Universitat Politècnica de Catalunya-BarcelonaTECH (UPC), Statistics and Operations Research Department

SEQUENTIA BIOTECH SL, Bioinformatics Department

PhD program in Bioinformatics

Genomics tools: the new frontier in omics data analysis

Dissertation submitted for the degree of Doctor in Bioinformatics

Doctoral thesis by:

Rosa Barcelona Cabeza

Thesis supervisors:

Walter Sanseverino

Riccardo Aiese Cigliano

Thesis tutor:

Guadalupe Gómez Melis

Barcelona, September 2021

Abstract

Substantial technological advancements in next generation sequencing (NGS) have revolutionized the genomic field. Over the last years, the speed and throughput of NGS technologies have increased while their costs have decreased, allowing us to achieve base-by-base interrogation of the human genome in an efficient and affordable way. All these advances have led to a growing application of NGS technologies in clinical practice to identify the genomics variations and their relationship with certain diseases. However, there is still the need to improve data accessibility, processing and interpretation due to both the huge amount of data generated by these sequencing technologies and the large number of tools available to process it. In addition to a large number of algorithms for variant discovery, each type of variation and data requires the use of a specific algorithm. Therefore, a solid background in bioinformatics is required to be able to select the most suitable algorithm in each case but also to be able to execute them successfully.

On that basis, the aim of this project is to facilitate the processing of sequencing data for variant identification and interpretation for non-bioinformaticians. All this by creating high-performance workflows with a strong scientific basis, while remaining accessible and easy to use, as well as a simple and highly intuitive platform for data interpretation.

An exhaustive bibliographic review has been carried out where the best existing algorithm has been selected to create automatic pipelines for the discovery of germline short variants (SNPs and indels) and germline structural variants (SVs), including both CNVs and chromosomal rearrangements, from modern human DNA. In addition to creating variant discovery pipelines, a pipeline has been implemented for *in silico* optimization of CNV detection from WES and TS data (isoCNV). This optimization pipeline has been shown to increase the sensitivity of CNV discovery using only NGS data. Such increased sensitivity is especially important for diagnosis in the clinical settings. Furthermore, a variant discovery workflow has been developed by integrating WES and RNA-seq data (varRED) that has been shown to increase the number of variants identified over those identified when only using WES data. It is important to note that variant discovery is not only important for modern populations, the study of the variation in ancient genomes is also essential to understand past human evolution. Thus, a germline short variant discovery pipeline from ancient WGS samples has been implemented. This workflow has been applied to a human mandible dated between 16980-16510 calibrated years before the present. The ancient short variants discovered were reported without further interpretation due to the low sample coverage. Finally, GINO has been implemented to facilitate the interpretation of the variants identified by the workflows developed in the context of this thesis. GINO is an easy-to-use platform for the visualization and interpretation of germline variants under user license.

With the development of this thesis, it has been possible to implement the necessary tools for a high-performance identification of all types of germline variants, as well as a powerful platform to interpret the identified variants in a simple and fast way. Using this platform allows non-bioinformaticians to focus on interpreting results without having to worry about data processing with the guarantee of scientifically sound results. Furthermore, it has laid the foundations for implementing a platform for comprehensive analysis and visualization of genomic data in the cloud in the near future.

Key words: NGS, Genomics, Variants, Bioinformatics, SNPs, Indels, CNVs, SVs, Platform

Resumen

Los avances tecnológicos en la secuenciación de próxima generación (NGS) han revolucionado el campo de la genómica. El aumento de velocidad y rendimiento de las tecnologías NGS de los últimos años junto con la reducción de su coste ha permitido interrogar base por base el genoma humano de una manera eficiente y asequible. Todos estos avances han permitido incrementar el uso de las tecnologías NGS en la práctica clínica para la identificación de variaciones genómicas y su relación con determinadas enfermedades. Sin embargo, sigue siendo necesario mejorar la accesibilidad, el procesamiento y la interpretación de los datos debido a la enorme cantidad de datos generados y a la gran cantidad de herramientas disponibles para procesarlos. Además de la gran cantidad de algoritmos disponibles para el descubrimiento de variantes, cada tipo de variación y de datos requiere un algoritmo específico. Por ello, se requiere una sólida formación en bioinformática tanto para poder seleccionar el algoritmo más adecuado como para ser capaz de ejecutarlo correctamente.

Partiendo de esa base, el objetivo de este proyecto es facilitar el procesamiento de datos de secuenciación para la identificación e interpretación de variantes para los no bioinformáticos. Todo ello mediante la creación de flujos de trabajo de alto rendimiento y con una sólida base científica, sin dejar de ser accesibles y fáciles de utilizar, así como de una plataforma sencilla y muy intuitiva para la interpretación de datos.

Se ha realizado una exhaustiva revisión bibliográfica donde se han seleccionado los mejores algoritmos con los que crear flujos de trabajo automáticos para el descubrimiento de variantes cortas germinales (SNPs e indels) y variantes estructurales germinales (SV), incluyendo tanto CNV como reordenamientos cromosómicos, de ADN humano moderno. Además de crear flujos de trabajo para el descubrimiento de variantes, se ha implementado un flujo para la optimización *in silico* de la detección de CNV a partir de datos de WES y TS (isoCNV). Se ha demostrado que dicha optimización aumenta la sensibilidad de detección utilizando solo datos NGS, lo que es especialmente importante para el diagnóstico clínico. Además, se ha desarrollado un flujo de trabajo para el descubrimiento de variantes mediante la integración de datos de WES y RNA-seq (varRED) que ha demostrado aumentar el número de variantes detectadas sobre las identificadas cuando solo se utilizan datos de WES. Es importante señalar que la identificación de variantes no solo es importante para las poblaciones modernas, el estudio de las variaciones en genomas antiguos es esencial para comprender la evolución humana. Por ello, se ha implementado un flujo de trabajo para la identificación de variantes cortas a partir de muestras antiguas de WGS. Dicho flujo se ha aplicado a una mandíbula humana datada entre el 16980-16510 a.C. Las variantes ancestrales allí descubiertas se informaron sin mayor interpretación debido a la baja cobertura de la muestra. Finalmente, se ha implementado GINO para facilitar la interpretación de las variantes identificadas por los

flujos de trabajo desarrollados en esta tesis. GINO es una plataforma fácil de usar para la visualización e interpretación de variantes germinales que requiere licencia de uso.

Con el desarrollo de esta tesis se ha conseguido implementar las herramientas necesarias para la identificación de alto rendimiento de todos los tipos de variantes germinales, así como de una poderosa plataforma para visualizar dichas variantes de forma sencilla y rápida. El uso de esta plataforma permite a los no bioinformáticos centrarse en interpretar los resultados sin tener que preocuparse por el procesamiento de los datos con la garantía de que estos sean científicamente robustos. Además, ha sentado las bases para en un futuro próximo implementar una plataforma para el completo análisis y visualización de datos genómicos en la nube.

Palabras clave: NGS, Genómica, Variantes, Bioinformática, SNPs, Indels, CNVs, SVs, Platforma

Preface

The basis for this thesis originally stemmed from my passion for research and genomics, as well as my motivation for their clinical application. In addition to researching and achieving results, it is important to successfully implement those results for the benefit of society. For this reason, the completion of an industrial doctorate in collaboration with the company SEQUENTIA BIOTECH SL and the Universitat Politècnica de Catalunya was the best option to undertake this project.

Next generation sequencing technologies has made it possible to sequence human DNA in a relatively fast and affordable way. However, these technologies generate big data that is difficult to process and manage without a strong bioinformatics knowledge, creating a gap between data production and analysis. It is my goal and my passion not only cover this gap, but also to achieve it using the best tools to procure both a solid scientific base and easy usability.

First of all, I want to thanks all my colleagues at SEQUENTIA BIOTECH SL, but especially to my supervisors, Riccardo Aiese Cigliano and Walter Sanseverino, for laying the foundations of such an exciting project and giving me the support and tools to carry out this thesis satisfactorily. I would also like to thanks my tutor, Guadalupe Gómez Melis, for her advice and guidance throughout the development of this thesis. In addition, thanks to the Spanish Government because this project could not have been accomplished without their financial support (DI-17-09652, Ministerio de Ciencia e Innovación). And above all thanks to my family and friends, who supported me with affection and understanding throughout these years.

Contents

List of figures

List of tables

Acronyms and abbreviations

AB	Allelic Balance
aCGA	Array Comparative Genomic Hybridization
ACMG	American College of Medical Genetics and Genomics
aDNA	Ancient Deoxyribonucleic Acid
AF	Allele Frequency
AL	Allele
AS	Assembly
ASCII	American Standard Code for Information Interchange
ASE	Allele-specific expression
BAM	Binary Alignment Map
BED	Browser Extensible Data
BEDPE	Browser Extensible Data Paired-End
BP	Basepair
BQSR	Base Quality Score Recalibration
cDNA	Complementary Deoxyribonucleic Acid
CEO	Chief Executive Officer
CHR	Chromosome
CI	Confidence interval
ClinGen	Clinical Genome Resource
CNN	Convolutional Neural Networks
CNV	Copy Number Variation
CSO	Chief Scientific Officer
CSS	Cascading Style Sheets
CTX	Interchromosomal Translocation

dbNSFP	Database for nonsynonymous SNPs' functional predictions
dbSNP	Single Nucleotide Polymorphism database
DDD	Deciphering Developmental Disorders
ddNTP	Dideoxynucleotides
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic Acid
DP	Depth of coverage
DRAGEN	Dynamic Read Analysis for GENomics
EGA	European-Genome phenome Archive
ESP	Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FISH	Fluorescence in situ hybridization
FOD	Focal osseous dysplasia
FPGA	Field-programmable gate array
FS	FisherStrand
GA4GH	Global Alliance for Genomics and Health
GATK	Genome Analysis Toolkit
GiaB	Genome in a Bottle
GQ	Genotype Quality
GT	Genotype
GVCF	Genomic Variant Call Format
HOM	Homology
HTML	HyperText Markup Language
IGSR	International Genome Sample Resource
IGV	Integrative Genomics Viewer

Indel	Insertion and deletion
isoCNV	*In silico* optimization of copy number variant detection from targeted or exome sequencing data
Kb	Kilobase
Mb	Megabase
MGRB	Medical Genome Reference Bank
MLPA	Multiplex ligation-dependent probe amplification
MQ	RMSMappingQuality
MQRankSum	MappingQualityRankSumTest
mtDNA	Mitochondrial Deoxyribonucleic Acid
NCBI	National Center for Biotechnology Information
nDNA	Nuclear Deoxyribonucleic Acid
NGS	Next Generation Sequencing
NHLBI	National Heart, Lung, and Blood Institute
NIST	National Institute of Standards
NPV	Negative Predictive Value
OMIM	Online Mendelian Inheritance in Man
PMD	Postmortem DNA damage
PPV	Positive Predictive Value
QD	QualByDepth
RD	Read-depth
ReadPosRankSum	ReadPosRankSumTest
RefSeq	Reference sequence
RNA	Ribonucleic acid
RNA-seq	RNA sequencing

ROI	Region Of Interest
RP	Read-pair
SAM	Sequence Alignment Map
SB	Strand Bias
SNP	Single Nucleotide Polymorphism
SOR	StrandOddsRatio
SR	Split-read
SRA	Sequence Read Archive
SV	Structural variation
TCAG	The Centre for Applied Genomics
TP	True Positive
TS	Targeted Sequencing
UCSC	University of California Santa Cruz
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Chapter 1 | Introduction

1.1 Thesis outline

The current thesis has been developed by the student Rosa Barcelona Cabeza at the Bioinformatics Department of SEQUENTIA BIOTECH SL, under the supervision of Dr. Riccardo Aiese Cigliano, Chief Scientific Officer (CSO) and co-founder of the company, and Dr. Walter Sanseverino, Chief Executive Officer (CEO) and co-founder, and under the academic tutelage of Dr. Guadalupe Gómez Melis, professor at the Statistics and Operations Research Department of Universitat Politècnica de Catalunya-BarcelonaTECH.

This thesis has taken place with the financial aid for the training of doctors in companies (Industrial Doctors) contemplated in the State Training Subprogram of the State Program for the Promotion of Talent and its Employability, within the framework of the State Plan for Scientific and Technical Research and Innovation 2013-2016 (DI-17-09652, Ministerio de Ciencia e Innovación).

In this thesis we present multiple high-performance pipelines to identify each type of human genomic variation from next-generation sequencing (NGS) data and GINO, a platform to visualize and interpret variants. This chapter introduces the Genomics field, its value and limitations, and presents the motivation for the research. Chapter 2 goes through all the steps involved in improving and creating the tools to perform a fast and reliable pipeline for identifying germline genomic variants. Chapter 3 presents an integrated approach of genomic and transcriptomic data for accurate detection of genomic variants. Finally, Chapter 4 explains the implementation of GINO, a unique platform with a robust graphical environment to perform visualization and interpretation of variants.

1.2 Context

1.2.1 The human genome

A genome is the complete set of genetic material present in a cell or an organism, it consists of Deoxyribonucleic Acid (DNA) (or Ribonucleic Acid (RNA) in RNA viruses). The typical human genome consists of approximately 3 billion base pairs of DNA, divided among the 24 types of nuclear chromosomes (22 autosomes, plus the sex chromosomes, X and Y) and the mitochondrial chromosome.

Even if the human genome comprises around 3 billion base pairs, not all of them are functionally relevant. It has been estimated that only 20,000 protein-coding genes are present and their coding sequences (exons) comprise less than 2% of the genome. Most of the genome consists of non-coding DNA, while

some initially referred to it as “junk,” many of the non-coding sequences act as regulatory elements of gene activity and are of evolutionary importance [1].

In the human genome, genes are relatively sparse and distributed quite non randomly along the different human chromosomes, ranging from approximately 3 genes/Mb of DNA in gene-poor chromosomes to more than 20 genes/Mb in gene-rich chromosomes (excluding the Y chromosome and the mitochondrial chromosome).

There are coding and non-coding genes in the human genome. Most genes known or thought to be clinically relevant are protein-coding and their products comprise the list of enzymes, structural proteins, receptors, and regulatory proteins that are found in various human tissues and cell types. However, there are also genes that do not encode for proteins and they represent as many as a half of all identified human genes (Table 1). These non-coding genes have different functions in the cell and many do not have any identified function, yet.

Table 1. Characteristics of the Reference Human Genome. From Ensembl, database GRCh38, patch release 13 (accessed June 2021).

Length of the human genome (base pairs)	3,096,649,726
Number of known protein-coding genes	20,442
Number of non-coding genes	23,982

Multiple techniques have been developed to visualize the human genome (Figure 1). Karyotyping of patient chromosomes has been a valuable and routine clinical laboratory procedure for a half century, however the resolution of chromosomal changes detectable by karyotyping is typically a few megabases, falling well short of most pathogenic DNA variants (Figure 1). The ultimate resolution comes from direct sequence analysis, which enables the search for novel variants or mutations that might be of clinical importance.

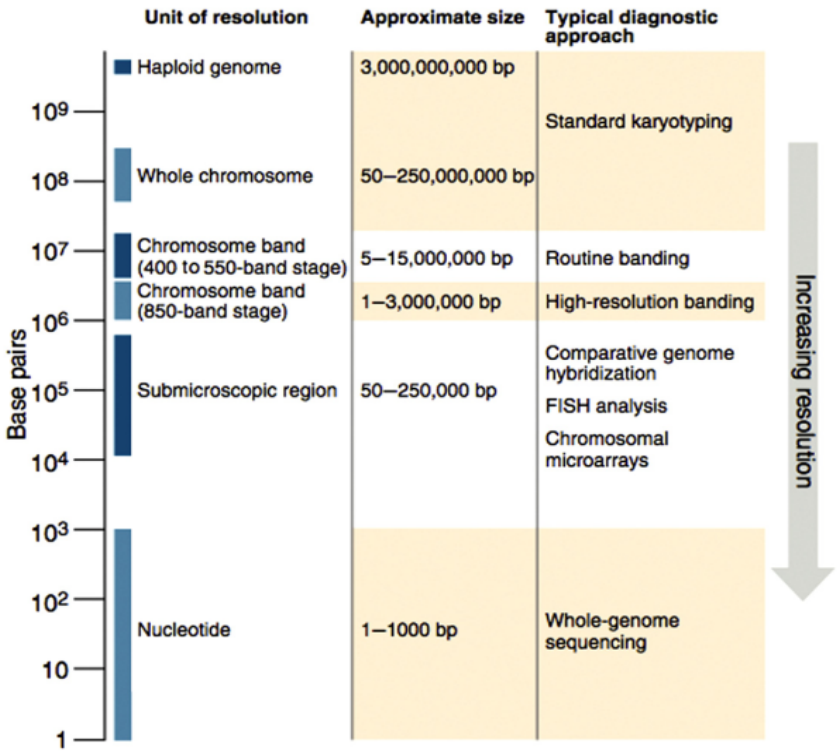

Figure 1. Spectrum of resolution in chromosome and genome analysis. The typical resolution and range of effectiveness are given for various diagnostic approaches used routinely in clinical and research practice. FISH, Fluorescence in situ hybridization [2, p. 546].

1.2.2 Variation in the human genome

A genomic variation is an alteration in the most common DNA nucleotide sequence within a population. With the completion of the initial reference human genome sequence, attention was turned to the discovery and cataloguing of variations among different individuals and among different populations [3]–[5]. Any given individual carries 4-5 million sequence variants and the majority of these variants are very rare, many of which probably exist in only a single or a few individuals [6]. The majority of variations are benign while some are protective, conferring an advantage against certain conditions, and others can be harmful, increasing susceptibility for a condition or directly causing a disease [7].

Identification of variations is important for both modern and ancient genomes. Detection of variants in modern human genomes allow us to improve the diagnosis and prognosis of certain diseases [8]–[11]. Analysis and characterization of ancient human genomes, also called human paleogenomics, allows us to understand past human evolution and provides new insights into a wide range of topics such as demography, migrations or human adaptation [12]–[16].

There are two different classifications of genomic variations, based on cell type and based on type of alteration. The first classification divides genomic variations in germline and somatic variants. A germline variant occurs within the germ cells (egg or sperm), such that the alteration can be passed to subsequent

generations. A somatic variant occurs in any of the cells of the body, except the germ cells and therefore is not inherited. Identification and study of both is important, germline variants interpretation focuses on pathogenicity of a variant for a specific disease or disease causality whereas interpretation of somatic variants should focus on their impact on clinical care [17]. However, for the sake of this thesis, we will focus only on the identification of germline variations. The second classification leads to three different alterations: Single Nucleotide Polymorphism (SNPs), insertions and deletions (indels) and structural variations (SVs). Specifically, this thesis is focused on the germline discovery of these three types of alterations (SNP, indels and SVs) in the human genome. Any and all of them can influence disease and thus must be accounted for in any attempt to understand the contribution of genetics to human health.

1.2.2.1 Single Nucleotide Polymorphism (SNPs)

Single nucleotide polymorphisms (SNPs) are defined as single base substitutions variation, where two or more different nucleotides can be observed within a given population. They are by far the most frequent type of variation in the human genome, occurring once every several hundred base pairs throughout the genome [18].

The biologic impact of SNVs in coding regions depends on their type: synonymous and non-synonymous. Synonymous SNPs are those that have different alleles that encode for the same amino acid while non-synonymous SNPs lead to a change of the encoded amino acids. In non-coding regions, the biologic effect depends on their impact on RNA processing or gene regulation [19]. The vast majority of SNPs are located in non-coding regions, this is due to the selection pressure that reduces the overall frequency of SNPs in coding DNA and in associated regulatory sequences [20].

1.2.2.2 Insertion and deletion (indels)

Indels refer to insertion and/or deletion of nucleotides into genomic DNA and include events less than 1 kb in length [21]. Insertions or deletions of sequences larger than 1 kb are categorized as copy number variants (CNVs) and they are more appropriately referred to as amplifications, duplications, or deletions.

Accurate identification of indels is critical in clinical diagnosis, as they are commonly implicated in constitutional and somatic diseases. Additionally, indels are a common mechanism of kinase activation in cancer, a feature exploited clinically by targeted therapy with kinase inhibitors [22].

1.2.2.3 Structural variants (SV)

Structural variation (SV) is generally defined as a genetic variation that occurs over a region of DNA approximately 1 kb or larger in size [21], [23] and includes both copy number variation (CNV) and chromosomal rearrangement events.

Copy number variation (CNV) refers to the gain or loss of specific regions of DNA larger than 1 kb. CNVs are an important class of genetic variation due to their wide-ranging impacts in human disease, including inherited syndromes and cancer-acquired mutations [24], [25].

Chromosomal rearrangements include insertions, inversions and translocations. An insertion is the addition of a novel sequence with respect to a reference genome. An inversion occurs when a segment of DNA is reversed in orientation with respect to the rest of the chromosome. Translocations occur when a segment of DNA changes its position, intra- or inter-chromosomally. Identification of SVs is critically important for the diagnosis and prognosis of both hematologic malignancies and solid tumors [26], [27].

1.2.3 DNA sequencing

DNA sequencing is the process of determining the nucleotide order of a given DNA polynucleotide chain and it has become the gold standard for identifying genetic variations [28].

1.2.3.1 Sequencing technologies

Scientific advances in DNA sequencing proceeded through four major technological revolutions: first generation sequencing (Maxam-Gilbert and Sanger sequencing), second generation or next generation sequencing (NGS, high throughput sequencing), third generation sequencing (3G) and fourth generation sequencing (4G).

1.2.3.1.1 First Generation Sequencing- Sanger Sequencing

Maxam-Gilbert sequencing was developed by Allan Maxam and Walter Gilbert in 1976–1977 [29]. It was the first widely adopted method for DNA sequencing alongside Sanger sequencing, however it is no longer in common use.

Sanger sequencing was developed by Dr. Frederick Sanger in 1977 [30], it is based on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during *in vitro* DNA replication. The ddNTPs are radioactively or fluorescently labeled for detection in gels or automated sequencing machines, respectively.

Sanger sequencing was used to sequence the first human genome in the Human Genome Project [31]. Nowadays, Sanger sequencing technology remains very useful for applications where high throughput is not required, for example to verify plasmid constructs or PCR products. In addition, it is used to validate next-generation sequencing data or for projects focused on single genes or regions.

1.2.3.1.2 Second Generation Sequencing - NGS

Second generation sequencing, also known as next-generation sequencing (NGS), refers to several different technologies of high-throughput DNA sequencing [32]. NGS technologies are different from the Sanger sequencing as they perform a high-throughput and massive analysis in parallel of multiple samples at much reduced cost [33]. Millions to billions of DNA nucleotides can be sequenced in parallel, increasing throughput and speeding up the sequencing process [34].

NGS technologies require preparing amplified sequencing libraries before sequencing amplified DNA clones [35]. Parallelization of a large number of sequencing reactions by NGS was achieved through the miniaturization of sequencing reactions and, in some cases, the development of improved microfluidics and detection systems [36]. Time spent in generating gigabase (Gb) size sequences by NGS was reduced from many years to only a few days or even hours, with the corresponding massive price reduction. For example, the Sanger sequencing of the human genome took almost 15 years costing more than 100 million dollars [31], while the genome of J.D. Watson (winner of the Nobel Prize in 1962) was sequenced by NGS in only 2 months with approximately the same coverage and for approximately a price 100 times lower [37].

