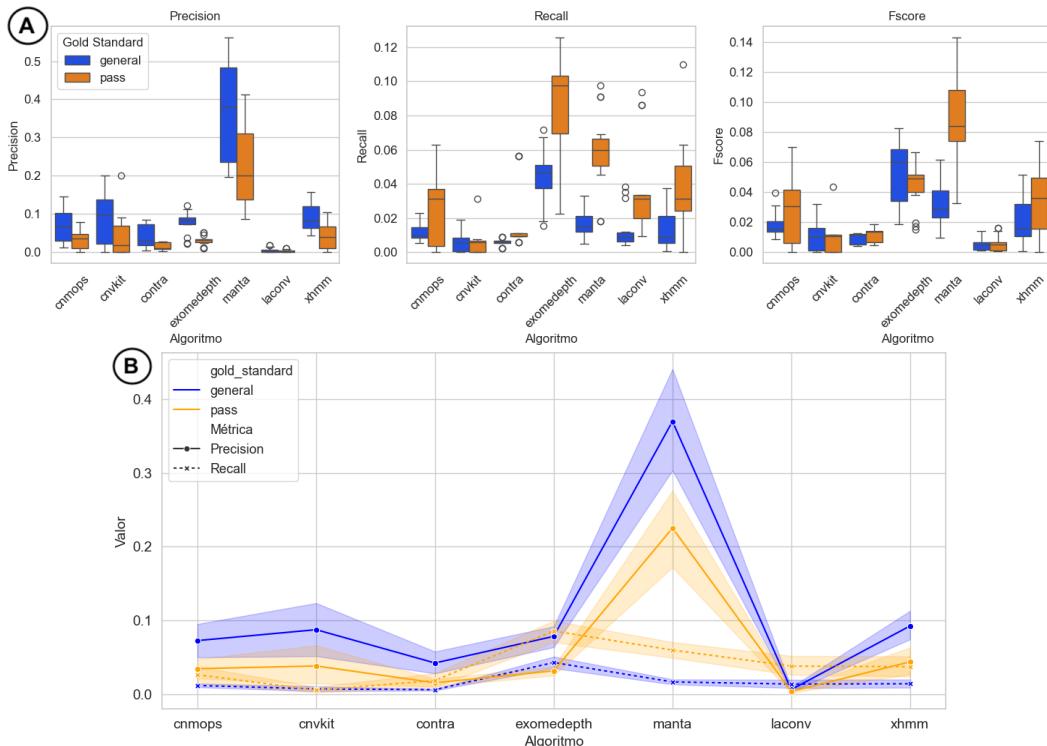


Figura 4.14: A. Boxplots que representan la distribución de las tres métricas para cada algoritmo. B. Lineplots que comparan las métricas de *precisión* (líneas rectas) y *recall* (líneas punteadas) para cada algoritmo. En azul el *gold standard general*, en naranja el *gold standard pass*.

Como primer análisis, la Figura 4.14 muestra diferencias significativas en el rendimiento de los algoritmos evaluados. Manta destaca como el algoritmo con las precisiones más altas en ambos *gold standards*, superando ampliamente al resto. Su *recall* es bajo, aunque presenta una mejora notable en el *gold standard pass*, lo que contribuye a un buen eficiencia global en este escenario.

El resto de los algoritmos muestran rangos de precisión más uniformes, aunque considerablemente inferiores a los reportados por Manta. ExomeDepth sobresale en términos de sensibilidad, especialmente en el *gold standard pass*, donde alcanza los valores más altos del estudio. Esto genera buenos valores de *F-score* en ambos *gold standards*.

cn.MOPS, CNVkit y XHMM presentan valores similares de precisión, pero difieren significativamente en sensibilidad. Entre ellos, XHMM alcanza los valores más altos de *recall*, mientras que CNVkit muestra métricas de sensibilidad notablemente bajas.

CONTRA tiene una eficiencia limitada, con valores bajos tanto de *precision* como de *recall*, especialmente en el *gold standard pass*. Finalmente, LACONv registra el peor rendimiento en términos de precisión, con valores cercanos a 0. Aunque mejora ligeramente en *recall* para el *gold standard pass*, su detección de variantes sigue siendo muy limitada.

Al comparar los resultados entre los dos *gold standards*, se mantiene la tendencia observada

4 Resultados

anteriormente: el *gold standard general* genera valores más altos de precisión, mientras que el *gold standard pass* favorece una mayor sensibilidad y significancia en la mayoría de los casos.

En la Figura 4.15 y sin hacer cuenta del *gold standard*, se desglosan las métricas de *precision* y *recall* reportadas por cada algoritmo en base a la *pipeline* de análisis utilizada y a cada réplica.

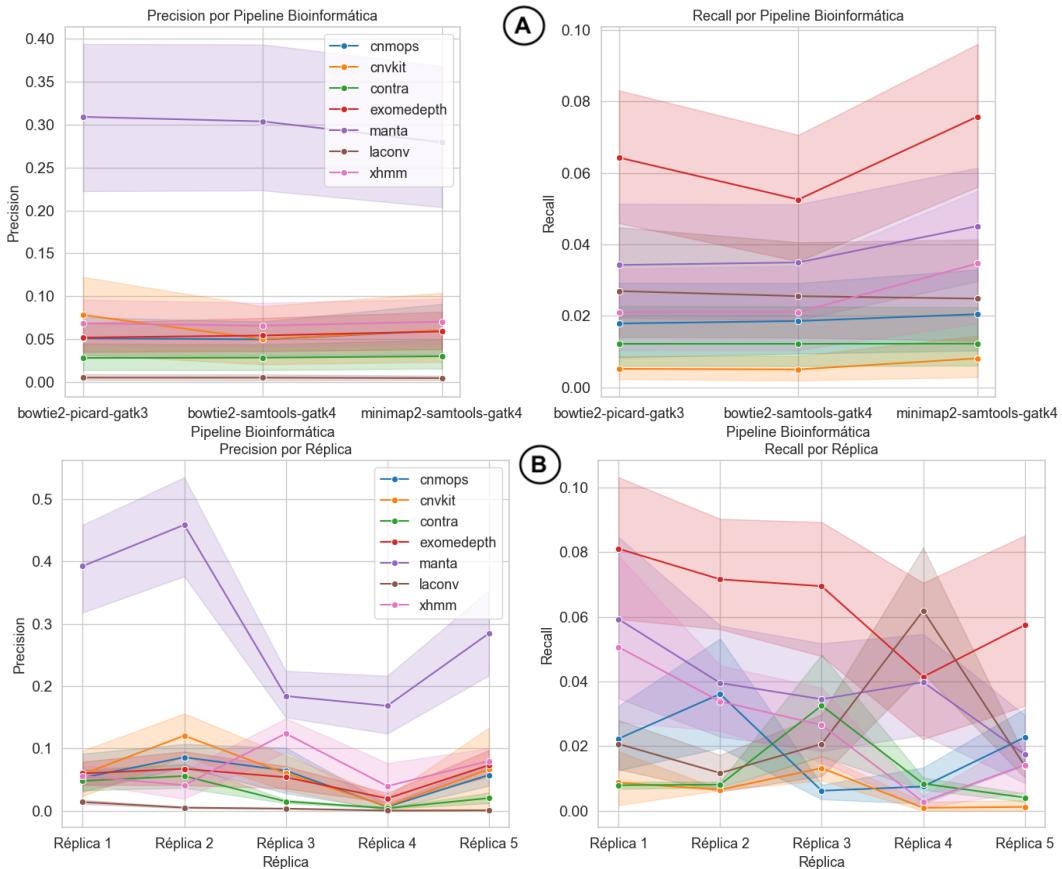


Figura 4.15: Lineplots que comparan las métricas de *precision* y *recall* para los algoritmos de detección de CNVs según: A. El análisis bioinformático empleado y B. La réplica analizada.

Las diferencias entre los análisis bioinformáticos en términos de precisión no presentan grandes variaciones globales. Sin embargo, destaca el caso excepcional de Manta, que muestra los valores de precisión más bajos con la *pipeline* Minimap2-SAMTools-GATK4. Al evaluar la sensibilidad, Minimap2-SAMTools-GATK4 genera los valores más altos en la mayoría de los algoritmos, mientras que Bowtie2-SAMTools-GATK4 tiende a presentar los valores más bajos.

Para los resultados en función de la réplica empleada, se vuelve a ver la dinámica obtenida en el apartado 4.2. Las réplicas 1 y 2, secuenciadas con IDT-V1, generan los mejores resultados globales, con valores más altos de precisión para la mayoría de los algoritmos. Los mayores valores de precisión se reflejan en la réplica 2, mientras que los peores los representa la réplica 4. Llama la atención como la réplica 5 supera ligeramente en precisión a la réplica 1 y 2 con ExomeDepth y XHMM, y como el mejor rendimiento de XHMM es con la réplica 3.

En términos de *recall*, se observa una amplia variabilidad sin patrones claros asociados a las réplicas. Las réplicas 1 y 2 vuelven a representar globalmente las sensibilidades más altas, mientras que la 4 genera las más bajas. Destacan el caso aislado de LACONV, que logra una sensibilidad notablemente alta en la réplica 4. La réplica 5 logra valores de precisión relativamente aceptables, aunque su sensibilidad sigue siendo limitada.