1.2.3.1.3 Third Generation Sequencing

Third generation sequencing methods allow sequencing of single DNA molecules without the need for a template amplification step although the sequencing step itself still involves sequencing-by-synthesis. The lack of an amplification step provides advantages over second-generation: it prevents artifactual DNA mutations and strand biases introduced by even limited cycles of PCR and allows higher throughput, higher consensus accuracy, faster turnaround times and longer read lengths (by some platforms) that enhance complex SVs mapping and *de novo* contig and genome assembly [38].

1.2.3.1.3 Fourth Generation Sequencing

Fourth generation sequencing can be described as the sequence analysis of single DNA molecules without prior amplification; the sequencing is performed without DNA synthesis, and so is free of nucleotide labeling and detection steps. Nanopore-based technologies are the best examples of fourth-generation platforms [38]. A range of other novel technologies are still in the development stage. They will continue to make DNA sequencing even faster and less costly [39].

1.2.3.2 Sequencing strategies

These technologies generate large amounts of complex data consisting of strings of bases which are called reads. In Illumina sequencing platforms, there are two different types of reads: single-end and paired-end reads. Single-read sequencing means sequencing DNA from only one end while paired-end sequencing involves sequencing both ends of a fragment.

The level of coverage and resolution of these technologies can be easily tuned, providing a high degree of flexibility. The term coverage (or depth) refers to the average number of reads that align to known reference bases. Additionally, these technologies can be tuned to sequence only a subset of genomics regions, allowing researchers to focus time and expenses.

There are three types of sequencing strategies which depend on the amount and type of DNA sequenced: targeted sequencing (TS), whole exome sequencing (WES), or whole genome sequencing (WGS).

1.2.3.2.1 Targeted Sequencing

Targeted sequencing (TS) approach focuses on selected sets of genes or genomic regions. It is a rapid and cost-effective way to detect known and novel variants in the targeted regions. However, variant detection is limited to the pre-selected regions and chromosomal rearrangements cannot be identified [40].

1.2.3.2.2 Whole Exome Sequencing

Whole exome sequencing (WES) is a targeted sequencing approach that interrogates all the exonic regions within a genome. It also may be extended to target functional non-protein coding elements as well as specific candidate loci [40], [41]. Since the exome comprises less than 2% of the genome, it can be sequenced at a greater depth than the genome at a lower price. The greater depth provides more confidence in the detection of low frequency alterations. Exome sequencing also reduces data storage costs and allows a quicker, cheaper and easier data analysis. WES has increasingly become the favored approach, both for research and for clinical care [5], [42], [43]. However, whole-exome sequencing and targeted panels only interrogate part of the variants as they focus on reduced areas of the genome and lead to a non-uniform read-depth distribution among regions caused by biases in sample batches, GC content, and targeting probes [44]–[46]. This creates regions with high depth (a waste of sequencing power) and with low coverage which in turn can lead to missing variant calls. Furthermore, non-uniform coverage hinders the identification of copy number variations.

1.2.3.2.2 Whole Genome Sequencing

Whole-genome sequencing (WGS) determines the order of the nucleotides in the entire genome [40]. It allows the analysis of all types of variation in the entire genome, it can take advantage of longer reads and allows the most uniform depth of sequencing, increasing the accuracy in the identification of structural variants. However, it requires a much greater expense than the other sequencing approaches to obtain the depth required to achieve reliable results.

1.3 State-of-the-art

The new sequencing technologies have allowed the field of human genomics to reach, in a short time, very diverse areas, including biomedicine, clinical diagnosis or evolutionary biology. In fact, the complete sequencing of the human genome has allowed us to understand not only its sequence but also its organization, its genetic variations, the differential expression of its genes and several aspects of its transcriptional regulation. Also, the sequencing of ancient human genomes has enabled us to track frequency changes of genomic variation over space and time and to understand how migration and admixture events produced current patterns of genetic variation. The current sequencing platforms are faster and have higher throughput than Sanger sequencing method but also, they generate big data that need to be sorted, curated, integrated, analyzed and interpreted.

Identification of variations is not only enabled by DNA sequencers (hardware) but also by variant callers (software) that combine the reads obtained by sequencing to identify where and how an individual's genome differs from a reference genome. Very few variant callers are versatile enough to call all types of genomics variants because they require very different algorithms, increasing the computational cost and making it more difficult for scientists [47].

Due to the rapidly growing technology, there is a lack of qualified personnel to process sequencing data, generating a niche in the market in terms of analysis, data storage and interpretation. The vast majority of the existing tools require the scientist to have knowledge of programming and scripting to be able to execute them. Most geneticists and biologists either fail to process their own data because of its size and complexity or because it requires spending a large amount of time. In addition, due to the diversity of genomic applications, informaticians usually do not have all the knowledge required to fully understand the biological "problem" of the researcher, making it more complicated to reach a practical solution. More and more publications and funds have been directed to the study of omics, the genomics market reached a size of $17.2 billion in 2019 and it has been estimated to reach $31.1 billion by 2027 [48], confirming the importance of finding a solution to efficiently manage the large amount of data produced.

In this section, the main algorithms and pipelines used in germline variant calling will be reviewed. Firstly, we will go through the variant callers used to identify SNPs and indels (short variant discovery) and then, to identify structural variants (structural variant discovery) as they require quite distinct algorithms. Finally, the current available genomic analysis platforms will be assessed.

1.3.1 Short variant discovery

Short variant discovery is a multi-step process by which SNPs and indels in sequence data are identified. A wealth of pipelines has been and are being developed for accomplishing this challenging task. Each of the pipelines mainly consist of the quality assessment, read alignment, variant discovery, variant filtering

and annotation [49], [50] (Figure 2), and different combinations of tools belonging to each step above-mentioned will result in varying performance of the pipelines affecting the interpretation of the short variants calls.

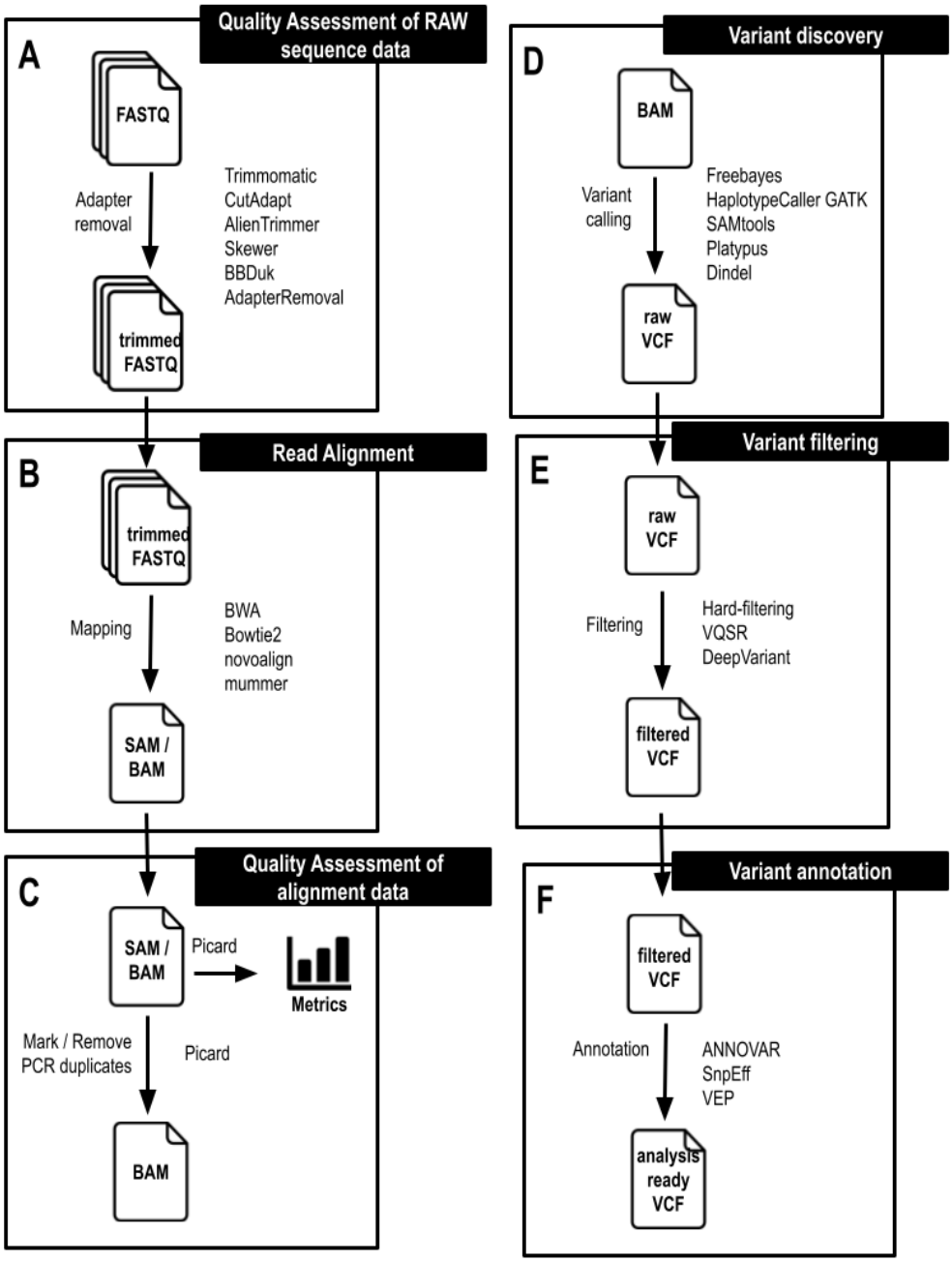

Figure 2. Short variant discovery pipeline. The short variant discovery pipeline mainly consists of six steps: A) Quality Assessment (QA) of RAW sequence data where low quality reads, uncalled bases, adapters and contaminant sequences are removed. B) Read alignment to the reference genome. C) Quality assessment (QA) of alignment data to verify sufficient sequencing coverage and marking or removal of duplicated reads. D) Variant discovery to identify short variants in sequence data. E) Variant filtering to reduce the false discovery rate. F) Functional variant annotation.

1.3.1.1 Acquisition of raw sequence data: the FASTQ file format

The first step before starting the variant calling process is obtaining the data. The datasets used in variant calling usually come from the sequencing of biological samples with any of the available sequencing technologies, although simulated data can also be a starting point [51]. The most common format file to start the process of variant calling is FASTQ format (Figure 3), thus, data from sequencing technologies in other formats, like the binary base call format (BCL) produced by Illumina should be transformed to this format.

The FASTQ format (Figure 3) is text-based and stores sequencing read data and its base quality score [52]. This type of file usually contains millions of records that belong to each of the sequenced reads. Each record is made of four lines, the first one corresponds to the ID of the read with information about the flow cell, lane, tile, tile coordinates and barcode, the second one hosts the nucleotide sequence of the read, the third one is a separator and the fourth one contains the read quality scores per base.