4 Resultados

Para conocer el rendimiento computacional de cada algoritmo, se escoge la réplica 2, secuenciada con el kit de exoma IDT-V1, para mostrar el tiempo promedio que necesita cada algoritmo para obtener sus resultados.

Algoritmo	cn.MOPS	CNVkit	ExomeDepth	LACONv	Manta	XHMM	CONTRA
Tiempo promedio (h:m:s)	04:33:26	01:10:39	03:16:07	00:22:15	00:37:03	00:23:54	33:04:01

Tabla 4.6: Tiempos promedio de algoritmo para obtener los resultados de la réplica 2

En la tabla 4.6 se observa que CONTRA es el algoritmo más lento con un promedio de más de 33 horas, superando significativamente al resto, seguido de cn.MOPS, con un tiempo promedio superior a 4 horas. Por el contrario, LACONv y XHMM destacan por su rapidez, completando las tareas en aproximadamente 22 y 24 minutos.

4.3.1. Análisis del tipo de variante reportada

Tras evaluar el rendimiento general de cada algoritmo, se profundiza en su capacidad para detectar delecciones y duplicaciones.

En la Figura 4.16, se muestra el número promedio de variantes que reporta cada algoritmo en cada réplica, diferenciando el tipo de variante. Este promedio es en base a los tres análisis bioinformáticos utilizados. Estos reportes son independientes de los dos *gold standards*.

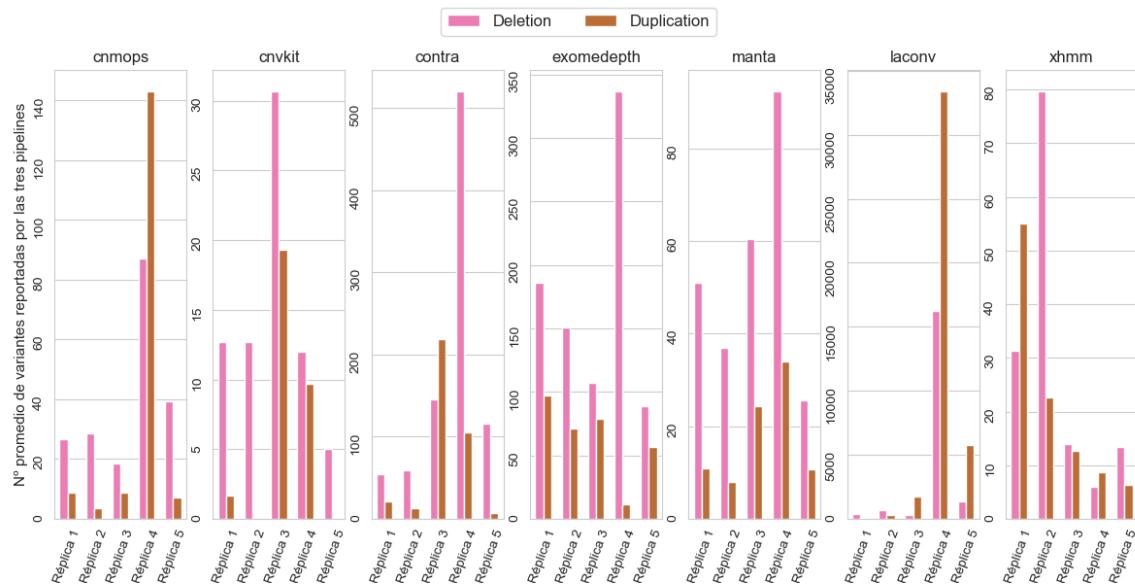


Figura 4.16: Promedio del número de variantes reportadas por cada algoritmo en cada réplica, diferenciado por tipo de variante. En rosa se representan las delecciones, en marrón las duplicaciones.

Se aprecia como en la mayoría de los algoritmos y réplicas, las delecciones son significativamente más numerosas que las duplicaciones.

Las réplicas 1 y 2, secuenciadas con IDT-V1, muestran un número más consistente de variantes detectadas en la mayoría de los algoritmos, proporcionando entre ellas un numero similar de delecciones y duplicaciones.

Generalmente destaca la gran cantidad de variantes detectadas en la réplica 4, muy desproporcionado al resto de réplicas. Concretamente, LACONv reporta un número desproporcionado de variantes

4 Resultados

detectadas en la réplica 4, alcanzando un promedio de 33.389, además de ser la mayoría duplicaciones. Este resultado está directamente relacionado con la alta cobertura que presenta esta réplica, detallado en la tabla 3.6.

La réplica 3 tiene un rendimiento muy variable en función del algoritmo utilizado. La réplica 5, aun con la baja cobertura obtenida, consigue reportar un mayor número de variantes que las réplicas 1 y 2 en los algoritmos de cn.MOPS, CONTRA y LACONV.

En términos generales, el algoritmo que más variantes detecta, de manera muy desproporcionada respecto al resto, es LACONV, seguido a distancia por ExomeDepth.

Por el contrario, CNVkit presentan el rendimiento más bajo en términos de número total de variantes detectadas. En particular, parece estar muy limitado en la detección de duplicaciones detectando una media de 2 duplicaciones en la réplica 1, y 0 en las réplicas 2 y 5. Seguido de este, XHMM tiene un bajo rendimiento en las réplicas 3, 4 y 5, y cn.MOPS en las réplicas 1, 2 y 3.

Llama la atención como XHMM, en contraste con el resto de algoritmos, detecta un mayor número de variantes en las réplicas 1 y 2 y para la réplica 4 reporta el menor número.

A continuación, se representan las métricas de calidad cada algoritmo en función del tipo de variante reportado mediante la Figura 4.17.

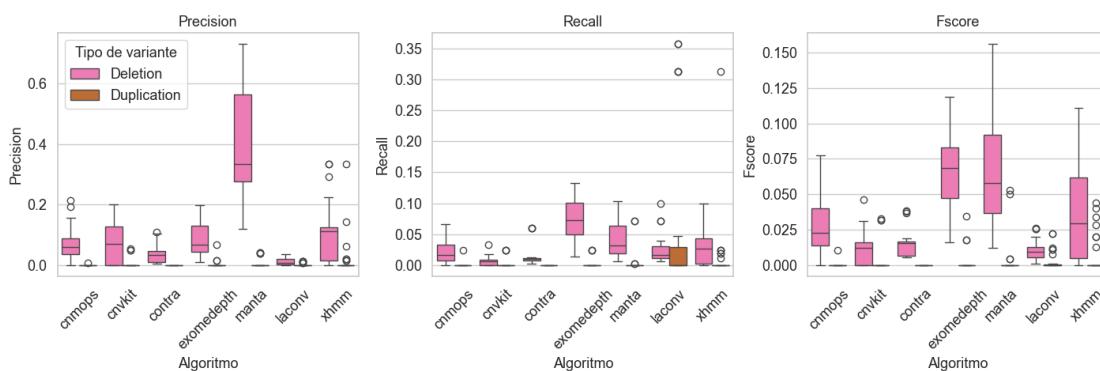


Figura 4.17: Boxplots que representan la distribución de las tres métricas para cada algoritmo en función del tipo de variante reportado. En rosa se representan las delecciones, en marrón las duplicaciones.

Se observa que todos los algoritmos presentan métricas significativamente mejores en la detección de delecciones en comparación con duplicaciones. En términos de precisión, destaca positivamente Manta, con los valores más altos entre los algoritmos. Por otro lado, cn.MOPS, CNVkit, ExomeDepth y XHMM muestran rangos de precisión similares y aceptables. Las peores precisiones para las delecciones se observan en CONTRA y, especialmente, en LACONV, cuyas métricas de precisión son prácticamente nulas. En cuanto a sensibilidad, ExomeDepth sobresale, lo que lo posiciona junto con Manta como los algoritmos con el mejor rendimiento para la detección de delecciones, seguidos por cn.MOPS y XHMM.

Por el contrario, las métricas para duplicaciones, tanto en precisión como en sensibilidad, son casi nulas en la mayoría de los algoritmos. El algoritmo con el mejor rendimiento global para duplicaciones, aunque sigue siendo bajo, es XHMM. Un caso particular es LACONV, que muestra una sensibilidad notablemente alta para duplicaciones en comparación con el resto de algoritmos. Sin embargo, su precisión casi nula hace que la eficiencia de este algoritmo sea muy baja. Este resultado está directamente relacionado con la baja cantidad de duplicaciones en los *gold standards* finales, y a las limitaciones de los algoritmos de identificar duplicaciones por su morfología.

Para conocer los valores exactos de *precision* y *recall* reportados por cada algoritmo, clasificados por réplica y *pipeline* bioinformática, así como diferenciados según el tipo de variante (delecciones y duplicaciones) y *gold standard* (general y *pass*), consulte las Figuras 8, 9, 10 y 11 del Anexo 2. Allí se presentan ocho mapas de calor organizados por métrica, *gold standard* y tipo de variante.

5 Discusión

A pesar del desarrollo múltiples algoritmos para la llamada de CNVs basados en distintas metodologías y técnicas de detección, su caracterización e interpretación continúan siendo un desafío[15, 18]. Esto se debe tanto a factores técnicos, como errores en la amplificación o limitaciones de las técnicas de secuenciación de lecturas cortas, como a restricciones propias del genoma, con regiones repetitivas o complejas que son especialmente difíciles de analizar [34, 36, 25]. Estas limitaciones dificultan la construcción de un conjunto de CNVs robusto y confiable que permita evaluar eficazmente el rendimiento de los algoritmos bioinformáticos. Aunque la muestra NA12878 se utiliza ampliamente como estándar para el análisis de diversas metodologías [46], aún no existe un conjunto de referencia robusto de CNVs [60]. Por ello, uno de los objetivos principales de este trabajo fue construir un *gold standard* interno para validar los algoritmos de detección de CNVs implementados.

Para ello, se seleccionaron cuatro conjuntos de SVs ampliamente reconocidos en la comunidad y se dividieron en dos categorías de filtrado: variantes de alta calidad (*PASS*), y variantes de baja calidad (*LowQual*). El análisis estadístico de los estudios reveló diferencias significativas entre ellos: las variantes *PASS* fueron más consistentes en todos los estudios, mientras que las variantes *LowQual* tendieron a ser más dispersas, menos representativas y menos fiables. Todos los estudios presentaron limitaciones específicas. En el caso de *1000 Genomes*, la mayoría de las variantes *LowQual* eran inserciones, lo que generó un sesgo hacia las delecciones y limitó su capacidad para validarlas en el *gold standard*. GIAB presentó una mediana de longitud de solo 2 bp para las variantes *PASS*, lo que reduce significativamente su utilidad, ya que las CNVs se definen como variantes mayores a 50 bp. MetaSV reportó más de 300.000 variantes *LowQual* con longitudes desproporcionadas y una alta dispersión, reflejando un alto nivel de ruido. Por último, en *svclassify*, la mayoría de las variantes eran delecciones, lo que sesga este conjunto hacia la detección de duplicaciones.

Posteriormente, se aplicaron criterios de filtrado a las variantes de ambas categorías, eliminando aquellas potencialmente artefactuales o redundantes. El filtrado de las variantes *PASS* resultó relativamente sencillo debido a su bajo nivel de ruido y al manejo de una cantidad limitada de variantes, reafirmando la alta calidad de las variantes *PASS* detectadas en estos conjuntos. En contraste, el análisis de las variantes *LowQual* mostró una alta proporción de variantes inconsistentes, especialmente en GIAB y MetaSV. Muchas de estas variantes presentaban tamaños inusuales para un individuo sano. Además, la presencia de múltiples variantes *LowQual* reportadas en la misma región genómica, muchas de ellas inconsistentes en tipo y tamaño, dificultó su procesado y uso en análisis posteriores. Para abordar esta limitación, se empleó un paso intermedio de fusión de variantes redundantes y solapantes. Aunque esta estrategia redujo el ruido de cada estudio, no estuvo exenta de problemas: las variantes fusionadas a menudo tenían puntos de corte imprecisos o tipos de variante indeterminados, introduciendo más incertidumbre en ciertos casos. Esto remarca la necesidad de realizar un análisis futuro más detallado de las variantes *LowQual* para minimizar los falsos positivos y mejorar la precisión de las variantes resultantes.

Con las variantes filtradas, se generaron el *gold standard pass* y el *gold standard general*. El conjunto general incluyó más variantes al incorporar las *LowQual*, aunque el análisis previo señaló que muchas de estas podrían ser falsos positivos, lo que compromete su capacidad como conjunto de validación y resalta la importancia de priorizar las variantes *PASS* para obtener resultados más confiables. En ambos conjuntos, el número de duplicaciones fue muy limitado, representando una importante restricción para validar duplicaciones detectadas por los algoritmos de CNVs. A su vez, ambos *gold standards* se restringieron a las regiones exónicas capturadas por cada uno de los tres kits de secuenciación. El número de variantes resultantes en cada restricción varió considerablemente

entre los kits, reflejando las diferencias en la cobertura genómica proporcionada por cada uno. Estas discrepancias reflejan la importancia de tener en cuenta estas variaciones al validar metodologías o diseñar experimentos.

Para evaluar el rendimiento de siete algoritmos independientes de detección de CNVs, se generaron cinco réplicas de la muestra NA12878 a partir de varias carreras de secuenciación y diferentes kits de exoma. Las métricas de calidad de estas réplicas revelan que las réplicas 1, 2 y 3 tienen una cobertura media estándar. Sin embargo, la réplica 4 presentó una cobertura excesivamente alta, lo que puede generar un exceso de datos que incrementa el ruido, además de aumentar el coste computacional de la secuenciación. La réplica 5 tuvo una cobertura relativamente baja, lo que puede limitar la precisión en la detección de variantes [93]. Las réplicas mostraron diferencias significativas en el rendimiento de los algoritmos. Las réplicas 1 y 2, secuenciadas con IDT-V1, obtuvieron las mejores eficiencias. La réplica 4 presentó métricas de precisión considerablemente bajas, probablemente debido al aumento de ruido y artefactos asociados a la sobrecobertura. En contraste, la réplica 5, con una cobertura baja, mostró resultados aceptables en precisión, aunque con menor sensibilidad. Esto podría indicar una buena eficiencia del kit ROCHE-V1, que, a pesar de la baja cobertura, fue capaz de reportar variantes con una calidad razonable. Las diferencias observadas entre las réplicas reflejan cómo la cobertura y la carrera de secuenciación pueden influir en el rendimiento, generando valores muy dispares incluso para una misma muestra.

Además de los sesgos técnicos, se evaluaron las diferencias entre los algoritmos de alineamiento y procesamiento de lecturas utilizando tres *pipelines* bioinformáticas. La combinación de Minimap2-Samtools-GATK4 destacó con las métricas más altas tanto en precisión como en sensibilidad en la mayoría de las réplicas. En contraste, las combinaciones Bowtie2-Picard-GATK3 y Bowtie2-Samtools-GATK4 mostraron un rendimiento más limitado en la mayoría de los casos. Estos resultados posicionan a Minimap2 como el alineador más eficiente en este análisis.

El rendimiento de los algoritmos fue más bajo de lo esperado, probablemente debido a las limitaciones y las problemáticas encontradas en la generación del *gold standard* interno. Los estudios de referencia utilizados reportan conjuntos de SVs, que al no corresponder exactamente con CNVs, podrían incluir variantes que no son detectables por algoritmos diseñados específicamente para CNVs. Además, estos estudios contenían un elevado número de variantes que podrían ser falsos positivos debido a las variantes *LowQual*. El paso intermedio de fusión de las variantes *LowQual* problemáticas dentro de cada estudio, como era de esperarse, introdujo un alto nivel de ruido. Por lo tanto, los resultados desfavorables, al menos en el caso del *gold standard* general, eran previsibles. Otro aspecto importante es que la mayoría de estos estudios no incluyen duplicaciones como categoría explícita. En consecuencia, en los *gold standards* generados, las inserciones se han considerado equivalentes a duplicaciones, lo que podría introducir un sesgo en su detección.

En relación con la sensibilidad, otro desafío es el alto número de variantes incluidas en el *gold standard*, con una media de 300 variantes en el *gold standard pass* y 2.000 en el general. Dado que NA12878 es un individuo sano, este número elevado podría reflejar un sesgo o un exceso de variantes incluidas, que no representan la realidad de este genoma. Además, algoritmos como CN.mops, CNVKit, Manta y XHMM reportan un promedio de variantes contenido entre 10 y 100 variantes para las réplicas con cobertura intermedia. Esto hace que, al comparar con un *gold standard* tan amplio, la sensibilidad quede baja para estas herramientas, ya que el conjunto de variantes esperado es desproporcionadamente mayor.

Manta mostró la mayor precisión en la mayoría de los escenarios, alcanzando su mejor valor con un 73 % en la réplica 2 utilizando el *gold standard* general. Esto lo posiciona como el mejor algoritmo de detección para los datos analizados en este trabajo. Seguido de este, ExomeDepth destacó en sensibilidad, alcanzando un valor de 12 % en la réplica 1 con el *gold standard pass*. CONTRA presentó valores de *precision* y *recall* generalmente inferiores a la media. Además, su tiempo de computación es significativamente mayor que el del resto de los algoritmos, con un promedio de 33 horas por análisis, lo que lo hace inservible para su implementación en entornos clínicos de rutina en análisis de exomas. LACONv, aun siendo el algoritmo que más variantes reportó, con un

promedio de 1.000 variantes en las réplicas con coberturas adecuadas, obtuvo los peores resultados en precisión. Esto indica que reporta un nivel muy alto de falsos positivos y presenta grandes limitaciones en su capacidad para detectar CNVs de forma confiable. El resto de los algoritmos mostraron un rendimiento más uniforme, pero con métricas que, en general, se mantuvieron en valores bajos. Generalmente, las mejores métricas se reportaron con las réplicas secuenciadas con el kit IDT-V1 y analizadas con la *pipeline Minimap2-Samtools-GATK4*.