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEE
```

Figure 3. Example of a single entry in a FASTQ file.

The quality of each base is codified in ASCII code (American Standard Code for Information Interchange) that translates each character into a number. This quality is given by the PHRED score (Q) that considers the probability (P) of that base of being a sequencing error [53]. The higher the quality score the lower the probability of sequencing error according to the following formula:

$$Q = -10 \log10 P$$

1.3.1.2 Quality assessment of raw sequence data

The first step of the analysis is the quality control, which is performed to remove low quality reads, uncalled bases, adapters and contaminant sequences. None of the available sequencing technologies are absent from making errors [54].

The most commonly used tool for evaluating and visualizing the quality of sequence data is FastQC [55], which provides comprehensive information about data quality, including but not limited to per base sequence quality scores, GC content information, sequence duplication levels and overrepresented sequences. Alternative tools for quality assessment of FASTQ files are fastqp [56] or PRINSEQ++ [57].

Adapters are artificial short DNA oligonucleotides, generally of known sequence. They are used with the scope of binding the DNA fragments to the flow cell. Since the adapter sequences are synthetic, adapter contamination often leads to NGS alignment errors and an increased number of unaligned reads.

Furthermore, recurrent untrimmed adapters at the same genomic position can lead to spurious variant calls. Hence, any adapter sequences need to be removed before mapping. There are plenty of tools for adapter removal, namely, Trimmomatic [58], CutAdapt [59], AlienTrimmer [60], Skewer [61], BBDuk [62] and AdapterRemoval [63]. In addition to adapter removal, trimming can be performed with these tools to discard low quality reads or low quality bases.

1.3.1.3 Read alignment

Reads that have passed the quality control are mapped against a reference genome. This step tries to determine the most likely region of the reference genome for each of the reads [64]. The aligner usually takes the reference genome of the species (in our case, *Homo sapiens*), its index and the FASTQ reads. For *Homo sapiens*, the most current and widely used reference sequences are GRCh37 (hg19) and GRCh38 (hg38). Depending on the mapping software used, additional files can be needed [64].

This step is one of the most time consuming and computational resource demanding steps, since the read mapper has to consider a high number of sequences and their similarity before assigning them to a specific region of the genome [65]. Repetitive regions make this task very difficult since the aligner might not be able to assign a read to a specific region [66]. Variants can also be a source of misalignment, for instance, indels can make that a substantial part of the read does not match the reference genome [67], [68].

There is an abundance of tools for alignment of sequences to the reference genome, some of them are BWA [69], Bowtie2 [70], novoalign [71], and mummer [72]. They differ on the algorithm used, the sensitivity, the memory requirements, the speed, and the sequence length requirements. The most widely used tools are BWA and Bowtie2.

The mapping step usually produces a Sequence Alignment Map (SAM) file consisting of the information of the reads aligned to the genome [73]. The SAM format is a text-based format, which contains a header and an alignment section having eleven mandatory fields with information relative to the alignment. Since this type of data can be very big and not very efficient to work with, SAM files are converted to BAM format, a binary version that supports quick retrieval of alignments [73].

At this time the alignment can be visualized using an alignment viewer for instance Integrative Genomics Viewer (IGV) [74]. The next step in the variant calling process is to sort the BAM file, so that its content gets to an arranged disposition and can be used with other sorted files containing chromosome information. This process also impacts greatly on the speed of later computations since the programs can focus just on the subset of rows of the regions of interest without the need of reading the whole file.

1.3.1.4 Quality assessment of alignment data

Quality control of alignment data should be performed prior to variant discovery to evaluate key sequencing metrics and to verify that sufficient sequencing coverage was achieved. The most used tool for this purpose is Picard [75].

Next, duplicated reads are marked or removed [76]. To that end, Picard [75] can also be applied. After the DNA is cut into pieces and adapters are ligated to each end of the fragments, PCR amplification occurs [32]. The sequencing of two or more duplicates of the same DNA sequence is what is called PCR duplicate. They can have errors resulting from the PCR application process and therefore affect allele frequencies since the reads supporting that mutation would be present in a higher number than others [77].

In ancient DNA (aDNA) samples it is also necessary to separate between authentic aDNA and modern contaminant DNA from microorganisms or present-day humans. After death, tissues are colonized by microbial decomposers, which can introduce microbial DNA contamination [78]. In addition, despite extensive precautions to avoid contamination during excavation and laboratory sample preparation, many ancient samples show contamination from living humans [79]–[82]. Since only minute amounts of DNA tend to be preserved, even a small contamination can overwhelm the original DNA. This is particularly problematic for contamination derived from modern humans due to their high genetic similarity with ancient humans.

One way to differentiate between aDNA and present-day contaminants is to evaluate the postmortem DNA damage (PMD) signatures from the read alignments. The PMD most commonly associated with aDNA is cytosine deamination at the single-stranded ends of aDNA [83], which has been shown to increase over time unlike other potential diagnostic patterns [84], [85]. The cytosine deamination pattern consists of cytosine to thymine substitutions that increases toward the 5′ end of the sequence reads, resulting in a complementary guanine to adenine pattern in the 3′ end caused by enzymatic repair [83], [86]. Several tools have been developed to detect these PMD patterns, including mapDamage [87] or DamageProfiler [88]. In addition, mapDamage [87] can be used to recalculate the quality scores of bases likely to be affected by PMD to mitigate its impact on subsequent analysis. However, while observing this pattern suggests the presence of aDNA, it does not discard the presence of modern DNA contamination. Therefore, specific tools have been developed to separate ancient DNA from present-day contamination, such as PMDtools [89], Schmutzi [90], ContamLD [91] or AuthentiCT [92]. PMDtools is an effective method to isolate aDNA using a likelihood-ratio test and has been the tool of choice in multiple studies [93]–[97].

At this point, the files are ready to proceed to most of the variant calling softwares [98], [99]. However, some variant callers such as Genome Analysis Toolkit (GATK) need an extra step that helps to increase the precision of the results: base quality score recalibration (BQSR) [76], [100], [101], which adjusts the base quality scores of sequencing reads applying machine learning to detect and correct any systematic bias.

1.3.1.5 Variant discovery

After all the preprocessing steps are done, variant identification can take place. Several approaches have been implemented in variant callers to identify variants in sequence data. Many algorithms use a Bayesian probabilistic approach such as Freebayes [98], GATK [102], SAMtools [73], Platypus [103] or Dindel [104]. Other software uses a Poisson-binomial distribution like LoFreq [105], an approximation based on frequencies such as SNVer [106] and mixed methods based on heuristics and statistics like VarScan [107]. In recent years, other methods have arisen due to the development and implementation of artificial intelligence, DeepVariant [99] is the first variant caller based on the deep convolutional neural network.

With many variant callers available, several benchmarking studies have been conducted to assess the performance of different variant calling pipelines in detecting accurate variants. Liu *et al.* compared the performance of four variant callers and reported that GATK performed best on real and simulated exome data [108]. In Pirooznia M *et al.*, a study based on the read-depth, allele balance and mapping quality, GATK outperformed SAMtools on low coverage exome data [109]. Kim BY *et al.* performed a comparative study of four variant callers (GATK, SAMtools, Dindel, and Freebayes) using human WES data and reported that GATK had the highest sensitivity for indel identification and that the performance of four algorithms was unaffected by indel size [110]. However, further studies have reported varying performance depending on indel size [67], [111]–[113]. Pei *et al.* assessed three germline variant callers (GATK, Sentieon and DeepVariant) and reported to have similar performance on NGS data while DeepVariant outperformed the others in indel calling with TS data [114].

For germline variant calling, GATK's HaplotypeCaller is one of the most commonly used callers [76], [100]. Sentieon has been designed as an accelerated software for GATK [115] and has become one of the most popular commercial variant callers. Sentieon reduced computing resource consumption and shortened the computation time without compromising the accuracy of the calling [114]. In addition, Sentieon showed the highest SNP recall and the highest indel precision using WGS data in the precisionFDA Truth Challenge v1 [116] and the best overall accuracy on identification of variants in difficult-to-map regions for PacBio sequencing technology and for a multi-technology approach (combination of Illumina, PacBio HIFI and Oxford Nanopore) in the precisionFDA Truth Challenge v2 [117].

The detection of variants constitutes in itself the main objective of many studies and projects [6], [118]–[120]. The standard format of representation of variants is Variant Calling Format (VCF). It is composed of a header with information about the columns and the parameters of the detected variants such as position, reference and alternative alleles, genotype, etc. [121].

1.3.1.6 Variant filtering

Following the variant calling step, raw VCF should be filtered to reduce the number of bad called variants. Most variant callers are tuned to be very sensitive, more tolerant to false positives than false negatives [122].

The current state-of-the-art filtration methods include hard-filtering [76], Gaussian Mixture Models as Variant Quality Score Recalibration (VQSR) from GATK [76], Random Forests [123] or Convolutional Neural Networks (CNNs) like DeepVariant [99].

Hard-filtering consists of choosing specific thresholds for one or more parameters and filtering out any variants above or below the set threshold. VQSR uses machine learning algorithms to learn from the data what are the profiles of variants that are likely to be real in a particular dataset. It assigns accurate variant quality scores to each variant that are then used for filtering. VQSR is more powerful than hard-filtering, however it requires multiple samples, a large number of variants and well-curated known variant resources. Random forest classifiers are trained on polymorphic variants to separate true variants from false positive artifacts. DeepVariant is a deep learning approach to filter variants based on CNNs.

1.3.1.7 Variant annotation

Variant annotation is the process of assigning functional information to genomic variants. It is a critical step as it can have a strong influence on the ultimate conclusions of a study. Improper or incomplete annotations can cause both to miss potentially relevant variants and to dilute interesting variants in a pool of false positives.

There are many different types of information that can be associated with variants, including but not limited to (i) genes or transcripts affected by the variants, (ii) the location of the variant (in coding sequence, in intronic sequence, in regulatory regions, etc.), (iii) the protein sequence consequence of the variant, (iv) the effects on protein structure and function, and (v) matching known variants in databases (dbSNP [124], 1000 Genomes Project [6], gnomAD [125], ClinVar [126], etc.).

The annotation process can be performed using a variety of annotation software and a variety of transcript sets (RefSeq transcript set [127], Ensembl transcript set [128], etc.). Both the choice of the annotation tool and transcript set can have a significant impact on variant annotation [129]. The most widely used annotation tools are ANNOVAR [130], SnpEff [131] and Variant Effect Predictor (VEP) [132].

1.3.2 Structural variant discovery

Structural variant (SV) discovery is the process of identification of both copy number variants and chromosomal rearrangements in sequence data. It mainly consists of a multi-step process including quality assessment, read alignment, SV discovery, filtering and annotation (Figure 4). While the quality assessment and the read alignment are the same used in the identification of short variants, SV discovery, filtering and

annotation require different algorithms. The choice of the different combination of algorithms will influence the performance of the SV discovery and the final interpretation of the results.

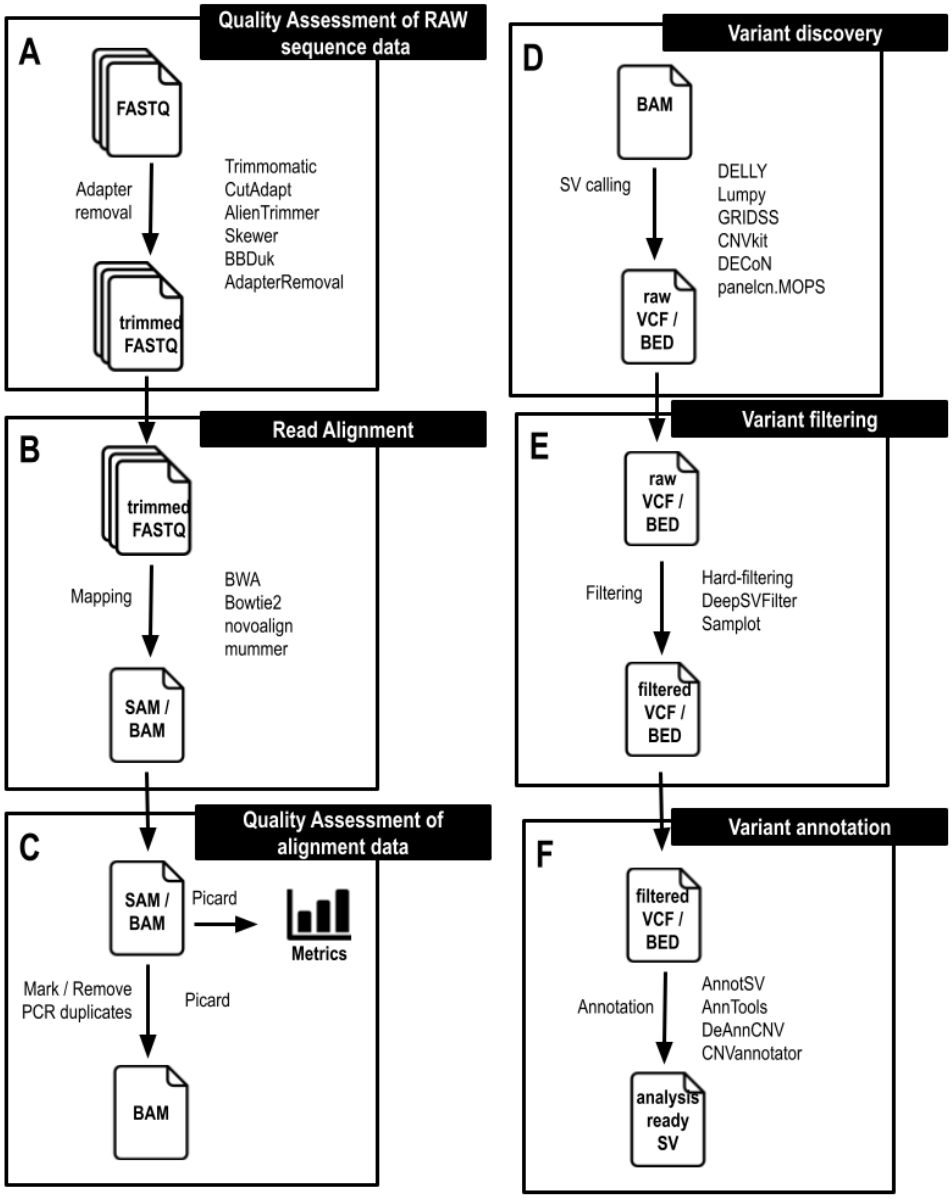

Figure 4. Structural Variant discovery pipeline. The SV discovery pipeline mainly consists of six steps: A) Quality Assessment (QA) of sequence data where low quality reads, uncalled bases, adapters and contaminant sequences are removed. B) Read mapping to the reference genome. C) Quality assessment (QA) of alignment data and marking or removal of duplicates reads. D) SV discovery in sequence data. E) SV filtering to reduce the false positive calls. F) Functional SV annotation.

1.3.2.1 Variant discovery

Many tools have been developed to detect structural variants from NGS data. These tools follow one (or a combination) of four methods: read-pair (RP), split-read (SR), read-depth (RD) and assembly (AS).

Read-pair (RP) approach involves the identification of discordance between mapped paired-reads. These discordantly mapped paired-reads may be indicators of different types of SV if they are: (i) further apart (or closer together) than expected (deletions or insertions), (ii) in wrong orientation (inversion), (iii) in incorrect order (tandem duplication) or (iv) mapping in different chromosomes (translocation) [133]. BreakDancer [134] was one of the first software using the RP method but it has been also implemented in other packages, such as DELLY [135], LUMPY [136], and Prism [137].

Split read (SR) method uses paired-end reads where only one read of the pair has a reliable mapping [138]. The unmapped reads are a potential source of breakpoints. Mapping of reads that span across a breakpoint of a SV provides the precise breakpoints with base accuracy. SR method is powerful for identifying small and medium-size variants such as insertions, deletions, inversions and translocations. However, SR approach is limited for large size variants or those in repetitive regions [139]. Some tools based on this approach are Pindel [140], Gustaf [141] and CREST [142].

Read-depth (RD) approach uses the correlation between depth of coverage and the copy number of the region [143]. It can only detect CNV being more effective for large size CNV, which are hard to detect with RP and SR methods [144]. In addition, RD can detect the exact number of CNV while RP and SR can only report their position. CNVnator [145], CNVkit [146], panelcn.MOPs [147] and DECoN [148] are some of the tools based on the RD approach.

The assembly (AS) method detects SV by aligning the contigs, assembled with the entire or unmapped sequencing reads, to the reference sequence [143]. All forms of SV can be detected by AS, however, they are less used in CNV detection due to their high demand on computational resources. The AS method is adopted by some tools such as Magnolya [149] and FermiKit [150].

Each of the four methods mentioned above has its own strengths and limitations. For this reason, there are a number of tools which have been implemented with more than one method aiming for higher specificity and sensitivity. Some tools based on the combined approach are DELLY [135], LUMPY [136], Manta [151] and GRIDSS [152].

The selected DNA sequencing technology has an enormous impact on the performance of the algorithms. Since WES and TS only cover a small portion of the whole genome, it is far more challenging to detect SV, especially if the breakpoints are not in the capture regions. While WGS technologies allow to detect all types of SV, WES and TS can only be used to detect CNV.

The performance of 69 algorithms for calling germline SVs from WGS data has been evaluated [153]. There, GRIDSS has been shown to call SVs for both simulated and real datasets with high precision and recall, being the best algorithm for identifying deletions, and the fourth for identifying duplications and inversions [153]. It has also shown a short run time and has achieved one of the highest accuracy values for

calling breakpoints for all size ranges of deletions and duplications [153]. In another comprehensive evaluation of 10 SVs callers, GRIDSS has been found to be one of the two best performing algorithms for SVs detection [154]. GRIDSS has been the algorithm of choice to detect germline SVs in the Medical Genome Reference Bank (MGRB) cohort [155].

Regarding the evaluation of germline CNVs calling from WES and TS data, DECoN [148] has shown a high performance [156], [157]. However, its performance is highly dependent on the selected parameters which should be optimized for each specific dataset to maximize sensitivity [157] and should not be used directly with data produced differently, i.e. with different sequencing technologies, targeting probes or capture protocol [157].