Resulta llamativo que Manta sea el único algoritmo que emplea la metodología de detección basada en lectura dividida y mapeo *paired-end*, y que no utilice controles, mientras que el resto de los algoritmos utilizan la metodología basada en profundidad de lecturas y requieren controles para el análisis. Esto sugiere dos situaciones: 1) para el conjunto de datos generado en este trabajo, la metodología SR podría ser más óptima para la detección de CNVs, al menos en las condiciones evaluadas, y 2), es posible que los controles seleccionados en este estudio podrían no ser los más apropiados o suficientemente representativos, lo que podría haber afectado negativamente el rendimiento de los algoritmos.

Al analizar los resultados de los algoritmos en función del tipo de variante, se observa que la detección de duplicaciones es muy deficiente. Esto puede atribuirse a la falta de duplicaciones consistentes en los estudios de referencia y, por ende, en los *gold standards* generados, que apenas contienen duplicaciones. Además, la detección de este tipo de variantes es técnicamente compleja debido a su naturaleza repetitiva en el genoma y a las limitaciones de las técnicas de lectura corta empleadas en este estudio [18, 36]. En contraste, la detección de delecciones muestra resultados más positivos en términos de métricas de evaluación, logrando resultados óptimos en algunas casuísticas, generalmente con Manta o ExomeDepth.

En cuanto a los dos *gold standards* utilizados en la validación, se observó que, en la mayoría de los casos, el *gold standard* general tenía la mejoría las métricas de precisión, mientras que el *gold standard pass* favorecía una mayor sensibilidad. Esto era esperable, ya que el *gold standard* general incluye un número mucho más elevado de variantes, lo que aumenta la probabilidad de coincidencia con los resultados de los algoritmos. Por otro lado, el *gold standard pass*, al centrarse exclusivamente en variantes de mayor calidad, permite evaluar con mayor rigurosidad la capacidad de los algoritmos para detectar variantes robustas.

En general, los resultados obtenidos en este estudio fueron peores de lo esperado, lo que evidencia la necesidad de abordar las limitaciones identificadas para mejorar el rendimiento en el proceso de detección de variantes en muestras clínicas WES. En consecuencia, será imprescindible realizar un análisis más exhaustivo para solventar estas deficiencias y alcanzar una productividad óptima de los algoritmos. Aunque el uso de varias réplicas para analizar sesgos técnicos de secuenciación es una metodología sólida, es fundamental evitar trabajar con coberturas extremas, tanto altas como bajas, ya que ambas parecen impactar negativamente en la detección de variantes. Entre las acciones necesarias, se debe implementar un filtrado más riguroso de las variantes para minimizar al máximo la introducción de ruido en el *gold standard*. Concretamente, para las variantes de baja calidad, se debe evitar usar herramientas que generar variantes fusionadas genéricas con puntos de corte inciertos. En términos de algoritmos, se deberá incluir algoritmos que empleen metodologías variadas y no solamente emplear profundidad de lectura, ya que, al menos en este estudio, no ha demostrado un buen rendimiento. Además, será fundamental evaluar los resultados con diferentes muestras control y con un conjunto más amplio, ya que la cohorte de controles elegida para este trabajo podría no haber sido adecuada.

Este estudio demuestra los desafíos asociados a la creación de un conjunto de validación robusta que abarque las variantes presentes en el exoma humano, así como las limitaciones actuales en la predicción de CNVs a partir de los conjuntos de variantes disponibles. Superar estas limitaciones y establecer las condiciones experimentales más óptimas para la llamada de CNVs requerirá una investigación más profunda. Sin embargo, las metodologías empleadas en este trabajo, junto con los conjuntos de validación generados, representan un recurso valioso para avanzar en el estudio de las CNVs y sus aplicaciones clínicas.

6 Conclusiones

- [1.] Los *gold standards* generados presentan limitaciones derivadas del uso de herramientas de fusión de variantes y del alto nivel de ruido asociado a las variantes de baja calidad reportadas por los estudios iniciales, lo que remarca la necesidad de un filtrado más riguroso para mejorar su confiabilidad.
- [2.] Las técnicas de captura de exoma, las coberturas obtenidas y los procesamientos bioinformáticos utilizados parecen influir significativamente en la detección de CNVs, generando resultados muy variables. En este estudio, el kit de exoma IDT-V1 junto con el alineador Minimap2 ofrecieron, en general, los mejores resultados.
- [3.] Los algoritmos de detección de CNVs mostraron, en general, un bajo rendimiento. Manta destacó por su mayor precisión, mientras que ExomeDepth sobresalió en sensibilidad. En contraste, LACONv presentó un desempeño deficiente, reportando un alto número de variantes pero con las peores métricas de precisión y sensibilidad, reflejando una elevada tasa de falsos positivos.
- [4.] Aun siendo la profundidad de lectura el enfoque predominante en la detección de CNVs, bajo las condiciones experimentales de este estudio, mostró un rendimiento inferior en comparación con la metodología de lectura dividida y mapeo *paired-end*, como la utilizada por Manta.
- [5.] La implementación de un flujo de trabajo automatizado mediante Snakemake y Singularity demostró ser una estrategia eficaz para estandarizar, escalar y reproducir el análisis de CNVs, facilitando la integración de múltiples herramientas bioinformáticas.

Anexo 1

Plataforma	Run	Nombre de muestra	Género	Tipo de muestra
HiSeq4000	201015	NGS27168-HiSeq4000-exoma-Run201015-HG0002	H	Control
HiSeq4000	201015	NGS27169-HiSeq4000-exoma-Run201015-HG0003	H	Control
HiSeq4000	201015	NGS27170-HiSeq4000-exoma-Run201015-HG0004	M	Control
HiSeq4000	200625	NGS26087-HiSeq4000-exoma-Run200625-87439	M	Control
HiSeq4000	200625	NGS26088-HiSeq4000-exoma-Run200625-87440	H	Control
HiSeq4000	200625	NGS26091-HiSeq4000-exoma-Run200625-87561	M	Control
HiSeq4000	200625	NGS26094-HiSeq4000-exoma-Run200625-82126-	M	Control
HiSeq4000	200625	NGS26095-HiSeq4000-exoma-Run200625-82128	H	Control
HiSeq4000	200625	NGS26148-HiSeq4000-exoma-Run200625-87507	M	Control
HiSeq4000	200625	NGS26149-HiSeq4000-exoma-Run200625-87506	H	Control
HiSeq4000	200625	NGS26152-HiSeq4000-exoma-Run200625-87764	H	Control
HiSeq4000	200625	NGS26156-HiSeq4000-exoma-Run200625-72852	M	Control
HiSeq4000	200625	NGS26157-HiSeq4000-exoma-Run200625-72853	H	Control

Tabla 1: Grupo 1 de muestras control secuenciadas con el kit de exoma **IDT-V1**, asociados con las réplicas 1 y 2.

Plataforma	Run	Nombre de muestra	Género	Tipo de muestra
NovaSeq6000	230119	Novaseq6000-Exome-v2-IDT-Run230119-HG-002	H	Control
NovaSeq6000	230119	Novaseq6000-Exome-v2-IDT-Run230119-HG-003	H	Control
NovaSeq6000	230119	Novaseq6000-Exome-v2-IDT-Run230119-HG-004	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-116906	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-116907	H	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-114118	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-117474	H	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-114720	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-118245	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-118246	H	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-118194	M	Control
NovaSeq6000	240121	NovaSeq6000-Exoma-V2-Run240121-NEUR-118192	H	Control

Tabla 2: Grupo 2 de muestras control secuenciadas con el kit de exoma **IDT-V2**, asociadas con las réplicas 3 y 4.

Plataforma	Run	Nombre de muestra	Género	Tipo de muestra
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-OFHI-112508	H	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-OFHI-112509	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-116440	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-120374	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-120369	H	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-121094	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-121091	H	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-121190	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-121192	H	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-123575	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-NEUR-123577	H	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-GE1194-123655	M	Control
NovaSeq6000	240722	NovaSeq6000-Exoma-V2-Run240722-GE1195-123674	H	Control

Tabla 3: Grupo 3 de muestras control secuenciadas con el kit de exoma **ROCHE-V1**, asociadas con la réplica 5.

Estudio	Número de variantes	Tipos de variantes	Filtrado de calidad de variantes
<i>The 100 Genomes Project</i>	3.260	DEL - 1.982 INS - 1096 INV - 28 DUP - 8 CNV - 146	<i>HighQuality</i> - 2.192 <i>LowQuality</i> - 1.068
<i>Genome in a Bottle Project</i>	41.785	DEL - 20.207 INS - 21.578	<i>HighQuality</i> - 10.594 <i>LowQuality</i> - 32.562
MetaSV	314.253	DEL - 16.987 INS - 101.777 INV - 195.489	<i>HighQuality</i> - 4.718 <i>LowQuality</i> - 309.535
svclassify	2.744	DEL - 2.676 INS - 68	<i>HighQuality</i> - 2.744 <i>LowQuality</i> - 0

Tabla 4: Resumen de las variantes estructurales reportadas en los cuatro estudios de referencia para NA12878, incluyendo el número total de variantes, su clasificación por tipo y su distribución según el filtrado de calidad (*HighQuality/PASS*, *LowQuality/LowQual*)

Anexo 2



Figura 1: Distribución de las variantes *PASS* filtradas del estudio de los **1000 Genomas**, por tipo y tamaño.



Figura 2: Distribución de las variantes *PASS* filtradas del estudio de **GIAB**, por tipo y tamaño.

Anexo 2

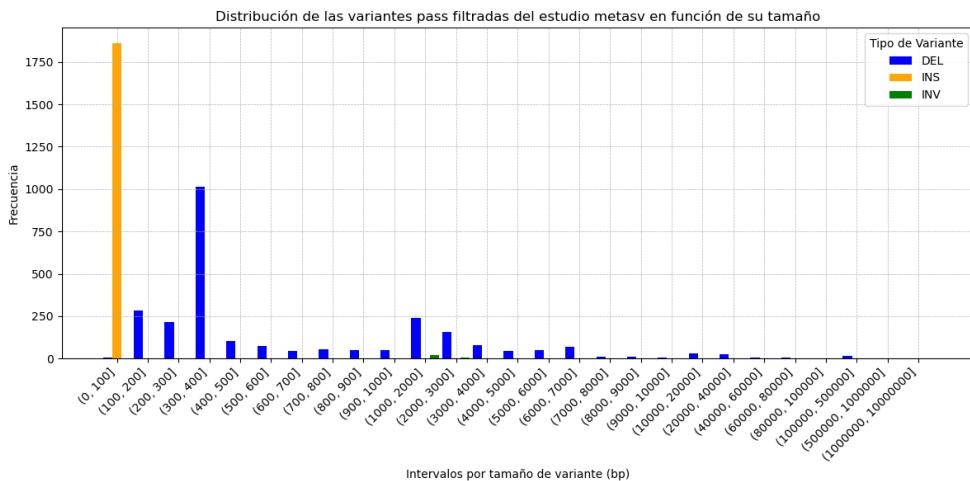


Figura 3: Distribución de las variantes *PASS* filtradas del estudio de **MetaSV**, por tipo y tamaño.

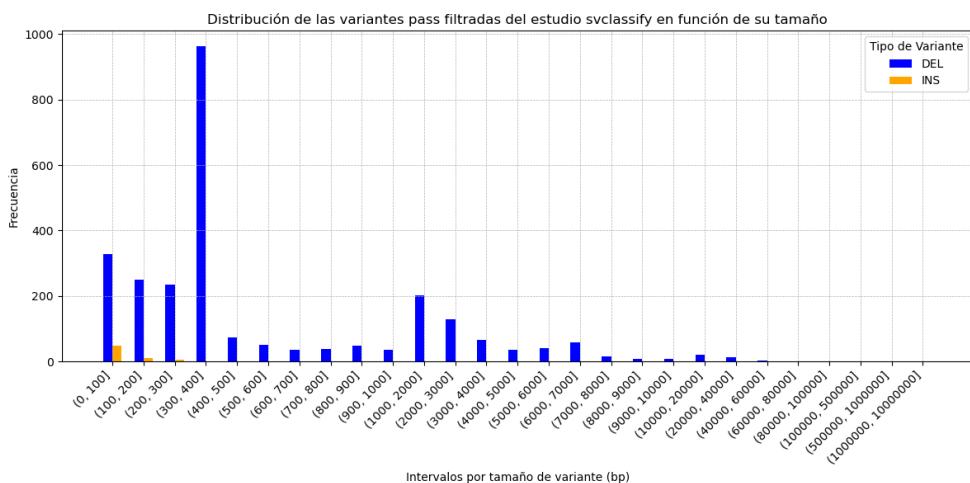


Figura 4: Distribución de las variantes *PASS* filtradas del estudio de **SVclassify**, por tipo y tamaño.



Figura 5: Distribución de las variantes *LowQual* filtradas del estudio de los 1000 **Genomas**, por tipo y tamaño.

Anexo 2

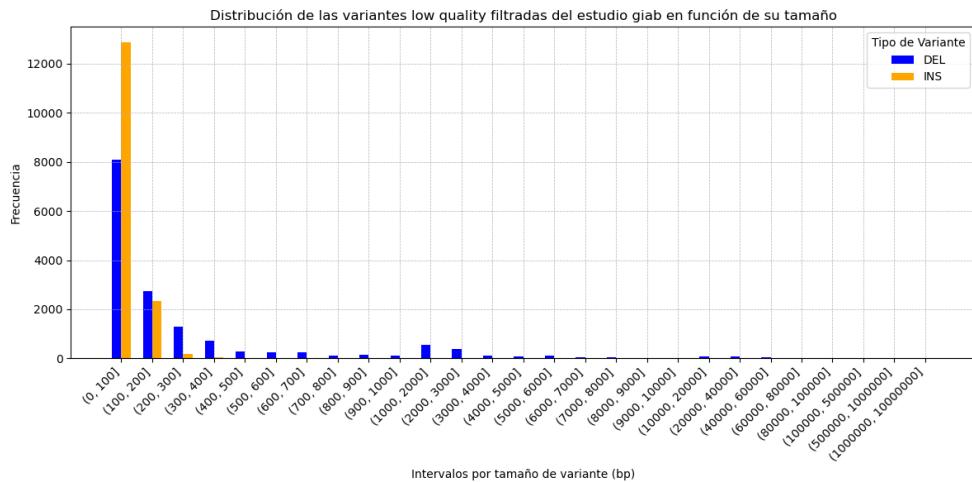


Figura 6: Distribución de las variantes *LowQual* filtradas del estudio de **GIAB**, por tipo y tamaño.

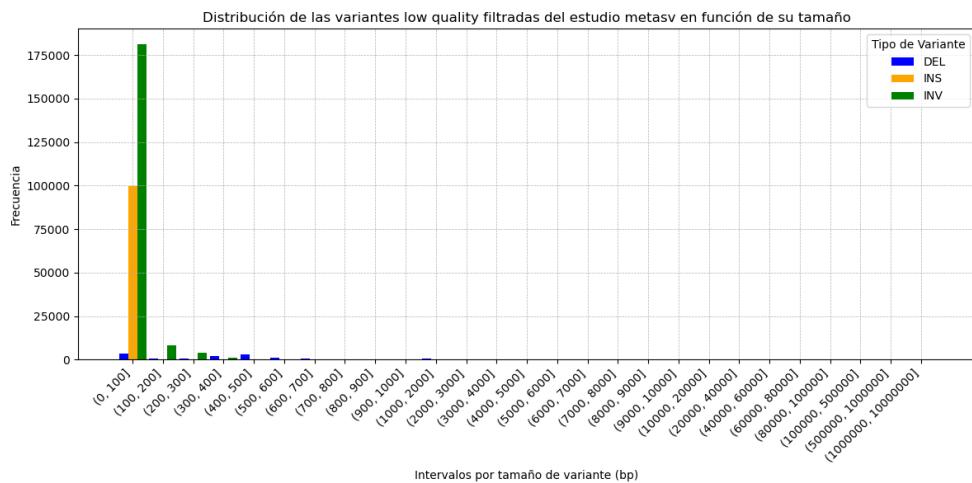


Figura 7: Distribución de las variantes *LowQual* filtradas del estudio de **MetaSV**, por tipo y tamaño.

Gold Standard	Kit de exoma	Cantidad	Longitud media	Desv. Est.	Mínimo	Mediana	Máximo
General	IDT-V1	1.914	90.473	101.920	53	58.165	505.675
	IDT-V2	1.791	92.985	107.221	53	58.165	505.675
	ROCHE-V1	2.285	96.847	107.938	53	58.165	505.675
PASS	IDT-V1	317	132.810	165.900	53	41.942	505.675
	IDT-V2	266	155.071	178.838	53	93.240	505.675
	ROCHE-V1	330	143.631	176.678	53	44.899	505.675

Tabla 1: Métricas de longitud de las variantes de los *gold standards* finales tras su intersección con las regiones exónicas capturadas por cada kit de captura.