The different SV variant callers store the variants in different formats: VCF as for short variant discovery, Browser Extensible Data (BED) format [158] or Browser Extensible Data Paired-End (BEDPE) format [159]. BED format is a flexible way to represent genomic features and annotations that supports up to 12 columns. BEDPE format is a modified version of BED format that allows storing inter-chromosomal features.

1.3.2.2 Variant filtering

SV calling algorithms still show a high false positive detection rate, so some filtering steps must be taken. There are computational approaches to filter false positive SV such as DeepSVFilter [160], a deep learning based tool designed for WGS data. Furthermore, manual curation can also be performed: SV can be validated by manually inspecting the aligned reads around the region using Samplot [161] or by choosing thresholds for one or more metrics of a specific algorithm and filter out any SV that do not meet that threshold.

1.3.2.3 Variant annotation

There are plenty of tools to functionally annotate SV, namely AnnTools [162], DeAnnCNV [163] or CNVannotator [164]. Among them is AnnotSV [165] that provides the most complete panel of annotation sources to date. It performs gene-based annotation, annotation with features overlapping the CNV, and annotation of the breakpoints. In addition, AnnotSV classifies SV according to their pathogenicity into one of the 5 classes proposed by the American College of Medical Genetics and Genomics (ACMG) guidelines [166], [167]: benign, likely benign, variant of unknown significance (VUS), likely pathogenic or pathogenic.

1.3.3 Genomic analysis data platforms

The large amount of data obtained from massive sequencing, especially if a large number of samples are analyzed, leads to two important obstacles in the field of genomics: (i) informatic issue, mainly in data storage and management and (ii) scientific issue, particularly in the interpretation and prioritization of data.

To solve these obstacles, multiple platforms have been developed to perform genomic analysis, both for the identification of variants and for their visualization and interpretation. Genomic platforms improve the management of the high amount of data produced by NGS technologies and provide users with automation, reproducibility and long-term data storage. The most used platforms to date are Illumina Dynamic Read Analysis for GENomics (DRAGEN) Bio-IT Platform [168] and VarSome [169].

DRAGEN Bio-IT Platform [168] provides short and structural variant discovery for both genome and exome sequencing data under user's license. It implements a field-programmable gate array (FPGA) hardware technique to dramatically speed up the analysis process, reducing the runtime from hours to minutes. Users can create their own pipeline within the platform, which provides a great level of flexibility and customization. However, this also complicates the use of the platform as it requires a high level of knowledge in genomics and variant calling algorithms, as well as a minimal knowledge of bioinformatics to use its command line interface. Furthermore, the DRAGEN platform does not perform full annotation of variants and does not have a variant inspector to browse, filter and prioritize the results. It is mainly focused on users who want to benefit from its computational resources, variant calling algorithms or its ultra-fast analysis and not on users whose main interest is the interpretation of the results.

VarSome [169] is a commercial web-based tool that allows short and structural variant discovery, annotation and interpretation for genomes, exomes and gene panels. It is intended for users whose main interest is the interpretation of results rather than data processing. VarSome offers cloud-based pipelines whose parameters are pre-configured and fixed. Its easy-to-use graphical interface allows users to browse and filter the variants of interest in a variant table and a genome browser. Even if it is a powerful and complete platform, there is always room for improvement. Variant discovery pipelines can be optimized, especially for identification of structural variants. Regarding the graphical interface, more features can be added to improve variant prioritization, such as user-specific allele frequencies. In addition, the usability can be improved to facilitate the user experience in the variant table as there are many subsections and tabs that complicates the interpretation.

Depending on factors such as the flexibility required during variant calling or how complete the annotation needs to be, users should choose the platform that best meets their needs and expectations.

1.4 Objectives

As it is possible to understand from the state of the art of genomics, the data obtained from the massive sequencing cannot be of interest without using informatics for their analysis and interpretation. Although platforms already exist to call and interpret genomic variants, there is still a lot of work ahead. Regarding the identification of variants, there is a wide variety of algorithms and tools available that need to be reviewed and optimized to obtain the most appropriate approach to ensure optimal performance. Concerning the interpretation of these variants, powerful platforms are needed that, in an accessible and scientifically robust way, help the researchers to extract interesting data from their experiments much faster.

Given these considerations, the main objective of this thesis is the development of workflows to identify germline variants from NGS data and the implementation of GINO, a platform to interpret the identified variants. The development of this platform with a user-friendly graphical interface, would make it possible to democratize, or make accessible, bioinformatics so that researchers, hospitals and scientific institutions can use genomic technologies to solve their scientific problems and advance knowledge to accelerate the development of new strategic scientific advances.

There are three specific objectives in this thesis that correspond to chapters 2, 3 and 4 respectively and are the following:

1. Development of genomics tools for the identification of variants of the human germline.
2. Integration of genomic and transcriptomic data to improve variant discovery.
3. Development of a genomic data analysis platform to interpret genomic variation.

A full description of the objectives in each thesis chapter is presented below.

1.4.1 Genomics tools

The main objective of this chapter is to develop fast and reliable variant discovery pipelines for identifying, or calling, SNPs, indels, CNVs and chromosomal rearrangement events from WGS, WES or TS data.

A relatively large number of open-source variant calling tools is now available, most of them are specific for one or few different types of alterations and feature different algorithms, filtering strategies and different outputs. However, the literature offers limited guidance to efficiently select the tools able to meet the standards of good clinical practice.

Without hesitation, the critical point of these analyses lies in the statistical power of the identification of variants, particularly for CNVs and chromosomal rearrangement events. In addition, the sensitivity and duration of the analysis are two of the most important characteristics that are taken into account for this type of processing.

1.4.2 Data integration

The integration of omics has an enormous potential that has been exploited in a wide range of research areas [170]. Several algorithms have been already developed to detect somatic variants using both WES and RNA-seq data [171], [172]. To our knowledge, there is not yet a tool for calling germline variants using WES and RNA-seq data in an integrated fashion.

Thus, we propose to integrate WES and RNA-seq data for germline short variant discovery. This will provide an orthogonal method to validate genomic variants and will allow to identify new variations in significantly expressed genes and outside the target regions of the WES analysis.

1.4.3 Genomics data analysis platform

The objective of this chapter is the development of a platform (GINO) to visualize and interpret genomic variations.

The germline variants identified through the workflows developed in the previous chapters will be displayed in a unique computer infrastructure with a simple, accessible and robust graphical environment. Users will be able to browse the results of all the samples from an experiment, visualize samples one at a time and even obtain comparative results between family members. This will allow them to interpret results in a more convenient and easier way and, consequently, to extract more easily conclusions from experiments. Moreover, users will avoid data storage issues as all data will be stored in the cloud.

A bioinformatic platform with these characteristics could lead to a paradigm shift where researchers do not use their time in the production/extraction of the data but in the interpretation of them. In the near future and outside the context of this thesis, this platform will continue its development to run the complete genomic analysis in the cloud.

Chapter 2 | Genomics tools

This chapter describes the different workflows implemented in this thesis to identify all the variations of the human germline. First, short variant discovery pipelines for both modern and ancient DNA samples are described. Then, structural variant discovery pipelines for modern DNA samples are detailed, including a pipeline for the identification of CNVs and chromosomal rearrangements from WGS data and a pipeline for *in silico* optimization of CNV detection using WES or TS data (isoCNV).

2.1 Short variant discovery

After reviewing the best performing algorithms and tools to identify germline SNPs and indels (see section 1.3.1), three workflows have been implemented: (i) a workflow for the analysis of modern DNA of individual samples sequenced using whole genome, whole exome or targeted sequencing strategies, (ii) a workflow for modern DNA of parent-child trios and (iii) a workflow for ancient DNA of individual WGS samples.

2.1.1 Modern DNA

2.1.1.1 Implementation of short variant discovery per sample

The workflow for modern short variant discovery from individual samples has been developed using Python 3.7. Although this workflow relies on existing algorithms and tools, as far as we know, they had not yet been brought together in an automated pipeline to perform the complete analysis of short variant discovery, from sequencing data to analysis-ready SNPs and indels.

The input to the pipeline is the raw sequence data for the sample of interest in FASTQ format and, for whole exome or targeted sequencing data, also their corresponding targeted regions in BED format. The output consists of a VCF file and table of variants in TXT format, both containing the SNPs and indels identified in the sample. Results are obtained in about 2,5 hours for 100x WES analysis (8 threads) and in 10 hours for 15x WGS analysis (8 threads).

2.1.1.1.1 Quality assessment of raw sequence data

Adapters and low quality bases are trimmed from raw sequence data in FASTQ format using Trimmomatic [58] (Figure 5). The minimum base quality thresholds depend on the sequencing platform used, being 25 for Illumina and 17 for Ion Torrent. After trimming, reads having a size less than 35 bp are excluded from further analysis in order to reduce the fraction of spurious alignments.

To assess the impact of the trimming process on the sequence data, the quality metrics of the reads before and after trimming are obtained with FastQC [55] (Figure 5).

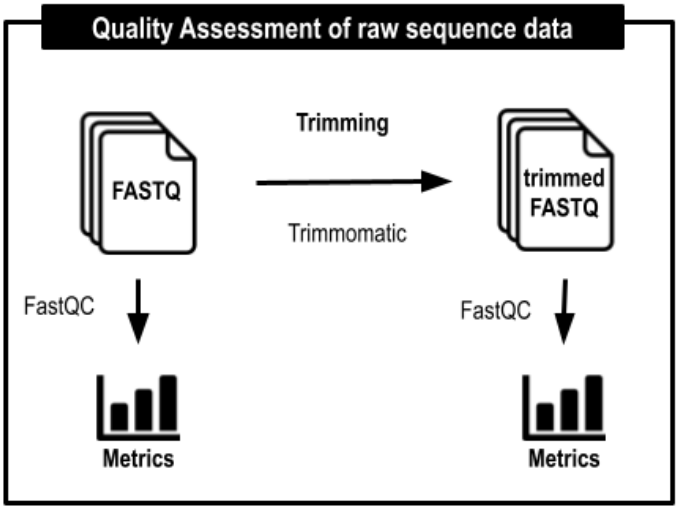

Figure 5. Quality assessment of raw sequence data in the short variant discovery of modern DNA. Trimmomatic [58] is applied to remove adapters and low-quality bases from raw sequence data. FastQC [55] is applied before and after trimming to assess the impact of the trimming process.

2.1.1.1.2 Read alignment

The high-quality reads resulting from the quality assessment and trimming process are aligned to the reference genome using the accelerated version of the BWA-MEM algorithm [69] developed by Sentieon [115]. Then, the aligned reads are sorted using the Sentieon sort utility v202010.02 [115] (Figure 6). Both GRCh37 (hg19) and GRCh38 (hg38) human genome assemblies can be used as reference genomes.

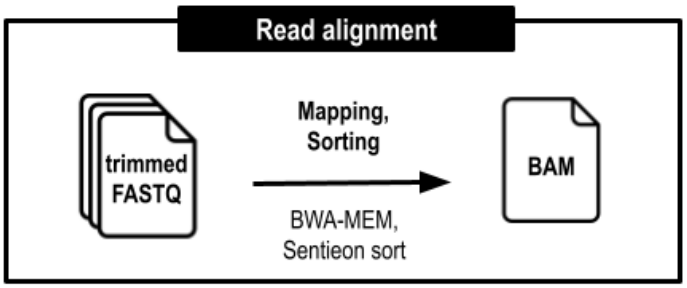

Figure 6. Read alignment in the short variant discovery of modern DNA. The trimmed sequence data is aligned to the reference genome and then sorted using the BWA-MEM algorithm and the Sentieon sort utility respectively.

2.1.1.1.3 Quality assessment of alignment data

Quality metrics of the alignment data are obtained using CollectHsMetrics and GcBiasMetrics from Picard [75] (Figure 7). The minimum threshold criteria for a sample to be considered of good quality are described in Table 2.

Table 2. Overview of the default quality thresholds for alignment data.

Field name in Picard [75]	**Description**	**Tool**	**Threshold criteria**
ZERO_CVG_TARGETS_PCT	The fraction of targets that did not reach coverage=1 over any base.	CollectHsMetrics	<= 2%
PCT_EXC_OVERLAP	The fraction of aligned bases that were filtered out because they were the second observation from an insert with overlapping reads.	CollectHsMetrics	<= 5%
PCT_EXC_DUPE	The fraction of aligned bases that were filtered out because they were in reads marked as duplicates.	CollectHsMetrics	<= 5%
FOLD_80_BASE_PENALTY	The fold over-coverage necessary to raise 80% of bases in "non-zero-cvg" targets to the mean coverage level in those targets.	CollectHsMetrics	<= 1.5%
PCT_TARGET_BASES_10X	The fraction of all target bases achieving 10X or greater coverage.	CollectHsMetrics	> 95%
PCT_TARGET_BASES_20X	The fraction of all target bases achieving 20X or greater coverage.	CollectHsMetrics	> 95%
PCT_TARGET_BASES_30X	The fraction of all target bases achieving 30X or greater coverage.	CollectHsMetrics	> 90%
PCT_TARGET_BASES_100X	The fraction of all target bases achieving 100X or greater coverage.	CollectHsMetrics	> 90%
AT_DROPOUT	A measure of how undercovered <= 50% GC regions are relative to the mean.	CollectHsMetrics	<= 5%
GC_DROPOUT	A measure of how undercovered >= 50% GC regions are relative to the mean.	CollectHsMetrics	<= 5%
AT_DROPOUT (For Illumina samples)	Illumina-style AT dropout metric.	GcBiasMetrics	<= 5%
GC_DROPOUT (For Illumina samples)	Illumina-style GC dropout metric.	GcBiasMetrics	<= 5%

Duplicated reads are removed with LocusCollector and Dedup algorithms from Sentieon software v202010.02 [115] (Figure 7) which are based on Picard's MarkDuplicates tool [75]. The base quality score recalibration (BQSR) is then performed with Sentieon's QualCal algorithm [115] (Figure 7) using three different resources as known sites: (i) the Single Nucleotide Polymorphism database (dbSNP) human build

152 [124], (ii) the 1000 Genomes Phase I indel calls [6] and (iii) the Mills and 1000G gold standard indels [6], [173]. Known sites are used by the QualCal algorithm to ensure that known locations do not get artificially low-quality scores.

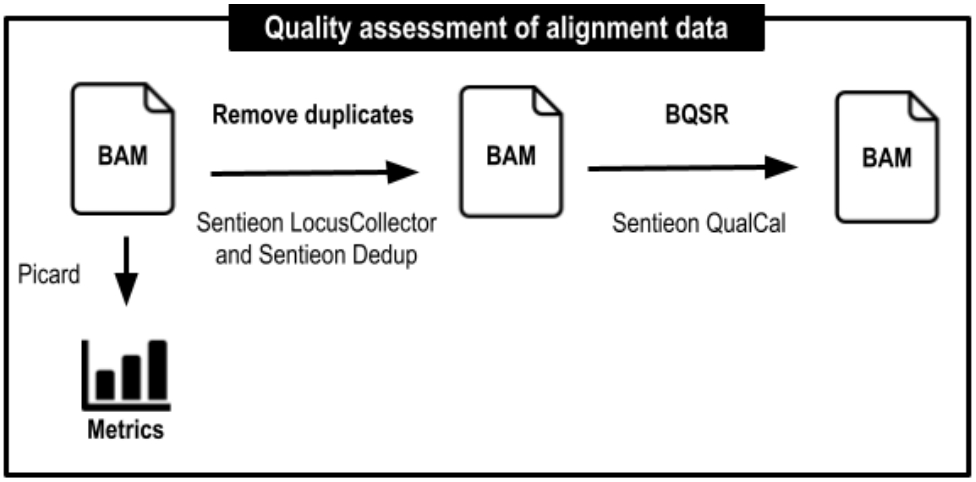

Figure 7. Quality assessment of the alignment data in the short variant discovery of modern DNA. Duplicate reads are removed with Sentieon LocusCollector and Sentieon Dedup [115]. Then, base quality score recalibration (BQSR) is performed with Sentieon QualCal [115].

2.1.1.1.4 Variant discovery

Depending on the DNA sequencing strategy used to obtain the data, genomic variants are called in the whole genome (WGS strategies) or only in the target regions provided in the BED file containing the capture bait locations (WES and TS strategies). By default, there is a 200 bp upstream and downstream padding of the target regions that can be tailored to the user's needs. In fact, due to the library preparation method, regions contiguous to the targets can also be captured and sequenced.

Short germline variants are called with Haplotyper algorithm from Sentieon v202010.02 [115] (Figure 8), which is the accelerated version of HaplotypeCaller algorithm from GATK 4.0 [102]. The dbSNP human build 152 [124] is used to label known variants.

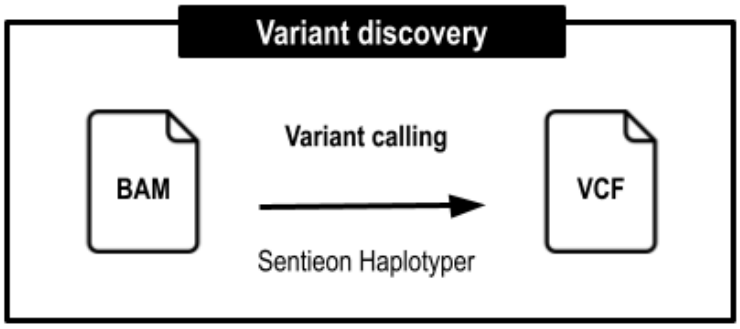

Figure 8. Variant calling in the short variant discovery of modern DNA. Germline variant calling is performed with Sentieon Haplotyper [115].

2.1.1.1.5 Variant filtering

Variant filtering is performed using the hard-filtering method [76] (Figure 9). The specific thresholds used for a variant to be considered a high-quality variant are described in Table 3.

Table 3. Overview of the default hard-filtering thresholds for short germline variants.

Field name in Sentieon [115]	**Description**	**Threshold for SNPs**	**Threshold for indels**
Depth of coverage (DP)	The number of filtered reads that support each of the reported alleles.	>= 6	>= 6
Genotype Quality (GQ)	The Phred-scaled confidence that the genotype assignment is correct.	>= 20	>= 20
QualByDepth (QD)	The variant confidence divided by the unfiltered depth of non-hom-ref samples.	>= 2	>= 2
FisherStrand (FS)	The Phred-scaled probability that there is strand bias at the site.	<= 60	<= 200
StrandOddsRatio (SOR)	Estimator of strand bias using a test similar to the symmetric odds ratio test.	<= 3	-
RMSMappingQuality (MQ)	The root mean square mapping quality over all the reads at the site.	>= 40	-
MappingQualityRankSumTest (MQRankSum)	The u-based z-approximation from the Rank Sum Test for mapping qualities. It compares the mapping qualities of the reads supporting the reference allele and the alternate allele.	>= -12.5	-
ReadPosRankSumTest (ReadPosRankSum)	The u-based z-approximation from the Rank Sum Test for site position within reads. It compares whether the positions of the reference and alternate alleles are different within the reads.	>= -8	>= -20