Anexo 2

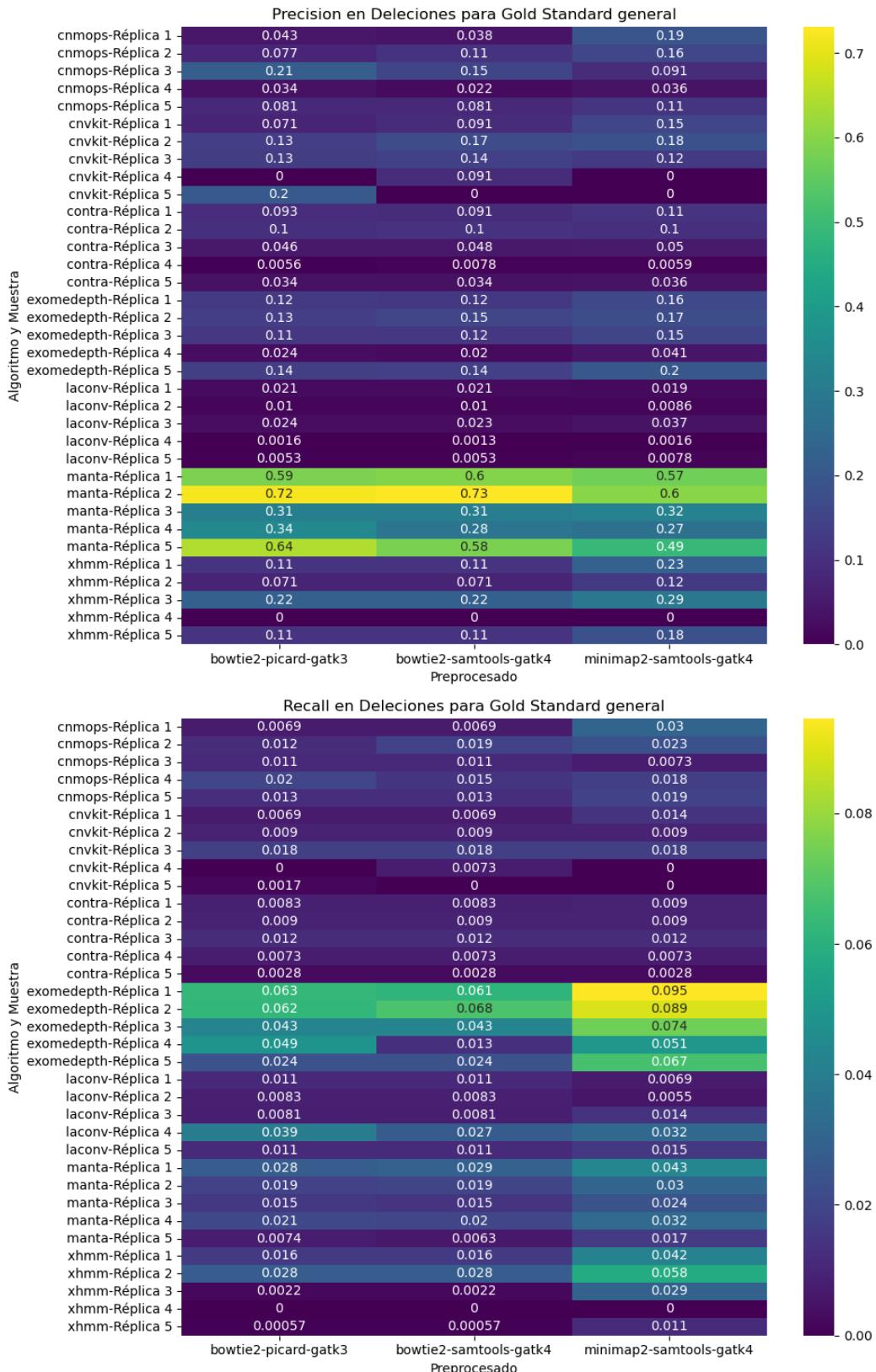


Figura 8: Mapas de calor de las métricas *precision* y *recall* para las delecciones validadas con el gold standard general, clasificadas por algoritmo y réplica, y pipeline de preprocesamiento

Anexo 2

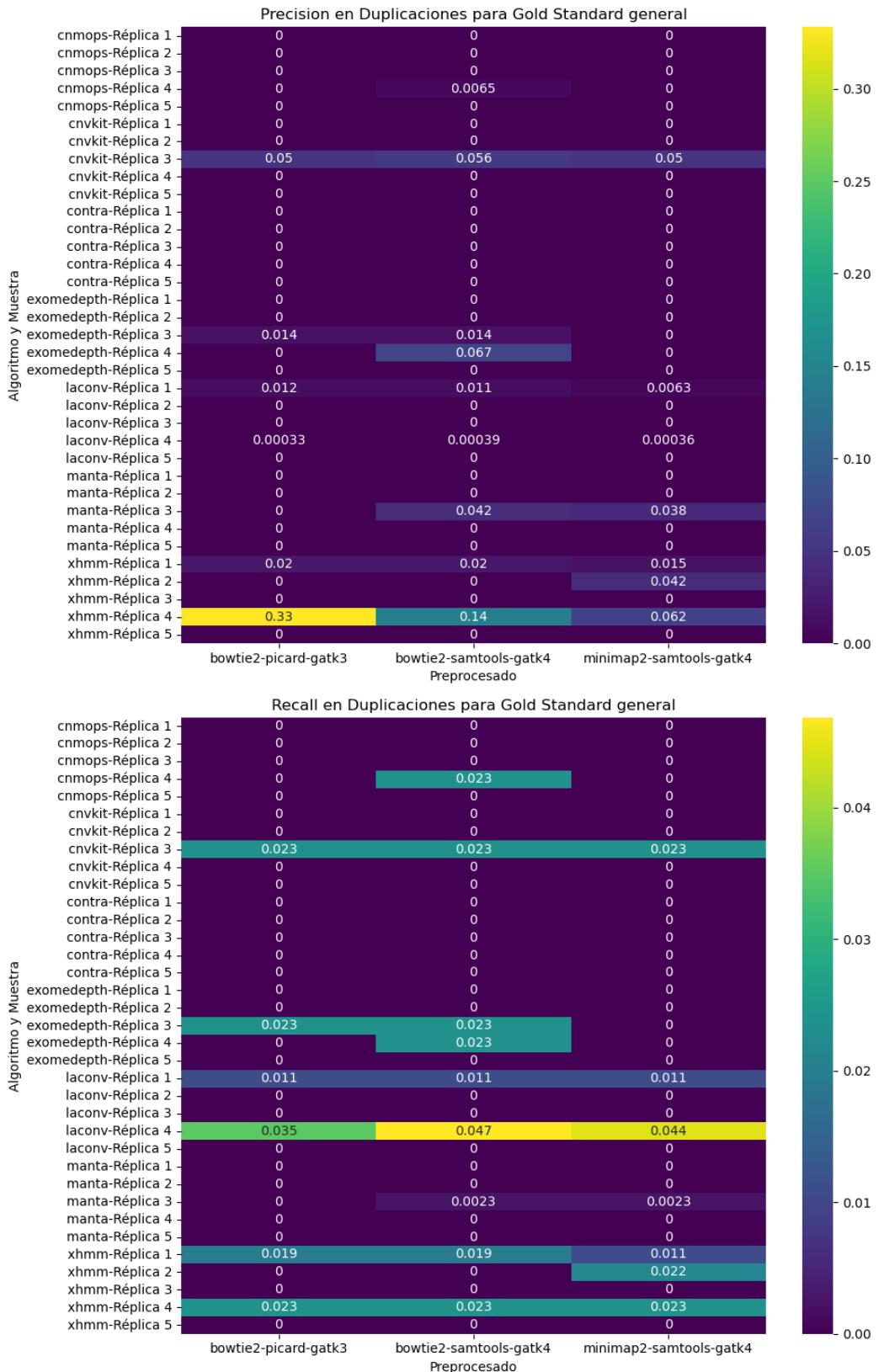


Figura 9: Mapas de calor de las métricas *precision* y *recall* para las duplicaciones validadas con el gold standard general, clasificadas por algoritmo y réplica, y pipeline de preprocesamiento

Anexo 2

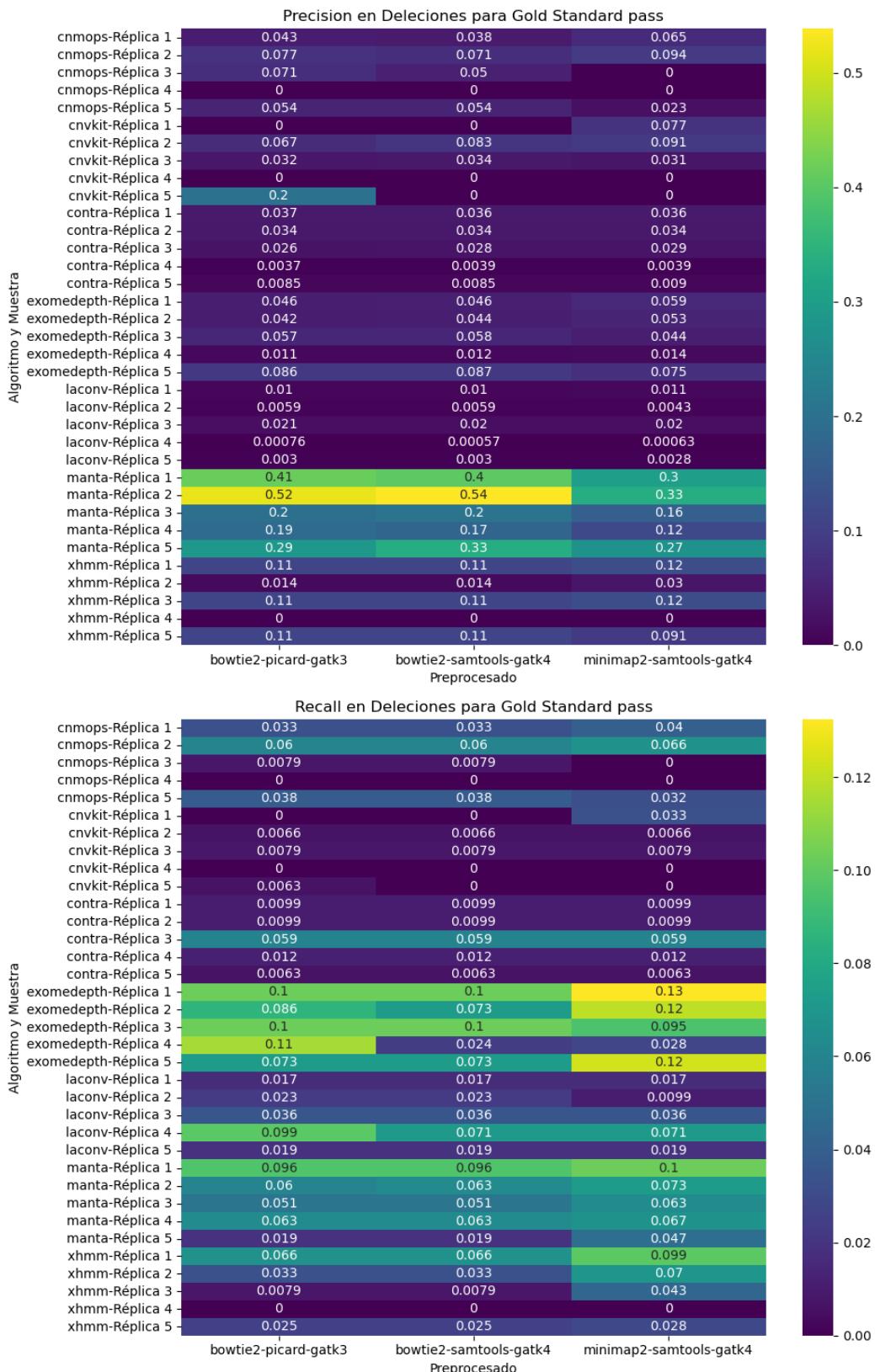


Figura 10: Mapas de calor de las métricas *precision* y *recall* para las delecciones validadas con el gold standard pass, clasificadas por algoritmo y réplica, y pipeline de preprocesamiento

Anexo 2

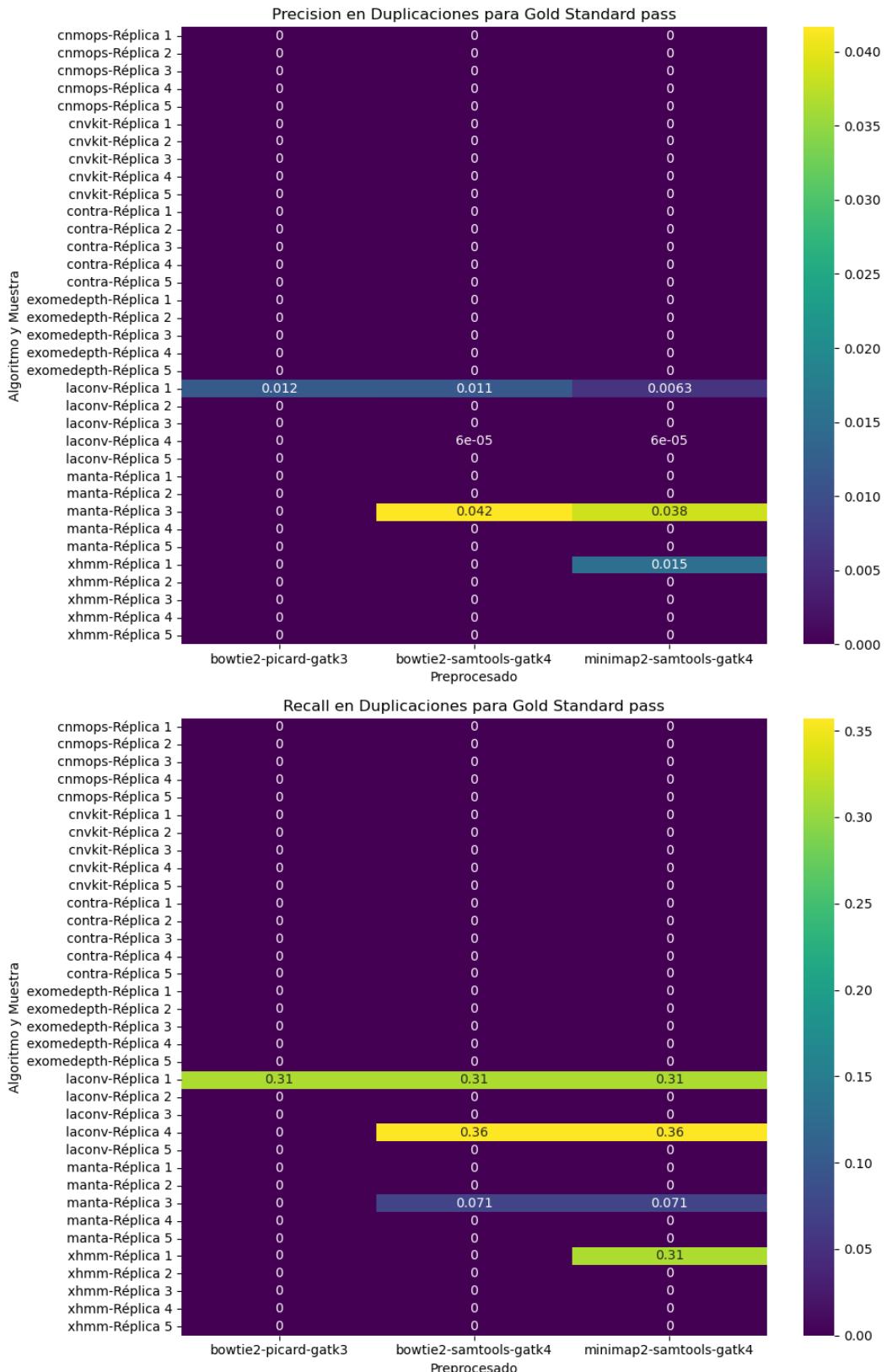


Figura 11: Mapas de calor de las métricas *precision* y *recall* para las duplicaciones validadas con el gold standard pass, clasificadas por algoritmo y réplica, y pipeline de preprocesamiento

Bibliografía

- [1] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.
- [2] Samuel Levy and Robert L Strausberg. Individual genomes diversify. *Nature*, 456(7218):49–51, 2008.
- [3] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.
- [4] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.
- [5] Philip N Benfey and Thomas Mitchell-Olds. From genotype to phenotype: systems biology meets natural variation. *Science*, 320(5875):495–497, 2008.
- [6] Nasheen Naidoo, Yudi Pawitan, Richie Soong, David N Cooper, and Chee-Seng Ku. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human genomics*, 5:1–46, 2011.
- [7] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.
- [8] Han Chang, Donald G Jackson, Paul S Kayne, Petra B Ross-Macdonald, Rolf-Peter Ryseck, and Nathan O Siemers. Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PloS one*, 6(6):e21097, 2011.
- [9] Steve S Ho, Alexander E Urban, and Ryan E Mills. Structural variation in the sequencing era. *Nature Reviews Genetics*, 21(3):171–189, 2020.
- [10] Melbourne Bioinformatics. Long-read structural variant calling. https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/.
- [11] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nature reviews genetics*, 12(5):363–376, 2011.
- [12] Sobin Kim and Ashish Misra. Snp genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, 9(1):289–320, 2007.
- [13] Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology*, 20:1–18, 2019.
- [14] Lorenzo Tattini, Romina D’Aurizio, and Alberto Magi. Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, 3:92, 2015.
- [15] Shu Mei Teo, Yudi Pawitan, Chee Seng Ku, Kee Seng Chia, and Agus Salim. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718, 2012.
- [16] Ke Lin, Sandra Smit, Guusje Bonnema, Gabino Sanchez-Perez, and Dick de Ridder. Making the difference: integrating structural variation detection tools. *Briefings in bioinformatics*, 16(5):852–864, 2015.
- [17] Ilkka Lappalainen, John Lopez, Lisa Skipper, Timothy Heffernon, J Dylan Spalding, John Garner, Chao Chen, Michael Maguire, Matt Corbett, George Zhou, et al. Dbvar and dgva: public archives for genomic structural variation. *Nucleic acids research*, 41(D1):D936–D941, 2012.
- [18] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14(Suppl 11):S1, 2013.