Before short variant annotation begins, the VCF file obtained from variant discovery and hard-filtering is pre-processed using a two-step strategy that includes decomposition of multi-allelic variants using vt decompose v0.5 [174] and left-normalization with BCFtools norm v1.9 [175], [176] (Figure 9). VCF is a format for describing locus, since multiple variants can be in the same locus (multi-allelic variants), a single line in a VCF file can describe multiple variants. However, our ultimate goal is to build a variant-centric display platform, therefore multi-allelic variants need to be decomposed into separate lines

so that each line contains one and only one variant. Furthermore, the number of possible combinations to represent the same genomic variant is non-unique. Thus, to have a unique way of describing a variant in a given reference genome, we perform left-normalization which means shifting the starting position of a variant to the left until it is no longer possible to do so.

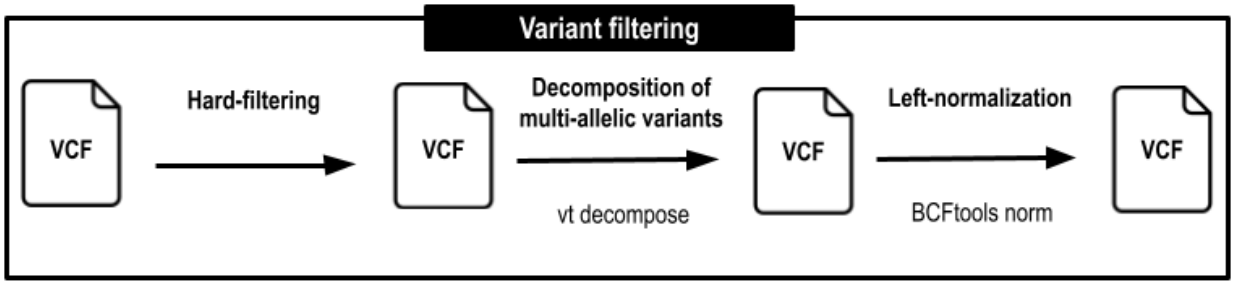

Figure 9. Variant filtering in the short variant discovery of modern DNA. Low-quality variants are removed by hard-filtering. Then, multi-allelic variants are decomposed with vt decompose [174] and left-normalization is performed with BCFtools norm [175], [176].

2.1.1.1.6 Variant annotation

The short variant annotation is performed with ANNOVAR v2019Oct24 [130] using the RefSeq transcript set [127] and 11 additional databases that are detailed in Table 4 (Figure 10). Information on the variant's distance to the closest exon-intron junction is also provided by an in-house annotation script in Python 3.7 and the RefSeq transcript set [127].

Table 4. Overview of databases for short variant annotation with ANNOVAR.

Name	Version	Description	Source	Availability
1000 Genomes Project [6]	Aug 2015	1000 Genomes Project allele frequency database	IGSR	Public
ClinVar [126]	Updated monthly on the 15th.	NCBI clinically significant variant database	NCBI	Public
Database for nonsynonymous SNPs' functional predictions (dbNSFP) [177]	3.5	Functional predictions and conservation scores database	dbNSFP	Public
Database for Single Nucleotide Variants within splicing consensus regions (dbscSNV) [178]	1.1	dbNSFP splice site variant database	dbNSFP	Public
dbSNP [124]	Build 152	NCBI SNP variant database	NCBI	Public

Exome Aggregation Consortium (ExAC) [123]	0.3	Allele frequency database	Broad Institute	Public
Exome Sequencing Project 6500 (ESP6500) [179]	6500	Exome Sequencing Project (ESP) allele frequency database	NHLBI	Public
Genome Aggregation Database (gnomAD) exomes [125]	2.1.1	Allele frequency data in exome collection	Broad Institute	Public
Genome Aggregation Database (gnomAD) genomes [125]	2.1.1	Allele frequency data in genome collection	Broad Institute	Public
genomicSuperDups [180]	14-Oct-2014	Segmental duplication Database	UCSC	Public
Online Mendelian Inheritance in Man (OMIM) database [181], [182]	Updated monthly on the 15th.	Database of human genes and genetic disorders	OMIM	License required
RefSeq [127]	96	NCBI transcript database	NCBI	Public

To further improve and facilitate the evaluation and prioritization of variants, TAPES v0.1.1 [183] is applied to assess the probability of pathogenicity of a variant and classify it into five categories following the ACMG guidelines [166], [167]: benign, likely benign, variant of unknown significance (VUS), likely pathogenic or pathogenic (Figure 10).

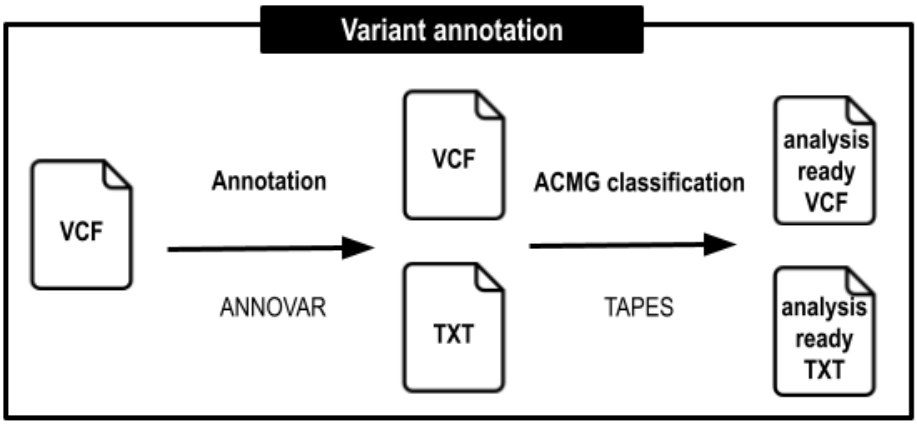

Figure 10. Variant annotation in the short variant discovery of modern DNA. The annotation of short variants is performed with ANNOVAR [130] and the classification into the pathogenic categories outlined by the ACMG is performed with TAPES [183].

2.1.1.2 Implementation of short variant discovery per parent-child trios

The short variant discovery of parent-child trios has been implemented using Python 3.7. The inputs to the pipeline are the raw sequence data for each family member, the family relationships in a PED file and, if applicable, the capture bait locations in BED format. The final SNPs and indels of the parent-child trio are output in a VCF file and table of variants in TXT format.

2.1.1.2.1 Individual calling per family member

The individual mapping and variant calling of each family member is performed following the strategy explained in Section 2.1.1.1, from the quality assessment of raw sequence data (Section 2.1.1.1.1) to the variant discovery (Section 2.1.1.1.4). The only modification to the workflow is to perform variant discovery with Sentieon Haplotyper [115] in GVCF mode instead of VCF mode. The main difference between a VCF and a GVCF is that the VCF only reports variations while the GVCF reports records for all sites, whether there is a genomic variation or not.

2.1.1.2.2 Joint calling of parent-child trios

Individual GVCF files from each family member are merged into a single multi-sample GVCF file with CombineGVCFs tool from GATK 4.0 [102] (Figure 11). Next, joint genotyping is performed in the multi-sample GVCF with the GATK 4.0 GenotypeGVCFs tool [102] and the most likely genotype combination for the parent-child trio is computed using the CalculateGenotypePosteriors tool from GATK 4.0 [102] (Figure 11).

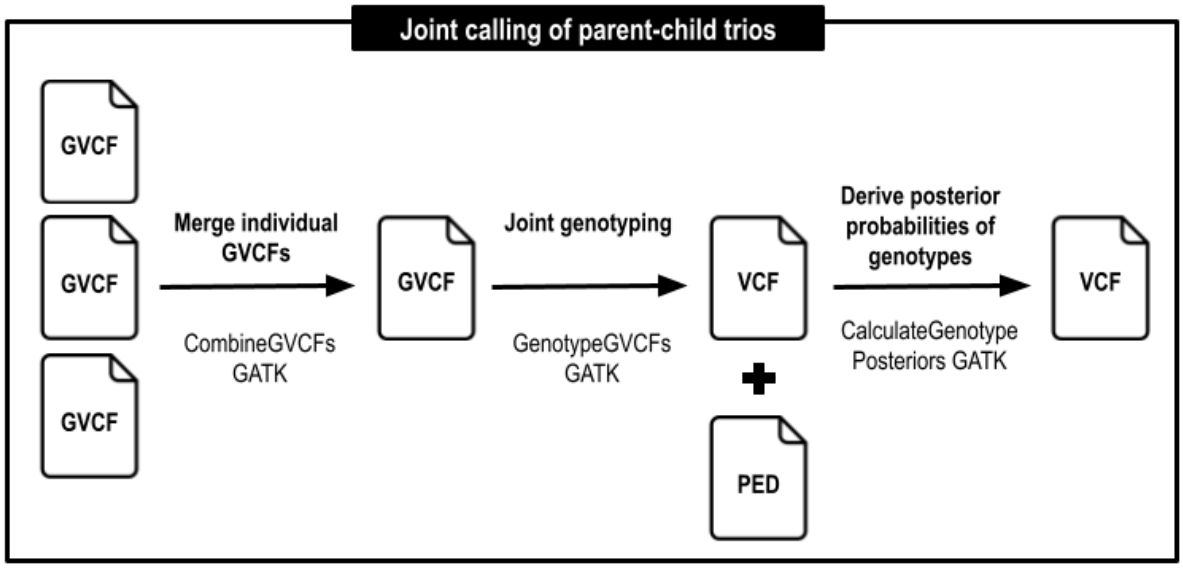

Figure 11. Joint calling of parent-child trios. The individual GVCFs containing the short germline variants of each family member are combined into a single GVCF file with GATK CombineGVCFs [102]. Then, joint genotyping is performed with GATK GenotypeGVCFs and posterior probabilities of genotypes are calculated with GATK CalculateGenotypePosteriors [102].

Variant filtering is performed as described in Section 2.1.1.1.5 but it should be noted that hard-filtering of a variant is only applied if all family members do not meet the minimum high-quality criteria (Table 3). Variant annotation follows the same workflow as for individual samples (Section 2.1.1.1.6).

2.1.2 Ancient DNA

A pipeline for the discovery of germline variants from ancient paired-end WGS data was implemented. Its input is the raw sequence data in FASTQ format and the outputs are a VCF file and a TXT file containing the identified variants.

This pipeline was applied to analyze DNA extracted from a human mandible dated between 16980-16510 calibrated years before the present. Germline variants (SNPs and indels) were reported without further interpretation due to the low coverage of the sample (0.28 x) and the degradation of ancient DNA. This analysis is part of an article published in the journal Current Biology where the doctoral student is a one of the co-authors [184].

2.1.2.1 Implementation

2.1.2.1.1 Quality assessment of raw sequence data

Adapters, low-quality bases (quality score < 15) and ambiguous nucleotides (Ns) at sequence ends are trimmed from raw paired-end reads and the overlapping sequences are merged into a single sequence by calling a consensus using AdapterRemoval tool v2.3.1 [63] (Figure 12). Partially overlapping paired-end reads are also merged into a single sequence if the overlap spanned at least 11 nucleotides, and a consensus is called on the overlapping stretch, selecting the most probable nucleotide in the case of mismatches in the overlapping region. Paired-end reads with less than 11 bases overlap are excluded from further analysis.

Only the collapsed sequences are further analyzed in an effort to exclude modern contamination under the assumption that such contamination would exhibit lower levels of fragmentation [81]. There are two different types of collapsed sequences: collapsed paired-end reads and collapsed truncated paired-end reads. Collapsed paired-end reads correspond to the sequences merged by AdapterRemoval [63] into a single sequence and are expected to represent the original template molecule. Collapsed truncated paired-end reads are like collapsed reads but were trimmed from either end of the collapsed sequence.

Sequences having a size less than 30 bp are excluded from further analysis to reduce the fraction of spurious alignments. The quality metrics of the reads before and after adapter trimming and sequence collapsing are obtained with FastQC [55] (Figure 12).

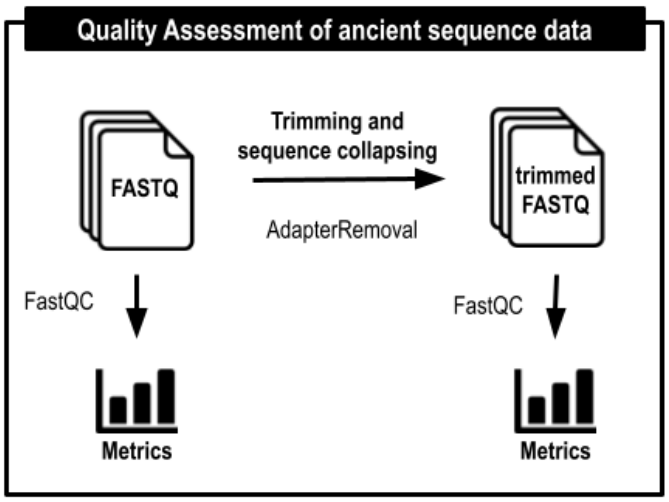

Figure 12. Quality assessment of paired-end data in the short variant discovery of ancient DNA. Adapter trimming and sequence collapsing is performed using AdapterRemoval [63]. FastQC [55] is applied on raw and trimmed reads to assess the impact of the trimming process.

2.1.2.1.2 Read alignment

Reads are aligned to the mitochondrial and nuclear genome separately. This is motivated by the presence of mitochondrial insertions in the nuclear genome (NUMTs). The presence of these sequences can hinder any attempts to call a consensus sequence for the mitochondrial data as they can result in: (i) both authentic and contaminant DNA sequences mapping the nuclear genome instead of the mitochondrial reference sequence or (ii) in a loss of sequence information since non-unique hits are generally discarded in downstream analyses.

Collapsed sequences are mapped in single end mode (as they represent the complete insert) to the nuclear human genome (GRCh37 assembly) and the mitochondrial human genome (GRCh38 assembly) using BWA aln algorithm v0.7.12 [69] with parameters that deactivate seeding and BWA samse v0.7.12 [69] (Figure 13). By default, BWA assumes that few differences will be observed between the query sequence and the reference within the first 32 bp. However, ancient DNA sequences often show an excess of nucleotide misincorporations at read termini due to the postmortem cytosine deamination. For this reason, we deactivate seeding as the mismatch expectations of a seeding approach could be too conservative and result in the loss of endogenous DNA sequences.

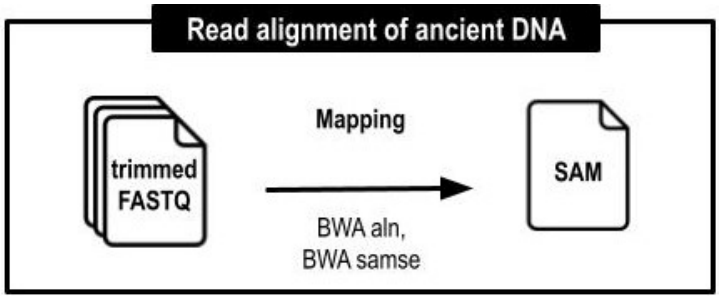

Figure 13. Read alignment in the short variant discovery of ancient DNA. Sequences are aligned to the reference genome using BWA aln [69] with seeding disabled and BWA samse [69].

2.1.2.1.3 Quality assessment of alignment data

The aligned reads are filtered based on SAM flags with SAMtools view v1.3 [73] to remove unmapped (flag 4) and QC-failed (flag 512) single reads. Then, the remaining reads are converted from SAM to BAM format with SAMtools view v1.3 [73] and sorted with SAMtools sort v1.3 [73] (Figure 14).

Different procedures are used to remove duplicates depending on the type of sequence to be analyzed: one approach for collapsed paired-end reads and another for collapsed truncated paired-end reads. Collapsed reads are filtered for duplicates by treating them as paired-end reads with the rmdup_collapsed tool from the PALEOMIX pipeline [185] (Figure 14) as both alignment termini can be considered specific features of an original template molecule. Collapsed truncated reads cannot be considered paired-end reads as they were trimmed. Thus, deduplication on collapsed truncated reads is performed in single-end mode using MarkDuplicates tool from Picard software [75] (Figure 14).

After removal of duplicates, sequences having a size higher than 60 bp are excluded from further analysis to exclude modern contamination.

Then, three main steps are followed to separate between endogenous ancient DNA and modern contaminant DNA: (i) the evaluation of the presence of ancient DNA, (ii) removal of modern contamination and (iii) microbial profiling to rule out the remaining presence of modern microbial contamination (Figure 14).

First, to determine the presence of ancient DNA in the sample of interest, patterns of nucleotide misincorporation and DNA fragmentation are plotted with mapDamage2.0 [87] looking for a postmortem DNA damage (PMD) pattern. The observation of such a pattern suggests the presence of ancient DNA but does not prove that modern contamination is absent. Modern day contamination is removed using PMDtools v0.60 [89] and then the microbial profile of the sample is evaluated with GAIA software v2.02 [186] to discard that microbial contamination remains. Base quality scores for misincorporations likely due to ancient DNA damage are recalculated with mapDamage2.0 [87] (Figure 14) to mitigate the impact of postmortem damage on downstream analyses.

Finally, local realignment around indels is performed with Sentieon Realigner v201911.01 [115] (Figure 14) both to achieve a consensus indel suitable for downstream analysis and to minimize spurious mismatches resulting from misalignments.

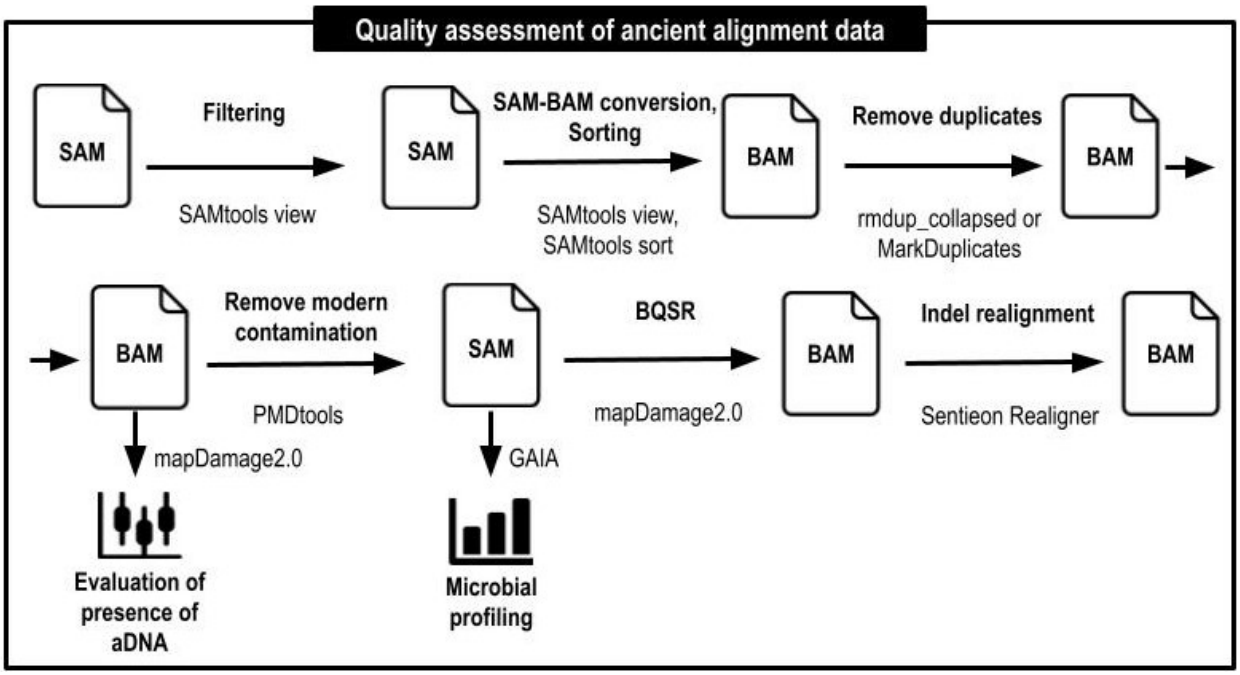

Figure 14. Quality assessment of alignment data in the short variant discovery of ancient DNA. Unmapped and QC-failed single reads are filtered with SAMtools view [73]. Then, SAM to BAM conversion is performed with SAMtools view [73] and reads are sorted with SAMtools sort [176]. Duplicates are removed using two approaches: rmdup_collapsed [185] for collapsed reads and MarkDuplicates [75] for collapsed truncated reads. The presence of ancient DNA is determined by the patterns of nucleotide misincorporation and DNA fragmentation obtained with mapDamage2.0 [87]. Modern contamination is removed with PMDtools [89] and the microbial profiling is obtained with GAIA [186]. Base Quality Score Recalibration (BQSR) is performed with mapDamage2.0 [87] and the realignment of indels with Sentieon Realigner [115].

2.1.2.1.4 Genetic sex estimation

Genetic sex is calculated using the yjasc_3752_ry_compute script from Skoglund *et al.* [187]. It is based on the estimation of the fraction of reads mapping to Y chromosome out of all reads mapping to either X or Y chromosome (R_Y). A sample is assigned as female if its confidence interval (CI) upper bound for R_Y is lower than 0.016 and as male if its R_Y CI lower bound is higher than 0.075. Only reads with a mapping quality greater than 30 are counted for genetic sex estimation.

2.1.2.1.5 Variant discovery and annotation

Short variant discovery and annotation for ancient DNA follows the same strategy as for modern DNA (Sections 2.1.1.1.4 and 2.1.1.1.6). Germline variants are called with Sentieon Haplotyper v201911.01 [115]

(Figure 15) and annotated with ANNOVAR software v2019Oct24 [130] (Figure 15) using 12 different databases (Table 4).

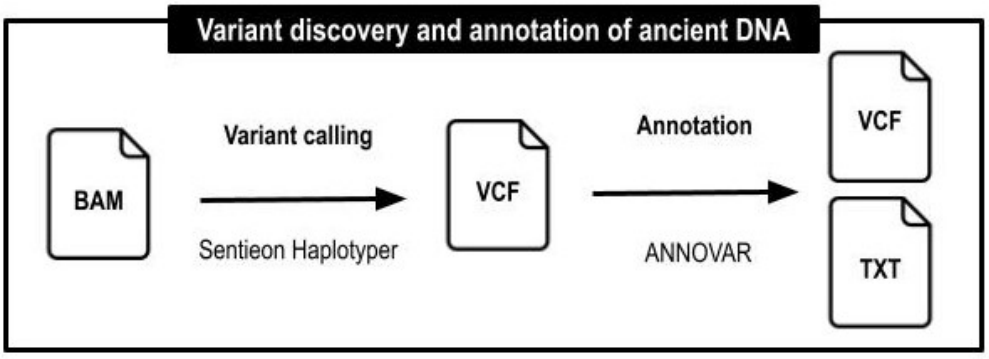

Figure 15. Variant discovery and annotation in the short variant discovery of ancient DNA. Short germline variants are called using Sentieon Haplotyper [115]. The ANNOVAR software [130] is applied for the annotation of variants.

2.1.2.2 Results

A total of 720,589,037 raw sequence reads are produced by whole genome sequencing of the ancient mandibular sample (Table 5). After the trimming and filtering procedures, 35,874,081 reads are mapped to the nuclear DNA (nDNA) and 19,780 to the mitochondrial DNA (mtDNA) (Table 5). Then, once duplication removal has been performed, 17,622,340 mapped reads in the nDNA and 18,243 in the mtDNA are retained (Table 5).

Table 5. Overview of total reads before and after trimming, mapping, and deduplication.

DNA type	Raw reads	Trimmed and filtered reads	Mapped reads			Duplicate filtered reads		
			Total	Collapsed reads	Collapsed truncated reads	Total	Collapsed reads	Collapsed truncated reads
nDNA	720,589,037	445,550,176	35,874,081	35,826,862	47,219	17,622,340	17,575,275	47,065
mtDNA			19,780	19,756	24	18,243	24	18,219

Regarding the evaluation of the presence of ancient DNA, the typical pattern for aDNA is observed in the nuclear genome (nDNA). Figure 16 provides base composition profiles within the ten first and ten last fragment positions, as well as in their respective flanking 10 bp regions in the reference genome. As can be observed, the base composition of the genomic positions immediately preceding the aDNA fragments starts is not random and is enriched in purines (A and G) as is observed in the top four plots (Figure 16). The bottom two plots of Figure 16 also show the expected pattern for aDNA, an increase in C to T and G to A mismatches when approaching the 5' and 3' termini, respectively. However, such patterns

are not observed in the mitochondrial genome (Figure 17) indicating a lack of ancient DNA and therefore excluded from further analysis.

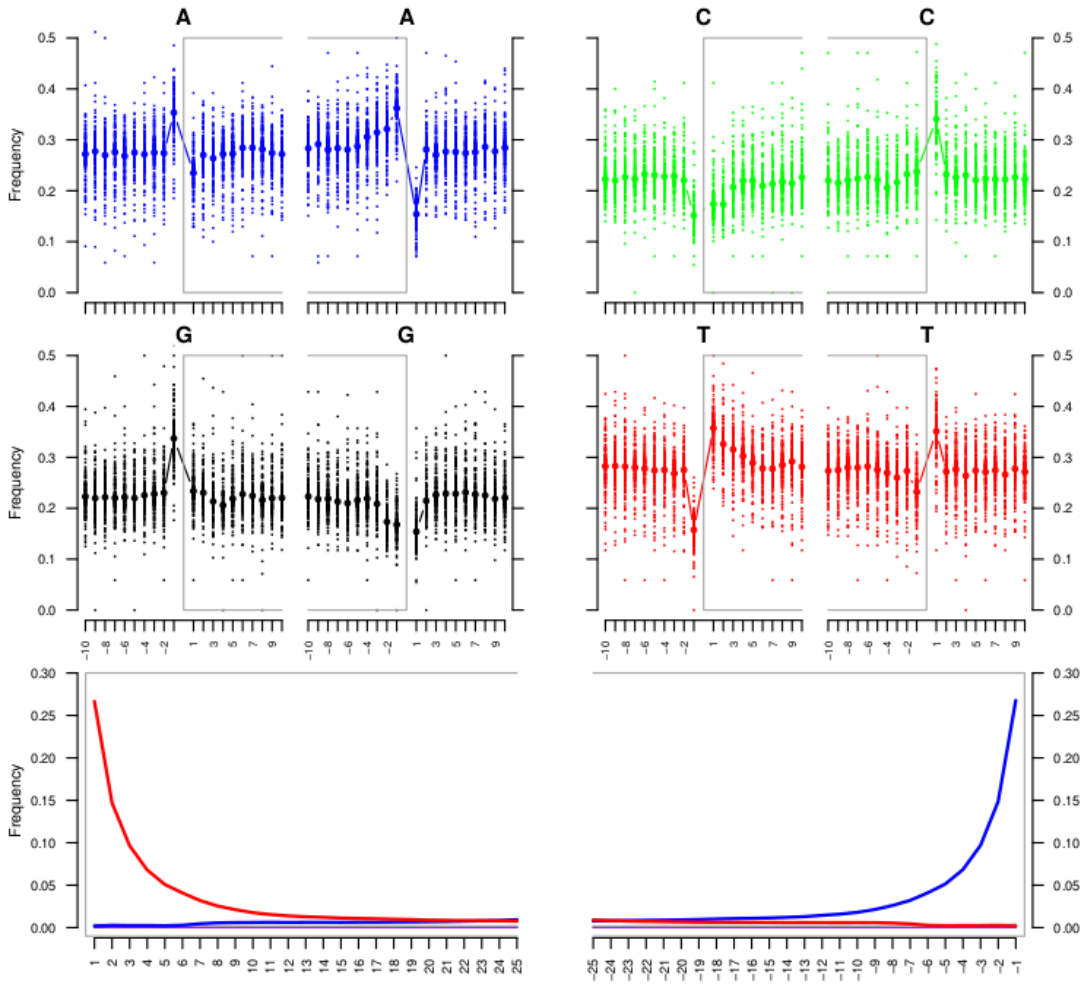

Figure 16. Fragmentation and misincorporation patterns in nDNA. The four upper mini-plots show the base frequency outside and inside the read (the open grey box corresponds to the read). The bottom plots are the positions' specific substitutions from the 5' (left) and the 3' end (right). The following color codes are used in the bottom plots: Red: C to T substitutions. Blue: G to A substitutions. Grey: All other substitutions. Orange: Soft-clipped bases. Green: Deletions relative to the reference. Purple: Insertions relative to the reference.

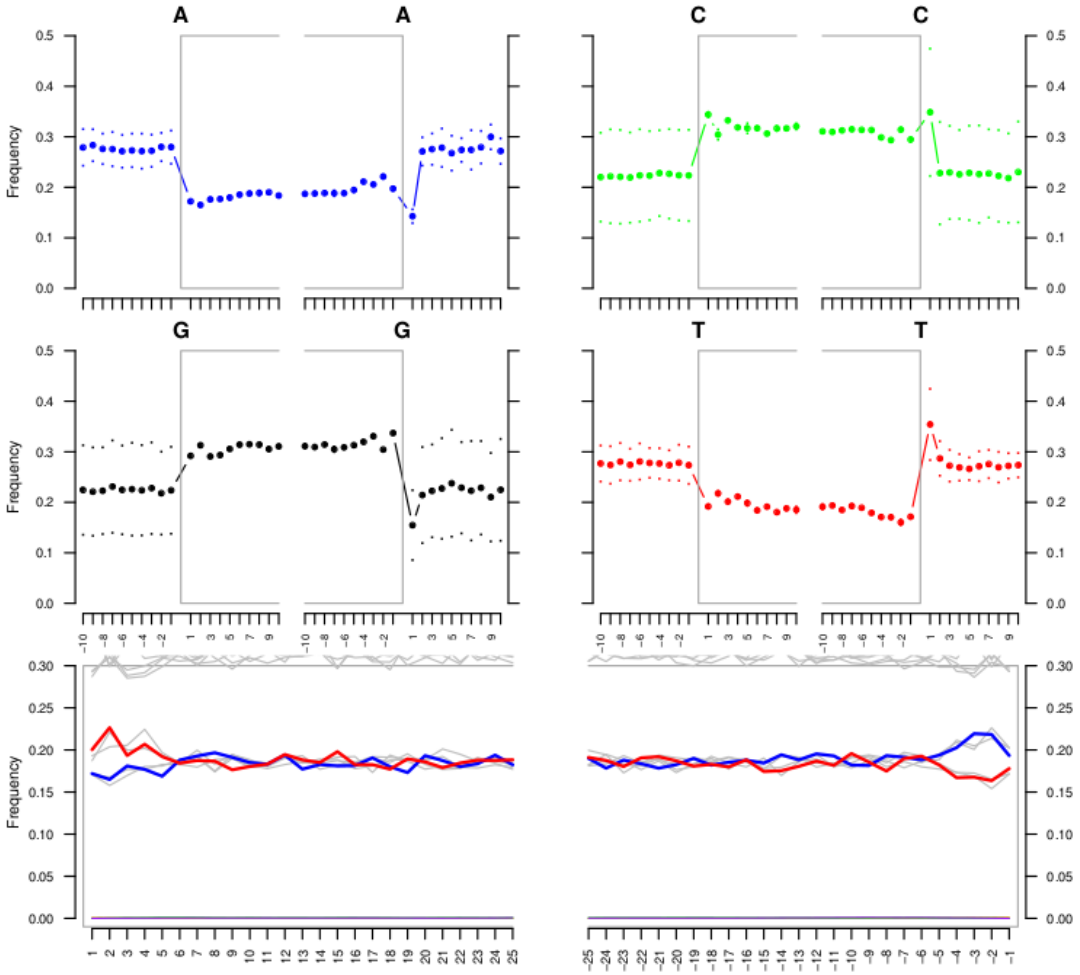

Figure 17. Fragmentation and misincorporation patterns in mtDNA. The four upper mini-plots show the base frequency outside and inside the read (the open grey box corresponds to the read). The bottom plots are the positions' specific substitutions from the 5' (left) and the 3' end (right). The following color codes are used in the bottom plots: Red: C to T substitutions. Blue: G to A substitutions. Grey: All other substitutions. Orange: Soft-clipped bases. Green: Deletions relative to the reference. Purple: Insertions relative to the reference.

After removing modern-day contamination, 3,398,029 endogenous ancient reads are obtained in the nDNA, of which only 413 are assigned to the bacteria domain. These low numbers ruled out microbial contamination in ancient DNA. Finally, genetic sex estimation and identification of variants is conducted in the endogenous nDNA. The sample is assigned as male since the CI lower bound for R_Y is 0.0826 (higher than 0.075) and a total of 1,774 germline variants are identified, of which there are 1742 SNPs and 32 indels.

2.1.2.3 Conclusion

This pipeline allows to identify endogenous ancient DNA, remove its modern contamination, assign the gender of the sample (male), and identify short germline variants. Given the low coverage of the sample

(0.28 x), the variants are reported without further interpretation, however, this pipeline could be applied to perform variant calling of ancient samples with sufficient coverage. The obtained results provided elements to the collaborator group to obtain new insights into the migration of ancient populations [184]. Specifically, the diffusion in Southern Europe of a genetic component linked to Balkan/Anatolian refugia that was previously believed to have spread during later major warming shifts was backdated by about 3000 years [184].

2.2 Structural variant discovery

In the context of this thesis, two pipelines have been implemented for germline structural variant discovery in modern DNA: (i) a pipeline for identification of copy number variants and chromosomal rearrangements using WGS data and (ii) a pipeline for *in silico* optimization of copy number variant detection from WES or TS data (isoCNV). Both pipelines require alignment data in BAM format as input, which is obtained following the same strategy applied for the discovery of modern short variants and described in Section 2.1.1.1: (i) quality assessment of raw sequence data (Section 2.1.1.1.1), (ii) alignment to the reference genome (Section 2.1.1.1.2) and (iii) quality assessment of alignment data (Section 2.1.1.1.3).

2.2.1 Whole Genome Sequencing

The structural variant discovery pipeline for WGS data has been implemented using Python 3.7. The only input required is a list with the full path to the BAM file(s) to be analyzed. The final duplications, deletions, insertions and inversions are output in a BED file and in a variant table in TXT format, while the final interchromosomal translocation (CTX) events are output in a BEDPE file.

Structural variant discovery from the alignment data is performed individually for each sample using GRIDSS [152] with default parameters (Figure 18). After review of the best performing tools for germline structural variant detection using WGS data (Section 1.3.2.1), GRIDSS [152] was the tool of choice because it is based on a combined approach and because of its high performance shown in several studies [153], [154].

The output of GRIDSS is a VCF file where SVs are described with breakend notations. The breakends are the junctions that define structural variants in the reference genome. In breakend notations, each SV has two positions in the reference genome except for inversions that have four records. To improve the interpretation of variants, we use SV types (insertion, deletion, duplication, etc.), also called simple events, instead of breakend notation. This notation conversion is performed using the simpleEvent_annotation R script provided by the GRIDSS software [152] (Figure 18).

To reduce the number of false positives, low-confidence calls are removed by following the default GRIDSS criteria (Figure 18). Thus, SVs with a low-quality score (retrieved from GRIDSS) or lacking any

supporting assemblies are filtered out. In addition, to eliminate overlap with the short variant discovery pipeline, duplications and deletions events are filtered to be at least 50 bp in length whereas insertions should be at least 30 bp. Although inversions are not called by the short variant discovery workflow, they are also filtered to meet a minimum size of 50bp. Then, the VCF file is split into two files: a BED file containing duplications, deletions, insertions and inversion and a BEDPE file containing interchromosomal events. (Figure 18). The BED file contains five columns corresponding to chromosome, start, end, SV type and sample name. The BEDPE file contains eight columns: first chromosome, start and end, second chromosome, start and end, SV type and sample name.

Finally, duplications, deletions, insertions and inversions are annotated with the AnnotSV tool v2.4 [165] while interchromosomal (CTX) events are retrieved without annotation since they are not supported by AnnotSV (Figure 18). AnnotSV provides annotations for 12 databases detailed in Table 6 and also classifies the SVs according to their pathogenicity following the ACMG guidelines [166], [167].

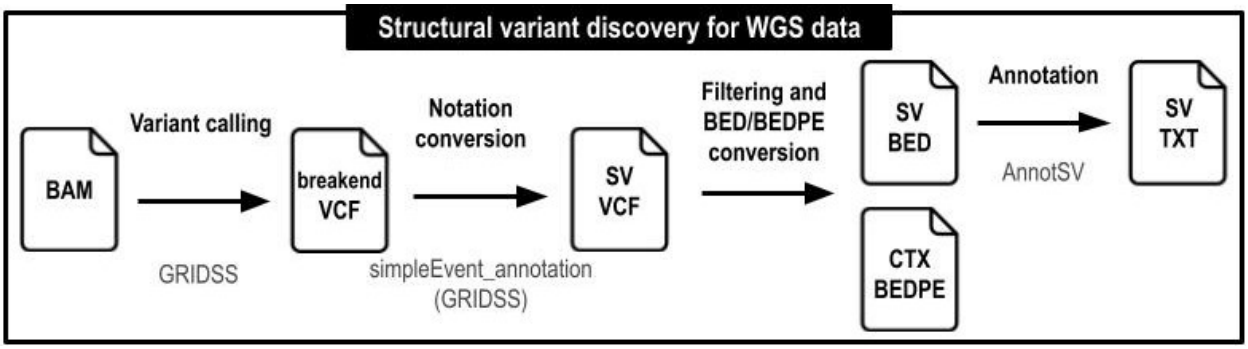

Figure 18. Structural variant discovery pipeline for WGS data. Structural variant calling is performed using GRIDSS [152]. The breakend notation in the VCF file is converted to SV types with the GRIDSS simpleEvent_annotation script. Then, low-confidence calls are removed, as well as duplications, deletions and inversions less than 50 bp in size and insertions less than 30bp. The VCF file is divided into a BED file containing duplications, deletions, insertions and inversion and a BEDPE file containing interchromosomal events (CTXs). Finally, the BED file is annotated using AnnotSV [165].

Table 6. Overview of databases for structural variant annotation with AnnotSV.

Name	Version	Description	Source	Availability
1000 Genomes Project [6]	2017	1000 Genomes Project allele frequency database	IGSR	Public
Clinical Genome Resource (ClinGen) [188]	13-Jul-2020	Haploinsufficiency and triplosensitivity scores	NCBI	Public
Database of Genomic Variants (DGV) [189]	May 2016	Gold Standard Variants	TCAG	Public
dbVar [190]	29-Jun-2020	NCBI structural variant database	NCBI	Public

Deciphering Developmental Disorders (DDD) Study [191]	13-Jul-2020	Human genome variants and phenotypes.	EMBL-EBI	Public
Exome Aggregation Consortium (ExAC) [123]	0.3	SV frequency database	Broad Institute	Public
Frequency annotations [192]	19-Dec-2019	DECIPHER frequency database	Sanger Institute	Public
Genome Aggregation Database (gnomAD) SV [193]	2.1.1	SV frequency data	Broad Institute	Public
Haploinsufficiency Predictions [194]	3	DECIPHER haploinsufficiency database	Sanger Institute	Public
Ira M. Hall's lab SV frequency [195]	31-Dec-2018	SV frequency data	Ira M. Hall	Public
Online Mendelian Inheritance in Man (OMIM) database [181], [182]	Updated monthly on the 15th.	Database of human genes and genetic disorders	OMIM	License required
RefSeq [127]	96	NCBI transcript database	NCBI	Public

2.2.2 Whole Exome Sequencing and Targeted Sequencing

A pipeline for *in silico* optimization of copy number variant detection from targeted or exome sequencing data (isoCNV) has been implemented. As mentioned in the state-of-the-art of structural variant discovery (Section 1.3.2.1), when using WES or TS data, only CNVs can be properly identified. This is due to the fact that the sensitivity of detection of other types of SVs is much lower, since only a subset of rearrangements with breakpoints in or near the capture regions can be detected and that the targeted capture method introduces inefficiencies [44]–[46].

After the bibliographic review of the CNV detection algorithms for WES or TS data (Section 1.3.2.1), DECoN [148] has been the tool of choice for this type of analysis. To maximize its sensitivity, the parameters of DECoN should be optimized for each specific dataset [157]. This parameter optimization process can be performed using an optimizer from the CNVbenchmarkeR framework [157] but requires a CNV validation set. The CNV validation set is usually generated using either multiplex ligation-dependent probe amplification (MLPA) or array comparative genomic hybridization (aCGH), which are gold standard methods [196] but are also time-consuming and expensive. For this reason, we have developed the isoCNV

pipeline, which optimizes the parameters of the DECoN algorithm using only NGS data. The parameter optimization process is performed using an *in silico* CNV validated dataset obtained from the overlapping calls of three tools: DECoN v1.0.2 with default parameters [148], CNVkit v0.9.6 [146] and panelcn.MOPS v1.12.0 [147].

The pipeline is a Python 3.7 software package comprising a command-line program, isoCNV.py. The inputs to the program are a batch of BAM files obtained under the same experimental conditions and the regions of interest (ROI) corresponding to the capture bait locations in BED format. The main outputs of the pipeline are a BED file with the unannotated CNVs and a variant table in TXT format that contains the annotated CNVs. isoCNV (*in silico* optimization of Copy Number Variant detection from targeted or exome sequencing data) is publicly available at https://gitlab.com/sequentiateampublic/isocnv.

The performance of isoCNV pipeline has been evaluated in both TS and WES real datasets and it has shown to increase the sensitivity of DECoN, which is especially critical when this tool is used as a screening step in a diagnostic strategy. An article on this work has been submitted to the BMC Bioinformatics Journal with the doctoral student as first author, and is currently under review prior to acceptance for publication.

2.2.2.1 Implementation

There are 5 main steps in the isoCNV pipeline: individual CNV calling using three different algorithms, creation of an *in silico* validation dataset, parameter optimization, CNV calling with optimized parameters and CNV annotation (Figure 19).

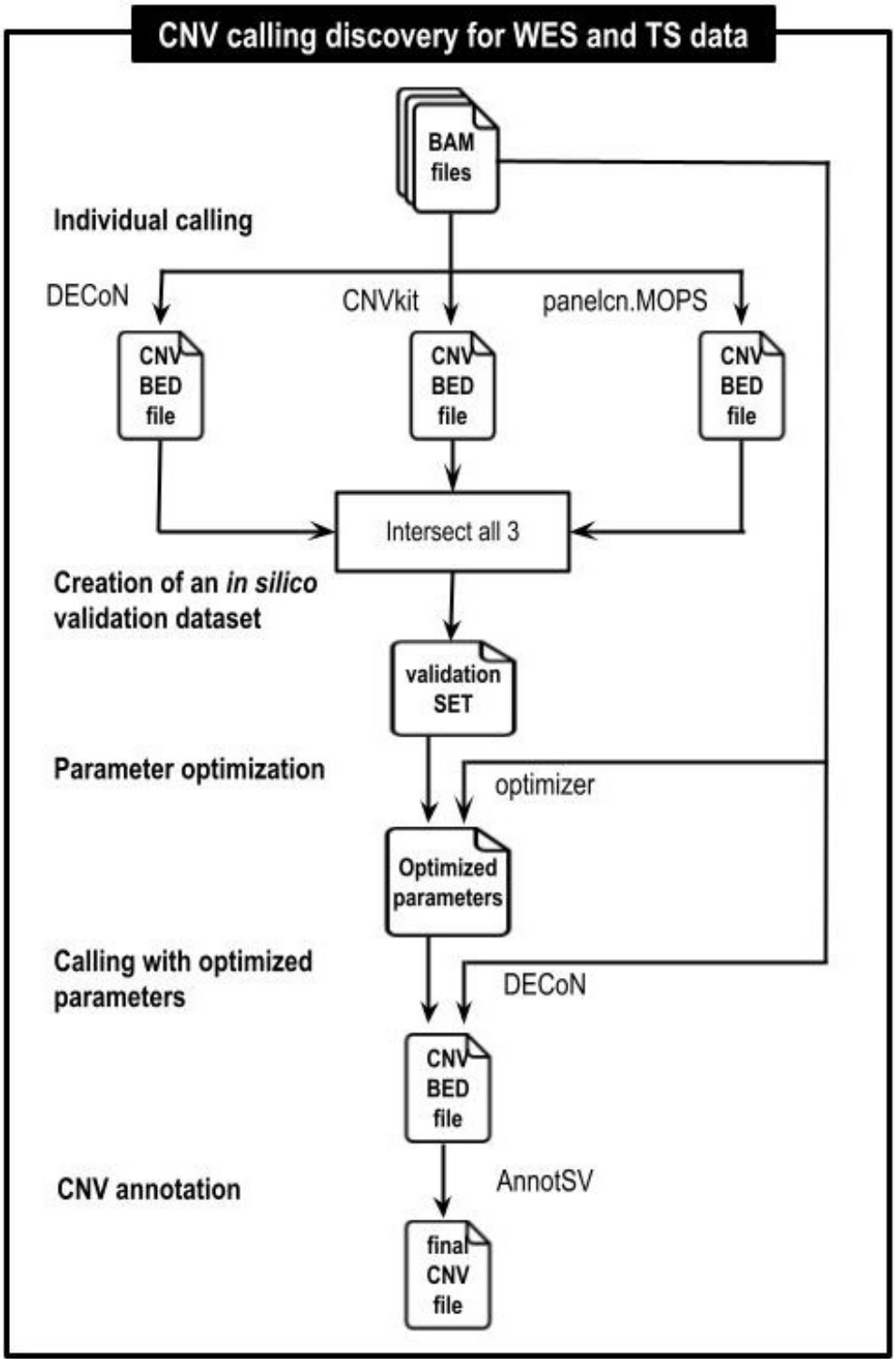

Figure 19. CNV discovery pipeline for WES and TS data. CNV calling is performed using three different tools: DECoN with default parameters [148], CNVkit [146] and panelcn.MOPS [147]. The CNV validation set is obtained from the overlapping calls of the three tools. Then, DECoN algorithm is executed using up to 22 different values for each parameter. The results obtained with each combination of parameters are compared with the validated set to obtain optimized parameters. Finally, CNV calling is performed using DECoN with optimized parameters and the final CNVs are annotated using AnnotSV [165].

2.2.2.1.1 Datasets

A targeted sequencing dataset and a whole-exome sequencing dataset have been selected to evaluate isoCNV performance: ICR96 exon CNV validation series [197] and NimbleGen set [198], respectively. Validated CNV information is available for both datasets, ICR96 has been validated by MLPA and NimbleGen by SNP microarray [199]. The FASTQ files for the ICR96 exon CNV validation series can be accessed through European-Genome phenome Archive (EGA) under accession number EGAS00001002428. The NimbleGen dataset can be downloaded from the Sequence Read Archive (SRA) [200] under accession number SRP010920.

2.2.2.1.2 Data pre-processing

The raw sequence data is pre-processed to obtain the alignment data following the same approach applied for the discovery of modern short variants (Section 2.1.1.1). First, trimming the sequence data with Trimmomatic [58] (Section 2.1.1.1.1). Secondly, alignment to the reference genome and sorting with Sentieon v202010.02 [115] (Section 2.1.1.1.2) and finally the evaluation of the quality of the alignment data (Section 2.1.1.1.3).

2.2.2.1.3 Individual CNV calling

Preliminary identification of CNVs is performed using three different algorithms: DECoN v1.0.2 [148] with default parameters, CNVkit v0.9.6 [146] and panelcn.MOPS v1.12.0 [147]. All three are based on the Read-depth (RD) approach (Section 1.3.2.1), so the gender of the samples is a critical factor in identifying copy number variants on sex chromosomes. The user can provide the gender of the samples as input or it will be inferred automatically using the CNVkit gender tool [146].

DECoN algorithm is applied using default parameters but with some modifications. DECoN creates a reference set for each sample of interest consisting only of those samples which are well correlated [148]. Our first implementation has been to add a list of related samples as an optional input to automatically exclude them from the reference set of their relatives, otherwise the common CNVs in the family would not be identified. Furthermore, DECoN has been modified to accept only a maximum of 10 samples in the reference set since it has been shown that its optimal size is between 5 and 10 samples [201]. By default, CNV calling is performed separately between male and female samples to allow detection of CNVs on the sex chromosomes. However, if there are less than 5 female or male samples in the batch, all samples are analyzed together. Optionally, only sex chromosomes can be analyzed separately between male and female samples, using the “batch2” option of isoCNV. Additionally, two optional filters have been added to the pipeline so that the user can easily select whether or not to apply them: filter by regions of interest (ROIs) or by sample. ROIs are removed if they are below the default minimum median coverage threshold (100) for any sample (measured across all ROI in the target) or region (measured across all samples). CNVs are

filtered out from samples that do not meet either the minimum coverage threshold (100) or the minimum correlation threshold (0.98). Samples which do not have a high correlation with other samples in the set are likely to have suboptimal detection across the entire target.

The default parameters also apply for calling CNVs with the CNVkit algorithm, except for the filtering process where the “cn” method is applied instead of “ci”. Here, the reference set consists of all female samples in the batch with a standard deviation (SD) between -2 and 2. However, if the sample of interest is female, it will be excluded from the reference. Two exceptions should be taken into account in the creation of the reference set: (i) if there are less than 5 female samples, then the males are used as reference and (ii) if there are less than 5 females and less than 5 males, all samples are used as reference so calls on the Y chromosome cannot be trusted. Additionally, the CNVkit thresholds for defining copy numbers 0 and 1 have been modified to be more restrictive: for CN0 the threshold range ($\log_2$ value up to) has changed from $\log_2 \leqq -1.1$ to $\log_2 \leqq -2$ and for CN1 from $-1.1 < \log_2 \leqq -0.4$ to $-2 < \log_2 \leqq -0.4$. Finally, the precise copy number values obtained by CNVkit (0, 1, 2, 3, etc.) are converted to deletion (DEL) or duplication (DUP) states taking into account the gender of the sample of interest and the gender of the references.

The identification of CNVs with panelcn.MOPS is also carried out using the default parameters of the tool. As with the DECoN algorithm, the analysis is performed separately between male and female samples, unless there are less than 5 females or males that all samples are analyzed together. Here, ROIs are removed if they are marked as "low quality" by panelcn.MOPS: their median read count across all samples does not meet the minimum default threshold (30) or if their read count shows a high variation across all samples as marked by the default behaviour of the algorithm.

2.2.2.1.4 In silico validation dataset

The output of each algorithm (DECoN with default parameters, CNVkit and panelcn.MOPS) is normalized to a single format, a tab-delimited BED file. This BED file has the same structure as the BED file for SV identification with WGS data (Section 2.2.1), contains five columns corresponding to chromosome, start, end, CNV type (DEL or DUP) and sample name. Then, using BEDTools utilities v2.29.2 [159] and pybedtools Python library v0.8.1 [202], the overlapping CNVs between call sets from the three algorithms are selected if meet two criteria (i) at least 60% of overlap with one of the call sets from the algorithms and (ii) a minimum size equivalent to the mean size of the target ROIs. If one of the tools reports no CNV in any sample, only the output of the other two algorithms is used to create the *in silico* validation set.

2.2.2.1.5 Parameter optimization

The parameter optimization process is implemented using the feature optimizer from the CNVbenchmarkeR framework [157]. This framework runs the DECoN algorithm against a validated dataset using up to 22 different values for each parameter. Then, it compares the results obtained from each combination of parameters with the validated copy number states to obtain the optimized parameters for the dataset.

The validated copy number states correspond to those obtained *in silico* from the overlapped calls between DECoN, CNVkit and panelcn.MOPS. However, validated information about normal copy number states is also necessary. To obtain this data, we select as validated regions those where a CNV has been found (and has been validated *in silico*) in any of the samples, and then, for each validated region, if validated CNV has not been found, we assign it a normal copy number state.

The DECoN parameters subject to optimization are the following: (i) the minimum correlation threshold between the sample of interest and any other sample to be considered well correlated (0.98), (ii) the minimum median coverage for any sample or ROI to be considered well-covered (100) and (iii) the transition probability between normal copy number state and either deletion or duplication state in the hidden Markov model (0.01).

2.2.2.1.6 CNV calling with optimized parameters

Final copy number variants are identified using our modified version of the DECoN algorithm (Section 2.2.2.1.3) with the optimized parameters obtained in the previous step instead of the default ones. Results are normalized in BED format with the following columns: chromosome, start, end, CNV type (DEL or DUP), sample name, reads ratio and the precise copy number value. Reads ratios are calculated by DECoN algorithm and copy number values are calculated based on reads ratio (Table 7). The calculation of the precise copy number values is based on the CNVkit threshold method [146] but with some modifications: log2 values are converted to absolute scale and then, adjusted empirically (Table 7).

Table 7. The thresholds map to integer copy number in CNVkit and in isoCNV. The $\log_2$ ratio thresholds assigned to copy number values in CNVkit [146] are converted to absolute scale and next, empirically adjusted to obtain the reads ratio thresholds of isoCNV.

Copy Number Value	Threshold Range		
	CNVkit ($\log_2$)	CNVkit (Absolute scale)	isoCNV (Reads Ratio)
0	$\log_2 \leqq -1.1$	Absolute scale $\leqq 0.4665$	Reads Ratio $\leqq 0.1$
1	$-1.1 < \log_2 \leqq -0.4$	$0.4665 <$ Absolute scale $\leqq 0.7579$	$0.1 <$ Reads Ratio $\leqq 0.8$
2	$-0.4 < \log_2 \leqq 0.3$	$0.7579 <$ Absolute scale $\leqq 1.2311$	$0.8 <$ Reads Ratio $\leqq 1.2$
3	$0.3 < \log_2 \leqq 0.7$	$1.2311 <$ Absolute scale $\leqq 1.6245$	$1.2 <$ Reads Ratio $\leqq 1.8$
4	-	-	$1.8 <$ Reads Ratio $\leqq 2.2$

Reads Ratio * 2	-	-	2.2 < Reads Ratio

2.2.2.1.7 CNV annotation

To facilitate prioritization of copy number variants of interest, CNVs are annotated using the AnnotSV tool [165] and the same 12 databases described in Table 6 for SV discovery from WGS data (Section 2.2.1).

2.2.2.1.8 Benchmark evaluation metrics