Bibliografía

- [19] Jacques S Beckmann, Xavier Estivill, and Stylianos E Antonarakis. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, 8(8):639–646, 2007.
- [20] Nathan Fortier, Gabe Rudy, and Andreas Scherer. Detection of cnvs in ngs data using vs-cnv. *Copy Number Variants: Methods and Protocols*, pages 115–127, 2018.
- [21] Rory Collins. What makes uk biobank special? *The Lancet*, 379(9822):1173–1174, 2012.
- [22] Chiara Auwerx, Maarja Jõeloo, Marie C Sadler, Nicolò Tesio, Sven Ojavee, Charlie J Clark, Reedik Mägi, Estonian Biobank Research Team Esko Tõnu Metspalu Andres Milani Lili Nelis Mari, Alexandre Reymond, and Zoltán Kutalik. Rare copy-number variants as modulators of common disease susceptibility. *Genome Medicine*, 16(1):5, 2024.
- [23] Nigel P Carter. Methods and strategies for analyzing copy number variation using dna microarrays. *Nature genetics*, 39(Suppl 7):S16–S21, 2007.
- [24] Chandra Shekhar Pareek, Rafal Smoczyński, and Andrzej Tretyń. Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52:413–435, 2011.
- [25] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021.
- [26] Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. Overview of next-generation sequencing technologies. *Current protocols in molecular biology*, 122(1):e59, 2018.
- [27] Beate M Crossley, Jianfa Bai, Amy Glaser, Roger Maes, Elizabeth Porter, Mary Lea Killian, Travis Clement, and Kathy Toohey-Kurth. Guidelines for sanger sequencing and molecular assay monitoring. *Journal of Veterinary Diagnostic Investigation*, 32(6):767–775, 2020.
- [28] Tracy Tucker, Marco Marra, and Jan M Friedman. Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2):142–154, 2009.
- [29] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15, 2014.
- [30] Illumina. Illumina official website. = <https://www.illumina.com>.
- [31] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.
- [32] Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human molecular genetics*, 19(R2):R227–R240, 2010.
- [33] Sen Zhao, Xi Cheng, Wen Wen, Guixing Qiu, Terry Jianguo Zhang, Zhihong Wu, and Nan Wu. Advances in clinical genetics and genomics. *Intelligent medicine*, 1(03):128–133, 2021.
- [34] Daniel C Koboldt. Best practices for variant calling in clinical sequencing. *Genome Medicine*, 12(1):91, 2020.
- [35] Hanns-Georg Klein, Peter Bauer, and Tina Hambuch. Whole genome sequencing (wgs), whole exome sequencing (wes) and clinical exome sequencing (ces) in patient care. *LaboratoriumsMedizin*, 38(4):221–230, 2014.
- [36] Sophia Yohe and Bharat Thyagarajan. Review of clinical next-generation sequencing. *Archives of pathology & laboratory medicine*, 141(11):1544–1557, 2017.
- [37] Petar Brlek, Luka Bulić, Matea Bračić, Petar Projić, Vedrana Škaro, Nidhi Shah, Parth Shah, and Dragan Primorac. implementing whole genome sequencing (wgs) in clinical practice: advantages, challenges, and future perspectives. *Cells*, 13(6):504, 2024.
- [38] Nathan D Olson, Justin Wagner, Nathan Dwarshuis, Karen H Miga, Fritz J Sedlazeck, Marc Salit, and Justin M Zook. Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics*, 24(7):464–483, 2023.
- [39] Mehdi Pirooznia, Fernando S Goes, and Peter P Zandi. Whole-genome cnv analysis: advances in computational approaches. *Frontiers in genetics*, 6:138, 2015.
- [40] Migle Gabreilaite, Mathias Husted Torp, Malthe Sebro Rasmussen, Sergio Andreu-Sánchez, Filipe Garrett Vieira, Christina Bligaard Pedersen, Savvas Kinalis, Majbritt Busk Madsen, Miyako Kodama, Gül Sude Demircan, et al. A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*, 13(24):6283, 2021.

Bibliografía

- [41] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports*, 5(1):17875, 2015.
- [42] Justin M Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nature biotechnology*, 32(3):246–251, 2014.
- [43] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.
- [44] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli L Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. 2003.
- [45] George M Church. The personal genome project. *Molecular systems biology*, 1(1):2005–0030, 2005.
- [46] Soobok Joe, Jong-Lyul Park, Jun Kim, Sangok Kim, Ji-Hwan Park, Min-Kyung Yeo, Dongyoong Lee, Jin Ok Yang, and Seon-Young Kim. Comparison of structural variant callers for massive whole-genome sequence data. *BMC genomics*, 25(1):318, 2024.
- [47] Integrated DNA Technologies. IDT Official Website. <https://eu.idtdna.com/page>.
- [48] Roche Sequencing. Roche Sequencing Official Website. <https://sequencing.roche.com/global/en/home.html>.
- [49] Illumina. bcl2fastq conversion software. <https://emea.support.illumina.com/sequencing/sequencing-software/bcl2fastq-conversion-software.html>.
- [50] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.
- [51] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- [52] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.
- [53] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.
- [54] National Center for Biotechnology Information. Genome assembly grch37.p13. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/.
- [55] Broad Institute. Human genome reference builds: Grch38 or hg38, b37, hg19. <https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19>.
- [56] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.
- [57] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giaboo8, 2021.
- [58] Brent S Pedersen and Aaron R Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 2018.
- [59] Eliot Cline, Nuttachat Wisittipanit, Tossapon Boongoen, Ekachai Chukenatirote, Darush Struss, and Anant Eungwanichayapant. Recalibration of mapping quality scores in illumina short-read alignments improves snp detection results in low-coverage sequencing data. *PeerJ*, 8:e10501, 2020.
- [60] Veronika Gordeeva, Elena Sharova, Konstantin Babalyan, Rinat Sultanov, Vadim M Govorun, and Georgij Arapidi. Benchmarking germline cnv calling tools from exome sequencing data. *Scientific reports*, 11(1):14416, 2021.
- [61] Genomes Project Consortium, A Auton, LD Brooks, RM Durbin, EP Garrison, and HM Kang. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Bibliografía

- [62] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.
- [63] Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
- [64] Python Software Foundation. Python programming language. <https://www.python.org>.
- [65] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780–786, 2015.
- [66] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289, 2015.
- [67] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O’Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050–1054, 2016.
- [68] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, et al. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, 10(6):563–569, 2013.
- [69] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC bioinformatics*, 13:1–18, 2012.
- [70] Gennady Denisov, Brian Walenz, Aaron L Halpern, Jason Miller, Nelson Axelrod, Samuel Levy, and Granger Sutton. Consensus generation and variant detection by celera assembler. *Bioinformatics*, 24(8):1035–1040, 2008.
- [71] Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611, 2015.
- [72] Hemang Parikh, Marghoob Mohiyuddin, Hugo YK Lam, Hariharan Iyer, Desu Chen, Mark Pratt, Gabor Bartha, Noah Spies, Wolfgang Losert, Justin M Zook, et al. svclassify: a method to establish benchmark structural variant calls. *BMC genomics*, 17:1–16, 2016.
- [73] Marghoob Mohiyuddin, John C Mu, Jian Li, Narges Bani Asadi, Mark B Gerstein, Alexej Abyzov, Wing H Wong, and Hugo YK Lam. Metasv: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, 31(16):2741–2744, 2015.
- [74] Anaconda. Conda. <https://anaconda.org/anaconda/conda>.
- [75] Aaron R. Quinlan and Ira M. Hall. Bedtools merge documentation. <https://bedtools.readthedocs.io/en/latest/content/tools/merge.html>.
- [76] Aaron R. Quinlan and Ira M. Hall. Bedtools multiinter documentation. <https://bedtools.readthedocs.io/en/latest/content/tools/multiinter.html>.
- [77] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.
- [78] Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arne Clevert, Andreas Mittrecker, Ulrich Bodenhofer, and Sepp Hochreiter. cn. mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research*, 40(9):e69–e69, 2012.
- [79] Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS computational biology*, 12(4):e1004873, 2016.
- [80] Jason Li, Richard Lupat, Kaushalya C Amarasinghe, Ella R Thompson, Maria A Doyle, Georgina L Ryland, Richard W Tothill, Saman K Halgamuge, Ian G Campbell, and Kylie L Gorringe. Contra: copy number analysis for targeted resequencing. *Bioinformatics*, 28(10):1307–1313, 2012.

Bibliografía

- [81] Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754, 2012.
- [82] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 2016.
- [83] Menachem Fromer and Shaun M Purcell. Using xhmm software to detect copy number variation in whole-exome sequencing data. *Current protocols in human genetics*, 81(1):7–23, 2014.
- [84] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017.
- [85] Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.
- [86] Docker. What is a container? <https://www.docker.com/resources/what-container/>.
- [87] Inc. Docker. Docker - empowering app development for developers. <https://www.docker.com>.
- [88] Richard M Stallman, Roland McGrath, and Paul D Smith. *GNU Make: A program for directing recompilation, for version 3.81*. Free Software Foundation, 2004.
- [89] Illumina. wittyer: A tool for comparing structural variants. <https://github.com/Illumina/witty.er>.
- [90] .NET Foundation. .net core release notes - version 6.0. <https://github.com/dotnet/core/tree/main/release-notes/6.0>.
- [91] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.
- [92] Philipp W Messer and Peter F Arndt. The majority of recent short dna insertions in the human genome are tandem duplications. *Molecular Biology and Evolution*, 24(5):1190–1197, 2007.
- [93] Xiaoqing Yu and Shuying Sun. Comparing a few snp calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, 14:1–15, 2013.