The performance of isoCNV is evaluated per region of interest (ROIs). The ROIs correspond to the regions in the target BED file of the dataset and are treated as independent entities. In addition, the evaluation takes into account the normal copy number states (no calls), since in a real diagnostic scenario, all no-call regions must be confirmed using an orthogonal method.

If the tool matches the result of the validation information it is classified as true positive (TP) or true negative (TN). If the tool identifies a CNV not present in the validation information, we consider it a false positive (FP) and if the tool misses a validated CNV it is a false negative (FN).

2.2.2.2 Results

2.2.2.2.1 In silico validation dataset

The total copy number variants per ROI identified by each algorithm (DECoN [148], CNVkit [146] and panelcn.MOPS [147]) are shown in a Venn diagram for each dataset (Figure 20). In both datasets, the total number of CNVs per ROI varies by algorithm, panelcn.MOPS identified the highest number of CNVs while DECoN identified the lowest number (Figure 20). The overlapping CNVs per ROI between the three call sets were 205 in the TS dataset (ICR96) and 693 in the WES dataset (NimbleGen) (Figure 20). From these, the validation dataset was composed from the CNVs that overlapped at least 60% with one of the call sets from the algorithms and that had a minimum size equivalent to the mean size of the target ROIs. Hence, 72 validated CNVs were obtained in ICR96 and 388 in NimbleGen. Regions with normal copy number state were also included in the validation set.

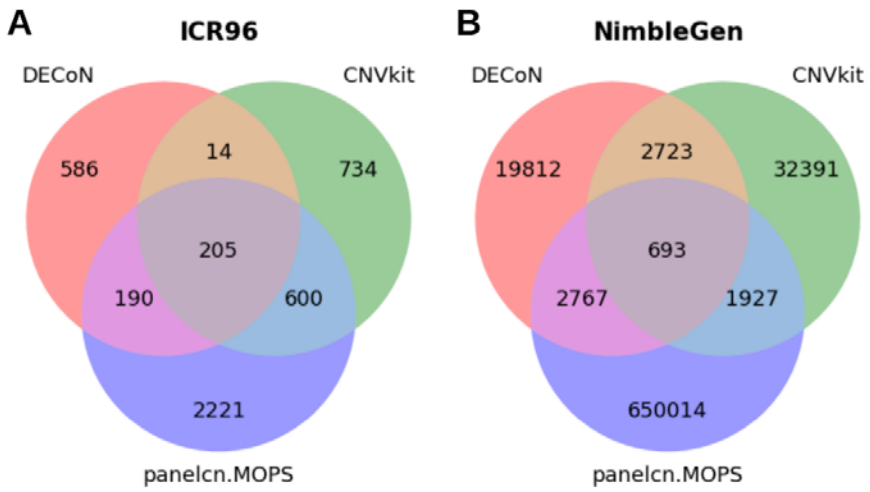

Figure 20. Number of CNVs per ROI detected by three callers. A. Venn Diagram of the CNVs in ICR96 dataset. B. Venn Diagram of the CNVs in NimbleGen dataset.

The copy number states in the validation set were compared with the real copy number information obtained by MLPA in ICR96 and by SNP microarray in NimbleGen set (Table 8). The *in silico* validation set of both datasets showed a specificity of 1 since no FPs were identified, while its sensitivity was quite low as a high number of FNs was found (Table 8). These results were expected, since the filters applied to define a copy number as validated are quite restrictive.

Table 8. Benchmark results for the individual callings and the *in silico* validation dataset.

Dataset	Method	TP	TN	FP	FN	Total	Sensitivity	Specificity	PPV	NPV	F-score
ICR96	DECoN	247	27330	60	49	27686	0.8345	0.9978	0.8046	0.9982	0.8192
	CNVkit	205	27225	91	165	27686	0.5541	0.9967	0.6827	0.9940	0.6156
	panelcn.MOPS	278	27061	18	329	27686	0.4580	0.9993	0.9392	0.9880	0.6157
	In silico validation	58	27390	0	238	27686	0.1959	1	1	0.9914	0.3277
NimbleGen	DECoN	220	7236	43	1138	8637	0.1620	0.9941	0.8365	0.8641	0.2714
	CNVkit	777	7274	582	4	8637	0.9949	0.9259	0.5717	0.9995	0.7262
	panelcn.MOPS	736	6891	619	391	8637	0.6531	0.9176	0.5432	0.9463	0.5931
	In silico validation	30	7278	0	1329	8637	0.0220	1	1	0.8456	0.0432

2.2.2.2.2 Benchmark evaluation

The CNV detection using DECoN with optimized parameters allowed the identification of 597 CNVs in ICR96 and 125,601 in NimbleGen. The parameter optimization process led to an increase in sensitivity and F-score for both datasets, but especially for NimbleGen. In the NimbleGen set, there was an increase in sensitivity from 16.2% to 84.5% and in F-score from 27.1% to 82.7% with a slight decrease in specificity from 99.4% to 96.3% (Figure 21 and Table 9). In both datasets, the Negative Predictive Value (NPV) was higher than the Positive Predictive Value (PPV) before and after optimization process (Figure 21 and Table 9) as expected in unbalanced datasets with a much larger number of negative elements (no calls) than positive ones.

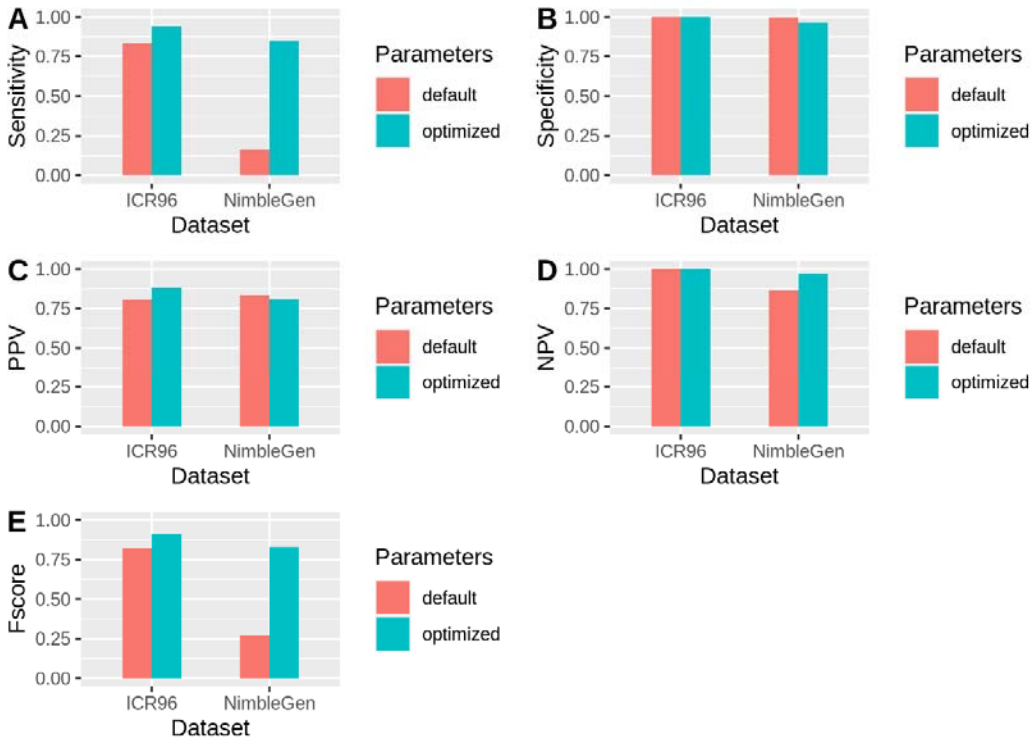

Figure 21. Benchmark results with default and optimized parameters. Shows sensitivity, specificity, PPV, NPV and F-score when executing DECoN with the optimized parameters in comparison to the default parameters.

Table 9. Benchmark results with default and optimized parameters.

Dataset	Parameters	TP	TN	FP	FN	Total	Sensitivity	Specificity	PPV	NPV	F-score
ICR96	Default	247	27330	60	49	27686	0.8345	0.9978	0.8046	0.9982	0.8192
	Optimized	279	27354	36	17	27686	0.9426	0.9987	0.8857	0.9994	0.9133
NimbleGen	Default	220	7236	43	1138	8637	0.1620	0.9941	0.8365	0.8641	0.2714
	Optimized	1147	7009	271	210	8637	0.8452	0.9628	0.8089	0.9709	0.8267

To evaluate if parameter optimization of DECoN allows to identify new CNVs only detected by the other two algorithms (CNVkit and panelcn.MOPS) when default parameters are used, the unique CNVs of CNVkit (identified by CNVkit but not by DECoN with default parameters) were obtained and compared to the final CNVs (identified by DECoN with optimized parameters). Within the final CNVs, it was found a total of 86 CNVs in ICR96 and 2,727 CNVs in NimbleGen that were identified by CNVkit but not initially by DECoN with default parameters. The same approach was applied to the unique CNVs of panelcn.MOPS and 88 CNVs were found in ICR96 and 68,569 CNVs in NimbleGen within the final CNVs that were not identified initially by DECoN with default parameters.

In addition, the performance of isoCNV was evaluated depending on the number of samples analyzed. This relates to the reference set as samples with a better correlation or a higher coverage may be included and could improve the performance of DECoN. The ICR96 set reached almost 100% specificity and NPV independently of the number of samples with both default and optimized parameters (Figure 22). An improvement in PPV and F-score can be observed in the ICR96 set when at least 20 samples were analyzed together and then, from 24 samples, both PPV and F-score remained fairly constant, being always higher when executing DECoN with optimized parameters (Figure 22). The sensitivity in the ICR96 set also remained quite constant and above 80% when at least 6 samples were analyzed with optimized parameters, whereas there was a decrease in the sensitivity when more than 86 samples were analyzed with default parameters (Figure 22). The NimbleGen set showed a fairly constant sensitivity, specificity, PPV, NPV and F-score with optimized parameters (Figure 23). However, sensitivity, F-score and NPV decreased considerably when analyzing more than 20 samples using default parameters (Figure 23).

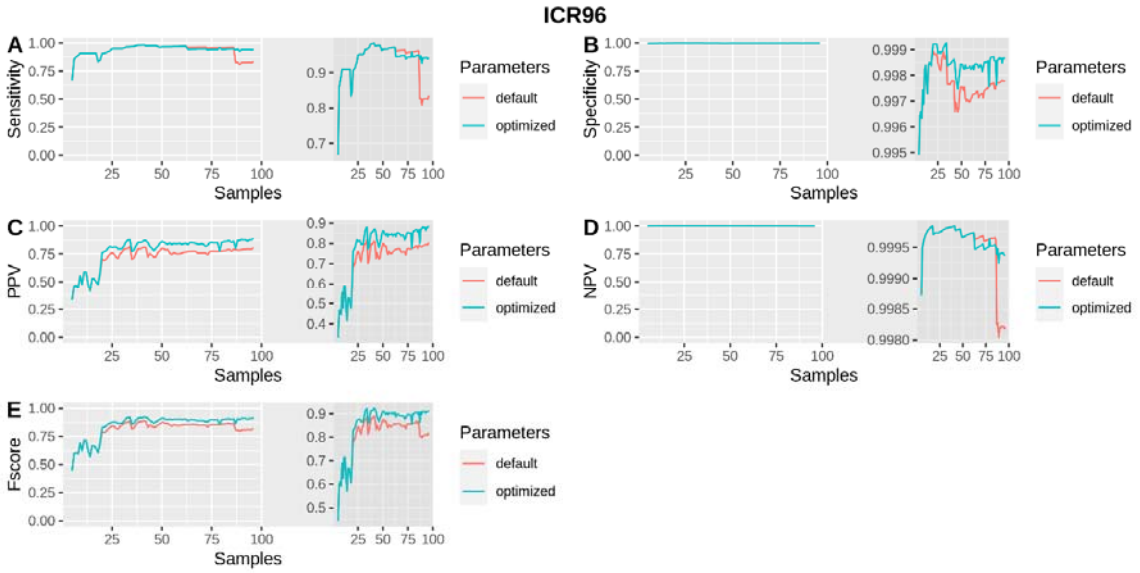

Figure 22. Benchmark results with default and optimized parameters when analyzing different numbers of samples in ICR96. Shows sensitivity, specificity, PPV, NPV and F-score when executing DECoN for different numbers of samples (from 5 to 96) with the optimized parameters in comparison to the default parameters.

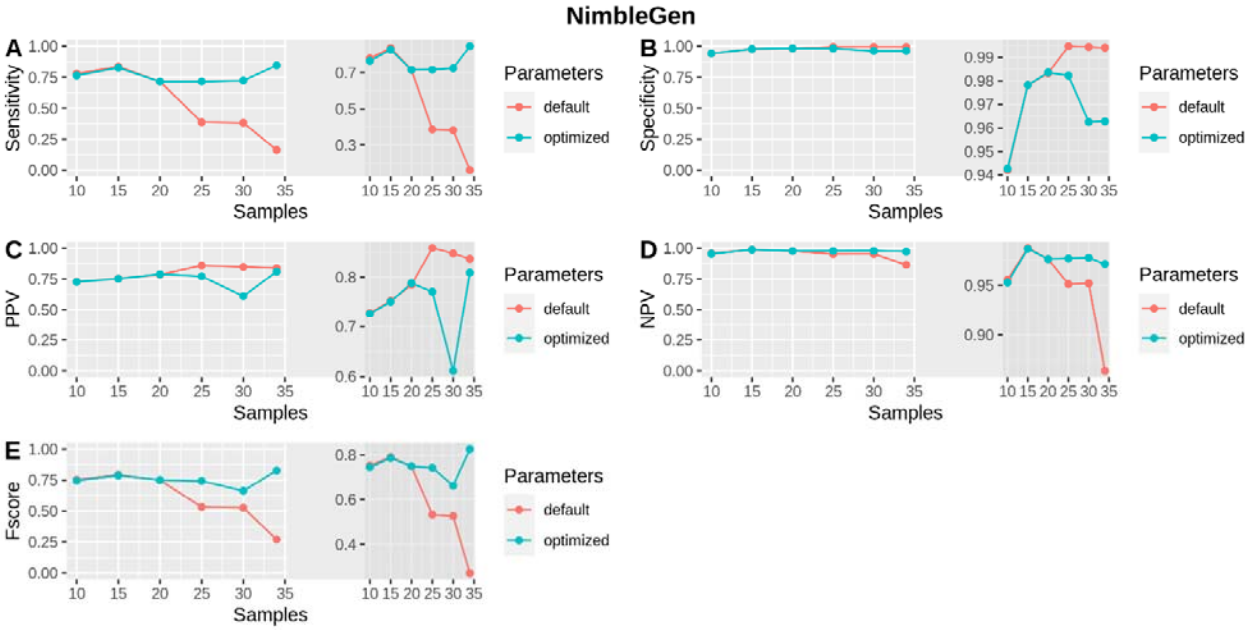

Figure 23. Benchmark results with default and optimized parameters when analyzing different numbers of samples in NimbleGen. Shows sensitivity, specificity, PPV, NPV and F-score when executing DECoN with the optimized parameters in comparison to the default parameters for different numbers of samples adding 5 each time.

To assess the effect of the number of target regions being analyzed, the isoCNV pipeline was tested by identifying CNVs by chromosome and by chromosome subset rather than analyzing all genomics regions in the target BED file at once, which is the default approach of isoCNV.

When applying isoCNV per chromosome, there was a decrease in isoCNV performance as well as DECoN performance. Regarding isoCNV performance, the parameter optimization process had no effect on the NimbleGen set, whose metrics remained the same before and after optimization (Table 10). In ICR96, optimization of parameters by chromosome also had practically no effect; there was only a small negative effect as two fewer TPs were identified after optimization (from 194 TPs to 192) (Table 10). Concerning the effect of the analysis by chromosome in DECoN, if we compare these results with those obtained when all genomics regions were analyzed at the same time (Table 9), we obtained worse results in the two datasets both before and after the optimization process (Table 10). There was a decrease in F-score from 91.3% when optimizing all genomics regions at the same time (Table 9) to 73.4% when optimizing by chromosome (Table 10) in ICR96 and from 82.7% (Table 9) to 76.7% (Table 10) in NimbleGen.

To apply isoCNV per chromosome subset, the total chromosomes found in the target BED file were split into four subsets. Here, parameter optimization by chromosome subset had a negative effect on both datasets. In ICR96, there was a decrease in sensitivity from 87.2% to 65.6% and in F-score from 81.4% to 72.4% (Table 10). In NimbleGen, there was a decrease in sensitivity from 73.0% to 66.3% and in F-score

from 80.1% to 78.6% (Table 10). These results were worse in all cases (before and after optimization) and in all datasets (ICR96 and NimbleGen) than when all genomics regions were analyzed at once. In ICR96 there was a decrease in F-score from 91.3% when parameter optimization was performed using all regions at the same time (Table 9) to 72.4% when parameter optimization was performed by chromosome subset and in NimbleGen (Table 10), from 82.7% (Table 9) to 78.6% (Table 10).

Table 10. Benchmark results for the isoCNV pipeline by chromosome and by chromosome subset.

Method	Dataset	Parameters	TP	TN	FP	FN	Total	Sensitivity	Specificity	PPV	NPV	F-score
By chr	ICR96	Default	194	27355	35	102	27686	0.6554	0.9987	0.8472	0.9963	0.7390
		Optimized	192	27355	35	104	27686	0.6486	0.9987	0.8458	0.9962	0.7342
	NimbleGen	Default	897	7196	82	462	8637	0.6600	0.9887	0.9162	0.9396	0.7673
		Optimized	897	7196	82	462	8637	0.6600	0.9887	0.9162	0.9396	0.7673
By chr subset	ICR96	Default	258	27310	80	38	27686	0.8716	0.9971	0.7633	0.9986	0.8139
		Optimized	194	27344	46	102	27686	0.6554	0.9983	0.8083	0.9963	0.7239
	NimbleGen	Default	991	7153	127	366	8637	0.7303	0.9826	0.8864	0.9512	0.8008
		Optimized	901	7246	32	458	8637	0.6630	0.9956	0.9657	0.9406	0.7862

Although the analysis by chromosome and by chromosome subset yielded suboptimal results, a new comparison was performed between the results obtained when all genomics regions were analyzed at once and when the analysis was performed by chromosome, but now taking into account only the sex chromosomes (Table 11). This comparison was conducted because the developers of ExomeDepth [201], the tool on which DECoN is based, recommend processing sex chromosomes separately rather than processing all chromosomes together but separating male and female samples.

When evaluating results on sex chromosomes in NimbleGen set, the optimization process negatively affected the performance of the algorithm regardless of the approach followed (all together or per chromosome). There was a decrease in F-score from 70.4% to 47.8% when all regions were analyzed together and a decrease in F-score from 58.5% to 50.7% with the analysis by chromosome (Table 11). Before parameter optimization, all metrics (sensitivity, specificity, PPV, NPV and F-score) were higher when all genomics regions were analyzed together than by chromosome (Table 11). However, after the parameter optimization process, all metrics except specificity were higher with the analysis by chromosome (Table 11). ICR96 was not evaluated as there were no validated CNVs available on the sex chromosomes.

In addition to the ICR96 dataset containing no validated CNVs on sex chromosomes, the NimbleGen dataset only contains 24. Given this low number of validated CNVs available on sex

chromosomes, the results obtained in this evaluation are not conclusive. Thus, the default approach for isoCNV was based on analyzing all chromosomes together but separating between male and female samples ("batch" option). In any case, an optional approach was added to the pipeline to process sex chromosomes separately ("batch2" option).

Table 11. Benchmark results for sex chromosomes in NimbleGen using "batch" and "batch2" option of isoCNV.

Method	Parameters	TP	TN	FP	FN	Total	Sensitivity	Specificity	PPV	NPV	F-score
All	Default	19	140	11	5	175	0.7917	0.9272	0.6333	0.9655	0.7037
	Optimized	16	124	27	8	175	0.6667	0.8212	0.3721	0.9394	0.4776
By chr	Default	19	129	21	6	175	0.7600	0.8600	0.4750	0.9556	0.5846
	Optimized	19	119	31	6	175	0.7600	0.7933	0.3800	0.9520	0.5067

2.2.2.3 Conclusion

We have implemented isoCNV, an easy-to-use pipeline to optimize DECoN algorithm using only NGS data. This pipeline can reduce the number of assays required per patient to reach a diagnosis, since orthogonal methods, such as MLPA or aCGH, are not required. We evaluated the performance of our tool and showed that it increases the sensitivity of DECoN in both TS and WES real datasets.

Chapter 3 | Data integration

Next-generation sequencing has greatly facilitated the identification of genomic variation. Among all NGS strategies, WGS is the best option to identify any types of genomic variants in the entire genome. WES covers only exonic regions within a genome and introduces biases due to hybridization and a non-uniform read depth distribution [44], [46]. However, WES is the preferred method for both research and clinical use [5], [42], [43] due to its lower price and faster data analysis. Here, we propose to complement WES with RNA sequencing (RNA-seq) to enhance the discovery of short variants.

RNA sequencing is the process of determining the order of nucleotides from a given RNA chain. It usually involves converting the RNA to be sequenced into cDNA fragments which are then analyzed by NGS or it can be sequenced directly without previous conversion (Nanopore technology). RNA-seq allows us to investigate and discover the gene expression patterns encoded within our RNA. Hence, RNA-seq would be able to detect variants within the expressed regions of the genome [203], [204].

The application of RNA-seq provides some advantages in variant discovery over WES. RNA-seq can identify new variations in highly expressed genes. Such highly expressed genes have a greater coverage in RNA-seq than in WES and, therefore, a greater statistical confidence to detect genomic variants. In addition, RNA-seq also allows us to identify variations in genes outside the target regions of the WES analysis and it can be more cost effective than WES as it bypasses the need for exome enrichment steps. However, the application of RNA-seq is not without its shortcomings and limitations. The transcriptome is specific to both tissue and cell type, therefore a transcriptome derived from a tissue type will not represent the entire exome. Furthermore, two significant considerations need to be accounted for: the inability of RNA-seq to detect variants in non-transcribed or low-expressed genes and its increased susceptibility to false positives calls due to errors during RNA to cDNA conversion, mapping mismatches, alternative splicing or gene fusion [203], [205].

Discovery of genomic variants using both RNA-seq and WES data increases the target regions where variants can be called and provides an orthogonal method to validate variations by complementing WES analysis with RNA-seq. In heterozygous variants, RNA-seq data can also allow us to identify the preferential expression of a parental allele, also called allele-specific expression (ASE). ASE can lead to heterozygous sites in WES-based calling being called homozygous in RNA-seq calling and lead to substantial error in monoallelic genes. In addition, RNA-seq data can be used to detect variants arising from RNA editing. RNA editing occurs after DNA transcription and synthesis by the RNA polymerase enzyme, so these changes cannot be detected by WES data. Identification of RNA editing events is important as they have been implicated in several disorders including cancer [206]–[208] and neurodegenerative diseases [209]. Moreover, the availability of RNA-seq data allows for additional analyses such as measuring

transcript expression levels or detecting novel fusion genes. These further analyses will allow for a more complete picture of the organisms of interest, supporting a better understanding of biological systems and, eventually, the development of successful precision medicine. All of this makes it more cost effective to use RNA-seq and WES data than just WGS data.

3.1 RNA-seq and WES integrated analysis for short variant discovery

Currently, there are tools to identify short variants using only RNA-seq data such as RVBoost [210], SNPiR [203], eSNV-Detect [211] or GATK Haplotypecaller [76]. Furthermore, there are tools to identify short somatic variants using both RNA-seq and WES data like RADIA [171] or VaDiR [172]. However, to our knowledge, there is no tool to identify short germline variants using both RNA-seq and WES data. For this reason, we have developed a Python 3.7 software package comprising a command-line program, varRED.py, to identify short germline variation from WES and RNA-seq data. varRED (variant discovery from RNA and Exome Data) is available at https://gitlab.com/sequentiateampublic/varred.

3.1.1 Implementation

varRED is a modular program that allows for running the complete analysis in batch or step-by-step. The inputs to the program are the RAW sequence data in FASTQ format from the WES and RNA-seq analysis and the capture bait locations used in the WES analysis. The pipeline consists of 3 main steps: WES calling, RNA-seq calling and joint variant calling. Joint variant calling includes genotyping, filtering and classification of short germline variations (Figure 24).

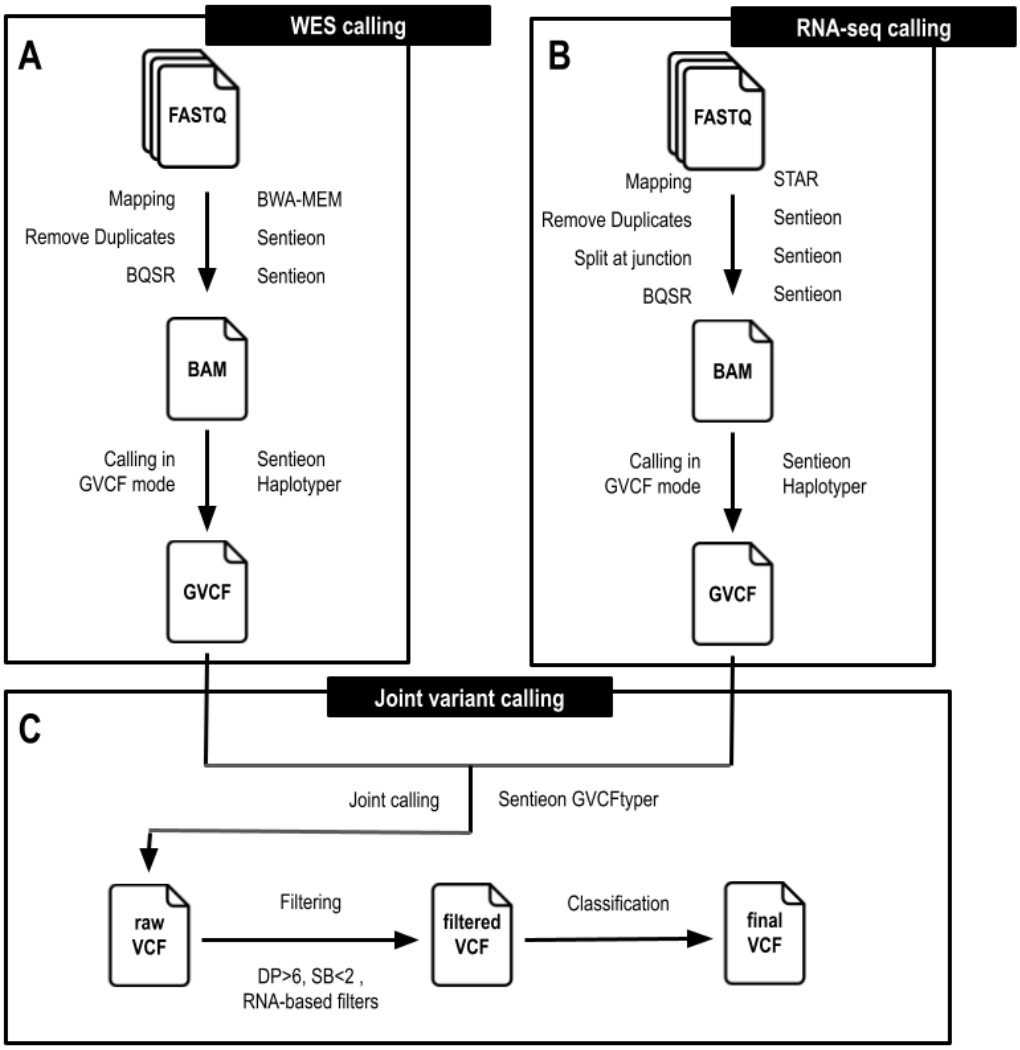

Figure 24. Overview of varRED workflow. The varRED workflow mainly consists of three steps: A) WES mapping and calling in GVCF mode. B) RNA-seq mapping and calling in GVCF mode. C) Joint calling genotyping of WES and RNA-seq data, filtering and classification of variants.

3.1.1.1 Datasets

Four samples were selected to evaluate the performance of varRED: NA12878 [or SAME123392], HG00171 [or SAME124961], HG00378 [or SAME124745] and NA20509 [or SAME124354]. All of them have publicly available WES data, RNA-seq data and high-quality variant information to validate true positive (TP) variant calls.

NA12878 sample is derived from the GM12878 cell line from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. WES and RNA-seq data of NA12878 were obtained from the Sequence Read Archive (SRA) [200] under accession numbers SRR2106342 and SRX082565, respectively. Truth data of NA12878 was generated by the Genome in a Bottle (GiaB) consortium, led by the National Institute of Standards (NIST) [212].

The remaining samples (HG00171, HG00378 and NA20509) are part of the 1000 Genomes Project [6] where their WES, RNA-seq and truth data were obtained. There are several RNA-seq datasets available from the 1000 Genomes Project, we selected the data generated by the Geuvadis consortium [213] to analyze the same type of data in the three samples, as it was the only one in common between them. Regarding the truth data, the phase III integrated variant set was used as high-quality variant information.

3.1.1.2 WES calling

The procedure is the same as for modern short variant discovery with only WES data (Section 2.1.1.1), but now variant discovery is performed in GVCF mode instead of VCF mode.

WES reads are trimmed using Trimmomatic v0.39 [58] and aligned to the human genome assembly using the BWA-MEM algorithm [69] implemented in the Sentieon utilities v202010.02 [115]. The Sentieon sort utility [115] is then used to sort and index the resulting alignment BAM files. Then, duplicate reads are removed and base quality score recalibration (BQSR) is performed using the Sentieon utilities v202010.02 [115]. Finally, variant calling is performed using Sentieon Haplotyper v202010.02 [115] in GVCF mode.

3.1.1.3 RNA-seq calling

RNA-seq calling is similar to WES calling and follows the GATK Best Practices workflow [76]. First, low quality portions of the RNA-seq reads are removed with BBDuk v35.85 [62]. Only reads with a minimum length of 35 bp and minimum base quality score of 25 are retained. The resulting reads are aligned to the reference genome using STAR v 2.7.3a [214] in two-pass mode to improve alignments around novel splice junctions [215]. Read groups are added to the sorted alignment BAM file using Picard's AddOrReplaceReadGroups [75]. Then, duplicate reads are removed, reads are split at splicing junctions into exon segments and base quality score recalibration (BQSR) is performed using the Sentieon utilities v202010.02 [115]. The split step is not performed in DNA calling and consists of splitting the RNA reads into exon segments by removing Ns and it also consists of hard-clipping any sequences overhanging into the intron regions and reassigning the mapping qualities from STAR. Finally, identification of short genomic variation is performed using Sentieon Haplotyper v202010.02 [115] in GVCF mode and with the "trim_soft_clip" option activated to exclude the soft clipped bases from the variant calling.

3.1.1.4 Joint variant calling of WES and RNA-seq data

3.1.1.4.1 Genotyping

Both WES and RNA-seq GVCF are collected and passed together to the joint genotyping tool, GVCFtyper from Sentieon v202010.02 [115] where the minimum phred-scaled confidence thresholds for calling and for emitting variants are adjusted to 20.

3.1.1.4.2 Filtering of genomic variants

Short variants are filtered out if there is a depth of coverage (DP) less than or equal to 6 reads and a strand bias (SB) greater than or equal to 2 in both the WES and RNA-seq data. It should be noted that the SB filter is only applied if the read counts for both the major allele and the minor allele are greater than or equal to 10. Sentieon automatically calculates the DP information during the variant calling whereas SB is calculated following the formula described previously in a mitochondrial heteroplasmy study [216]:

$$\left|\frac{b}{a+b} - \frac{d}{c+d}\right| / \left(\frac{b+d}{a+b+c+d}\right)$$

where a, c represent the forward and reverse reads counts of the major allele, and b, d represent the forward and reverse reads counts for the minor allele.

RNA-based variants require more filters than WES-based variants due to its greater susceptibility to false positives. Each RNA-seq calling tool uses different filtering strategies, but they are all based on similar criteria (Table 12).

Table 12. Overview of the filtering information used by different RNA-based calling tools.

Filters	**RADIA** [171]	**PMC7708150** [203]	**SNPiR** [203]	**varRED**
Repetitive regions	X	X	X	X
RNA editing sites		X	X	
Exon boundaries		X		X
Homologous regions	X	X	X	X
Not in the accessible genome	X			X
Strand Bias	X			X
Quality Control	X	X	X	X
Unique mapping	X		X	X
Extra filters	X	X	X	

Before RNA-based filtering can begin, genomic variations must be annotated. Annotation of variants is performed using ANNOVAR software [130] and different databases: (i) RepeatMasker track from the UCSC Genome Browser [158] to identify repetitive regions, (ii) the genomicsuperDups database [180] to obtain homologous regions and (iii) the ENCODE Blacklist [217] to annotate variants that are not in the accessible genome. Finally, the distance of each variant to its closest exon boundary is calculated using

BEDTools utilities v2.29.2 [159], pybedtools Python library v0.8.1 [202] and the Reference sequence (RefSeq) Gene database [127].

Finally, RNA-based variants are removed if they are found in homologous regions, interspersed repeats or low-complexity sequences, in the inaccessible genome, or within 5bp upstream of an exon start or downstream of an exon end site.

3.1.1.4.3 Classification of genomic variants

In order to facilitate the prioritization of variants, we have implemented a classification in six groups: Strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants (Figure 25). RNA-only variants do not meet the minimal threshold for DNA but meet the RNA thresholds (DP, SB and RNA-based filters) and have an RNA genotype quality (GQ) greater than 20. DNA-only variants meet the DNA thresholds (DP and SB) and have a DNA GQ greater than 20 but do not meet the RNA thresholds (DP and SB) and/or have an RNA GQ lower than or equal to 20. Strong-evidence variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and the same genotype in DNA and RNA data. ASE variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and are heterozygous for DNA but homozygous for RNA. RNA-editing variants meet the DNA and RNA thresholds, have a DNA GQ and an RNA GQ greater than 20 and are present in RNA but absent in DNA. RNA-rescue variants meet the DNA and RNA thresholds, have an RNA GQ greater than 50 but a DNA GQ lower than or equal to 20 (Figure 25).

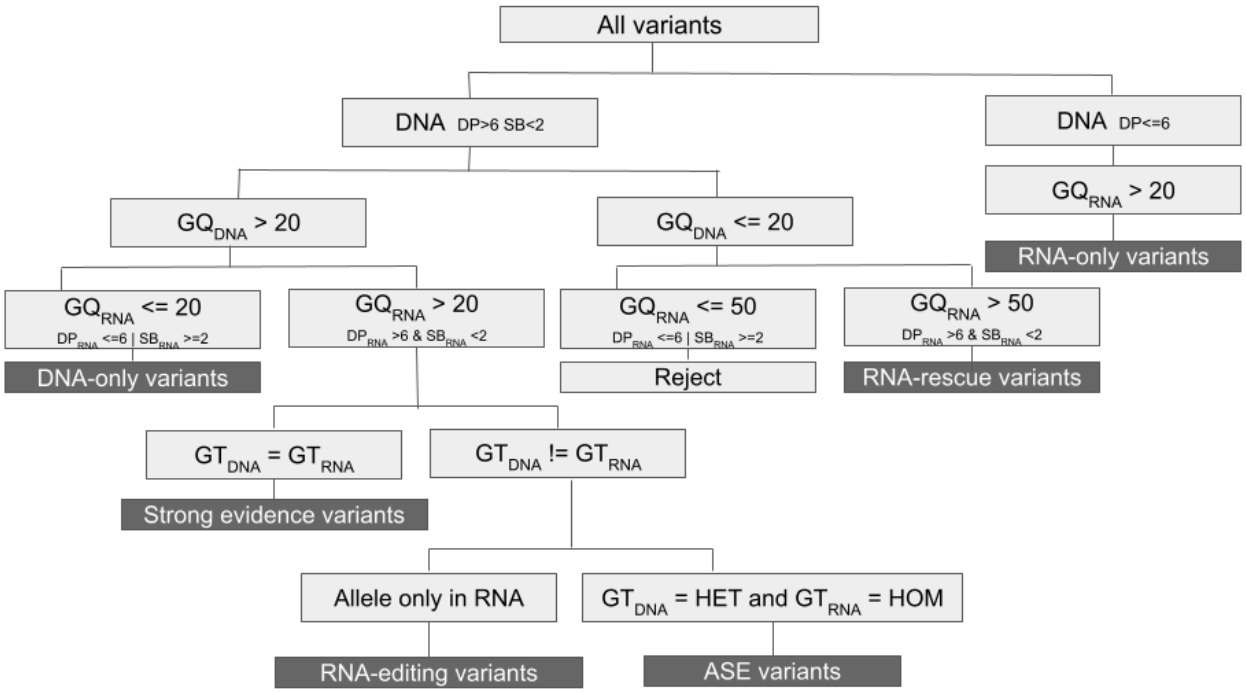

Figure 25. Overview of the varRED classification workflow.

3.1.1.5 Benchmark

The evaluation of performance of varRED was conducted according to the Global Alliance for Genomics and Health (GA4GH) best practices [218] using hap.py framework [219] with the vcfeval comparison tool [220] and the truth variant information publicly available for each sample.

We considered for evaluation two types of variant matches: (i) genotype (GT) match when the unphased genotype and alleles of a variant match in the truth and query set, and (ii) allele (AL) match when the truth and query set contain the same allele regardless of genotype. This evaluation was applied to the allele and genotype information obtained from WES data, to the allele and genotype information from the RNA-seq data, and to the allele information obtained by jointly considering both WES and RNA-seq alleles.

It should be taken into account that not all variants were considered for all evaluations. The variants to consider changed depending on the type of match (AL or GT match) and the source of information (DNA, RNA, or both). To evaluate the genotype information obtained from WES data, the RNA-only, RNA-rescue, RNA-editing and ASE variants were removed as they did not meet the minimum DNA GQ criteria (RNA-only and RNA-rescue) or they had a different allele expression (RNA-editing and ASE). To assess the genotype information obtained from RNA-seq data, the DNA-only variants were removed as they had an RNA GQ lower than or equal to 20 and the RNA-editing and ASE variants were also removed as they had a different allele expression. To evaluate the allele information, RNA-editing variants were removed from the analysis in all cases (DNA, RNA and both).

Furthermore, we perform an evaluation by type of variant according to the varRED classification: Strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants.

3.1.1.5.1 Comparison of varRED with the short variant discovery from WES data

To assess the impact of adding RNA-seq data for variant calling, we have applied the short variant discovery pipeline described in Section 2.1.1.1 to the WES data for each sample and evaluated their results using using hap.py framework [219] with the vcfeval comparison tool [220] as for varRED.

3.1.2 Results

3.1.2.1 Comparison of varRED with the short variant discovery from WES data

There was an increase in the total number of identified variants when varRED was applied instead of performing short variant discovery from WES data (WES calling) (Table 13). There are two types of variant matches considered for evaluation (GT match and AL match) and each includes different variants. For this reason, we evaluated the number of variants obtained from varRED depending on the type of match and compared it with the total variants obtained from WES calling.

Table 13. Overview of the total variants identified with varRED and with the short variant discovery from WES data.

Sample	Analysis	Variants		
		Total	SNPs	Indels
NA12878	varRED	273660	259503	14157
	WES calling	259692	224840	34852
HG00171	varRED	142760	122761	19999
	WES calling	130068	113298	16770
HG00378	varRED	150370	129610	20760
	WES calling	133523	116342	17181
NA20509	varRED	124512	110355	14157
	WES calling	97676	89030	8646

Evaluation of the genotype information extracted from DNA data included two types of variants: strong-evidence and DNA-only variants. These two types of variants were enough to exceed the total number of variants obtained with WES calling in all samples except NA20509 (Table 14). For evaluation of genotype information, RNA data includes strong-evidence, RNA-only and RNA-rescue variants. The number of variants detected for genotype matching from RNA data was much lower than those obtained from DNA data with varRED (Table 14).

Regarding the assessment of AL matches, all variant types except RNA-editing variants were considered. Here, the number of variants detected with varRED from DNA data was higher than with WES calling and even higher when the allelic information of both DNA and RNA was considered (Table 14).

Table 14. Overview of the number of variants identified with varRED and with the short variant discovery from WES data.

Sample	Analysis	Data type	Match	Variants		
				Total	SNPs	Indels
NA12878	varRED	DNA	GT	449974	224987	224987
			AL	266317	229854	36463
		RNA	GT	47288	42359	4929
			AL	68881	60859	8022
		DNA+RNA	AL	287333	248646	38687
	WES calling	DNA	-	259692	224840	34852
HG00171	varRED	DNA	GT	131453	113844	17609
			AL	134158	116030	18128
		RNA	GT	13295	11425	1870
			AL	19922	17403	2519
		DNA+RNA	AL	140257	120943	19314
	WES calling	DNA	-	130068	113298	16770
HG00378	varRED	DNA	GT	134219	116216	18003
			AL	138657	120023	18634
		RNA	GT	19761	17454	2307
			AL	27187	24154	3033
		DNA+RNA	AL	147339	127262	20077
	WES calling	DNA	-	133523	116342	17181
NA20509	varRED	DNA	GT	96406	87806	8600
			AL	101784	92342	9442
		RNA	GT	32702	28109	4593
			AL	42550	37076	5474
		DNA+RNA	AL	119228	106337	12891
	WES calling	DNA	-	97676	89030	8646

Evaluation of recall, precision and F-score in the identification of variants was performed with respect to the genotype information (GT match) and the allele information (AL match). There was an improvement in SNP genotyping from DNA and RNA data in most samples (HG00171, HG00378 and NA20509), but a decrease in the precision and F-score when genotyping indels from DNA (NA12878, HG00171 and HG00378) and RNA (NA12878, HG00171, HG00378 and NA20509) (Figure 26). Concerning the genotyping of SNPs from DNA data, the recall was the same with varRED or WES calling for all samples. Precision was the same in NA12878 (Figure 26 – A) but higher in varRED than in WES calling for HG00171, HG00378 and NA20509 (Figure 26 – B, C, D). F-score was higher in varRED than in WES calling for HG00171 ($F\text{-score}_{varRED} = 93.6\%$ and $F\text{-score}_{WES\text{-}calling} = 93.5\%$), HG00378 ($F\text{-score}_{varRED} = 93.3\%$ and $F\text{-score}_{WES\text{-}calling} = 93.1\%$) and HG00378 ($F\text{-score}_{varRED} = 93\%$ and $F\text{-score}_{WES\text{-}calling} = 92.7\%$) (Figure 26 – B, C, D) but lower for NA12878 ($F\text{-score}_{varRED} = 98.0\%$ and $F\text{-score}_{WES\text{-}calling} = 98.1\%$) (Figure 26 – A). Regarding the genotype of indels from DNA data, recall, precision and F-score were higher with varRED than with WES calling for NA20509 (Figure 26 – D), while for the rest of samples, they were lower (with the exception of the recall of NA12878 which is the same) (Figure 26). Concerning the genotyping of SNPs from RNA data, there was an improvement in recall, precision and F-score for HG00171, HG00378 and NA20509 with the exception of the recall for NA20509 which was the same (Figure 26 – B, C, D). However, all metrics for NA12878 were lower with varRED than with WES calling (Figure 26 – A). Regarding the genotype of indels from RNA data, there was a decrease in precision and F-score for all samples (Figure 26) and an increase in recall for NA12878, HG00171 and HG00378 (Figure 26 – A, B, C), but there was also a decrease in recall for NA20509 (Figure 26 – D).

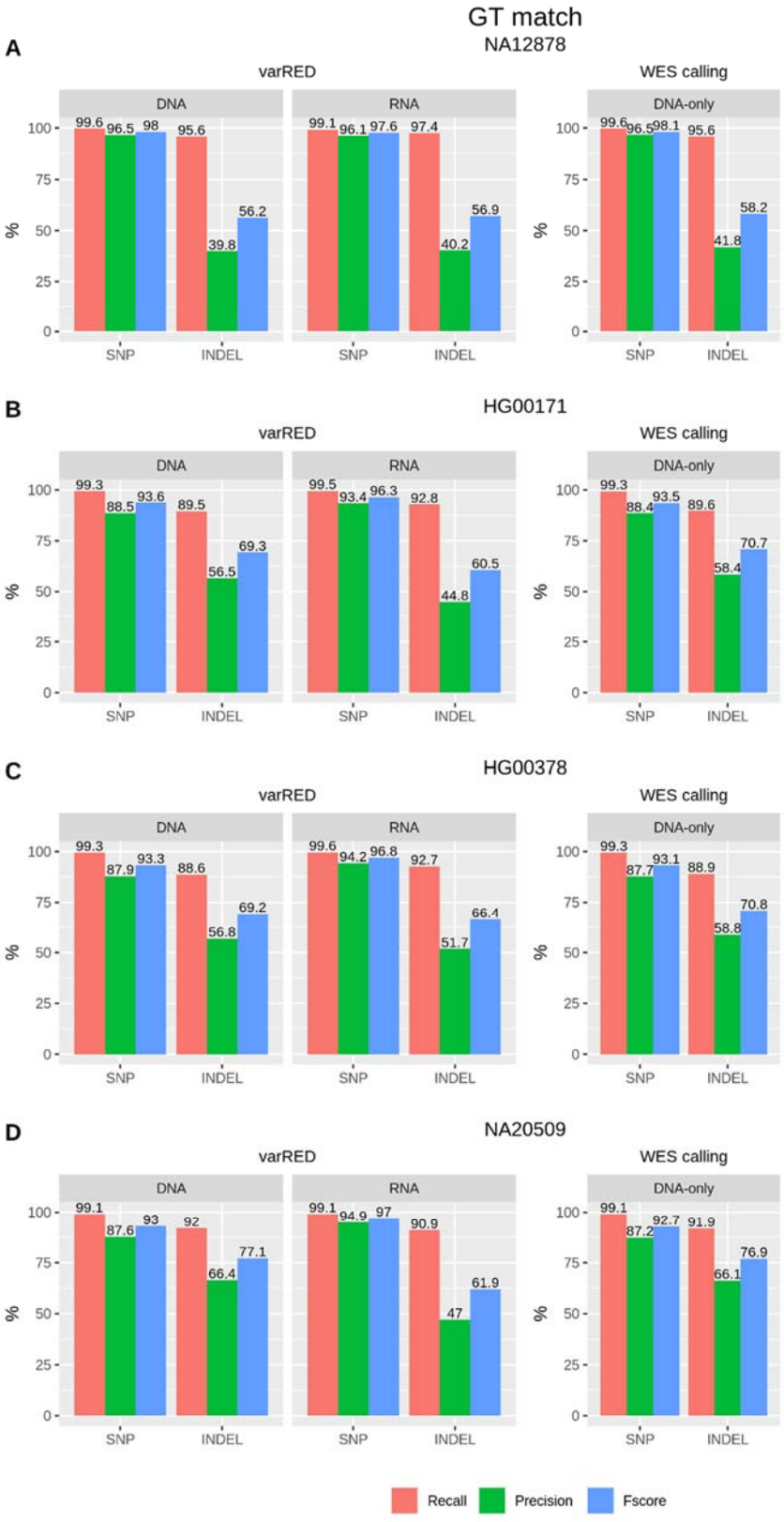

Figure 26. Benchmark results of the genotype match with varRED and with WES calling for different samples. Recall, precision and F-score in the identification of the genotype of SNPs and indels with varRED and WES calling for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D).

In general, there was a decrease in precision and F-score of allele identification of SNPs and indels when using DNA and RNA data with varRED instead of performing WES calling (Figure 27). With regard to the alleles of SNPs and indels of DNA data, precision and F-score decreased in all the samples (Figure 27). Concerning the alleles obtained from RNA data, their precision and F-score in identification of SNPs and indels decreased in all samples, but increased for the identification of SNPs in HG00378 and NA20509 (Figure 27). Regarding the alleles obtained jointly considering DNA and RNA data, there was an improvement in the recall, precision and F-score to identify SNPs in NA20509 (Figure 27 – D), however, there was a decrease in the precision and F-score in the identification of SNPs in the rest of samples (Figure 27 – A, B, C). In the detection of alleles in indels there was a decrease in the precision and F-score in all samples when considering jointly DNA and RNA data (Figure 27).

To sum up, there was an improvement in genotyping, but a decrease in allele match metrics when varRED is used instead of WES calling. However, these changes of the metrics are minor, so we can conclude that we were able to increase the number of detected variants without a great effect on performance.

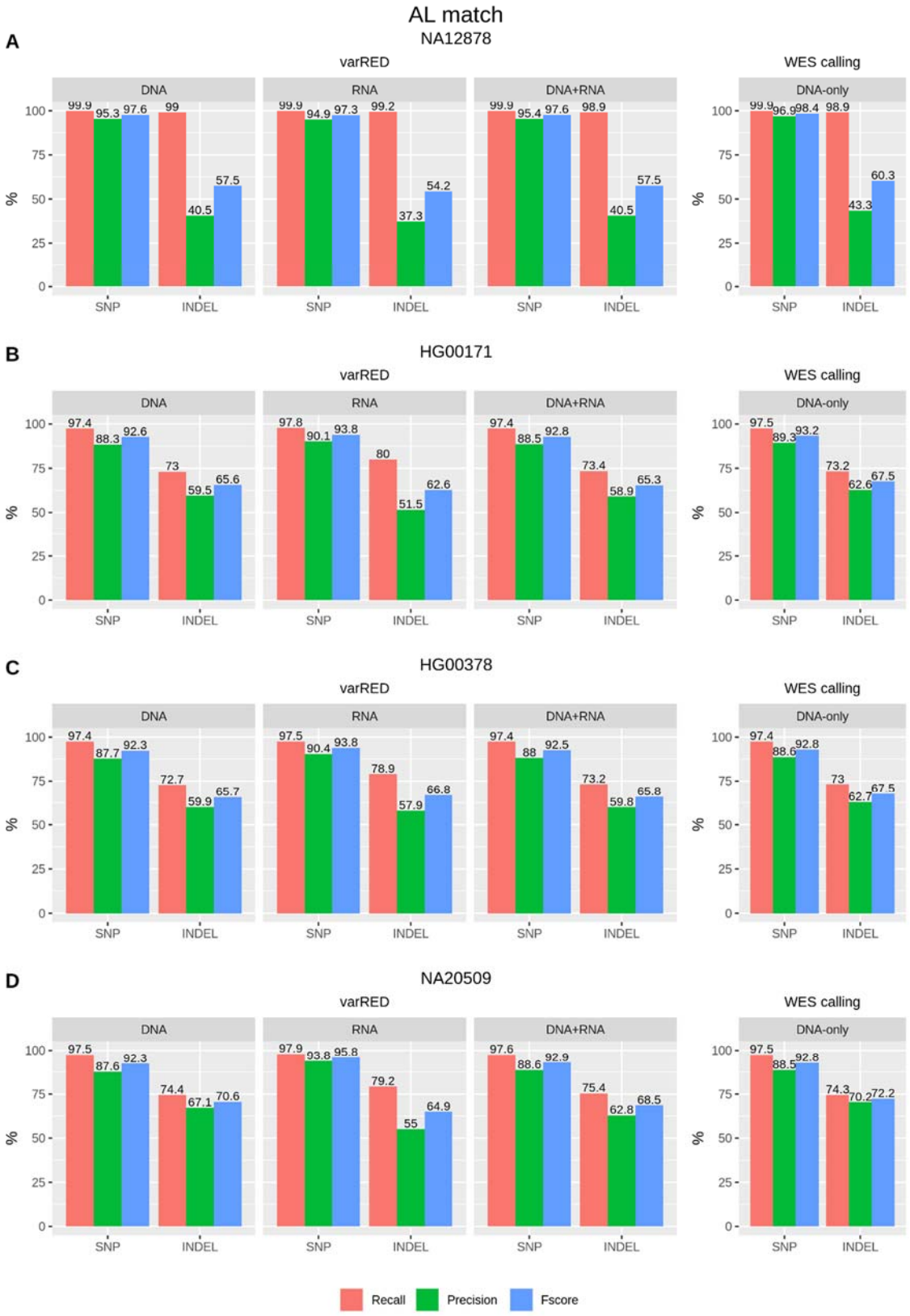

Figure 27. Benchmark results of the allele match with varRED and with WES calling for different samples. Recall, precision and F-score in the identification of the alleles of SNPs and indels with varRED and WES calling for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D).

3.1.2.2 Benchmark results by variant type

3.1.2.2.1 Strong evidence variants

Strong-evidence variants have a reliable DNA and RNA quality and the same genotype. Thus, the same alleles and genotypes will be evaluated regardless of the type of data and therefore we will obtain the same metrics.

We observed a high recall, precision and F-score in genotyping and identification of alleles for Strong-evidence SNPs and indels (Figure 28). As expected, these metrics were the highest of all the variant types (Supplementary Table A.1). Strong-evidence SNPs were identified with a recall, precision and F-score greater than 98% in all samples (Figure 28). Strong-evidence indels did not achieve as high metrics as SNPs but they did achieve the highest metrics of all types of variants (Supplementary Table A.1).

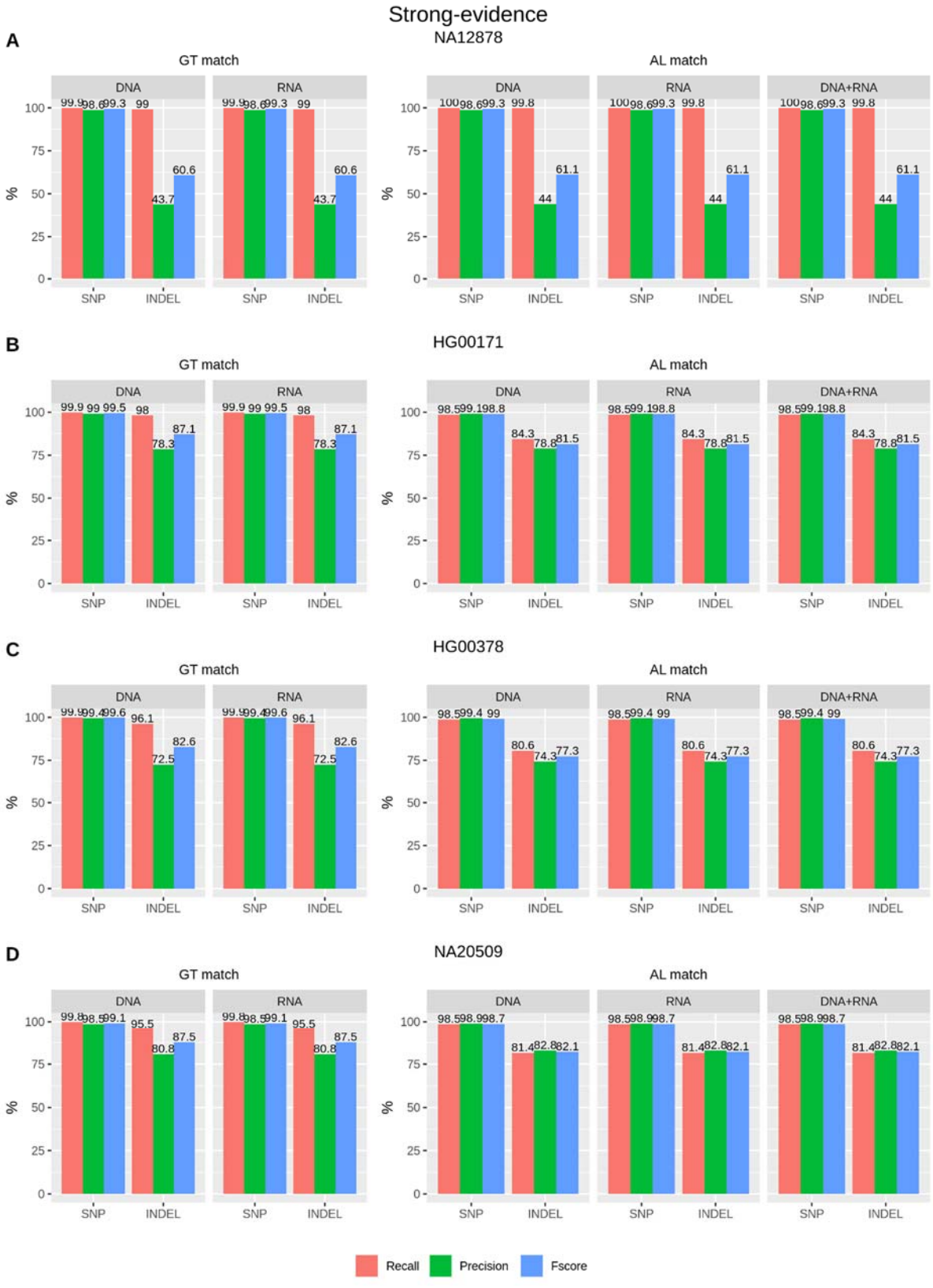

Figure 28. Benchmark results of Strong-evidence variants for different samples. Recall, precision and F-score in the identification of Strong-evidence variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.2.2.2 DNA-only variants

DNA-only variants have an RNA coverage lower than or equal to 6 and/or a strand bias in RNA greater than or equal to 2 and/or an RNA GQ lower than or equal to 20. For this reason, the alleles and genotype of the RNA data cannot be trusted, and therefore RNA-editing and ASE cannot be detected.

Regarding the genotyping from RNA data, there was a low recall, precision and F-score as it was expected (Figure 29). Furthermore, we observed a reliable recall, precision and F-score in genotyping SNPs and indels from DNA data but lower than the one observed in Strong-evidence variants (Supplementary Table A.1). DNA-only SNPs were genotyped with a F-score greater than 92% in all samples (Figure 29). When comparing the metrics in genotyping of DNA-only variants with the final GT metrics of varRED from DNA data, which includes Strong-evidence and DNA-only variants, we observed lower recall, precision and F-score in DNA-only variants compared to the final GT metrics (Supplementary Table A.1).

There was reliable recall, precision and F-score in identifying alleles of SNPs and indels from DNA data, RNA data and jointly using both (DNA+RNA) (Figure 29). Alleles of DNA-only SNPs were identified with a F-score greater than 91% in all the cases (Figure 29). We observed a decrease in recall, precision and F-score in identifying DNA-only SNPs alleles relative to total SNPs alleles, which includes all variant types except RNA-editing variants, when using DNA data, RNA data and both (Supplementary Table A.1). However, in general, there was a higher precision and F-score in the allele identification of DNA-only indels than of total indels in DNA, RNA and with the joint use of both (Supplementary Table A.1).

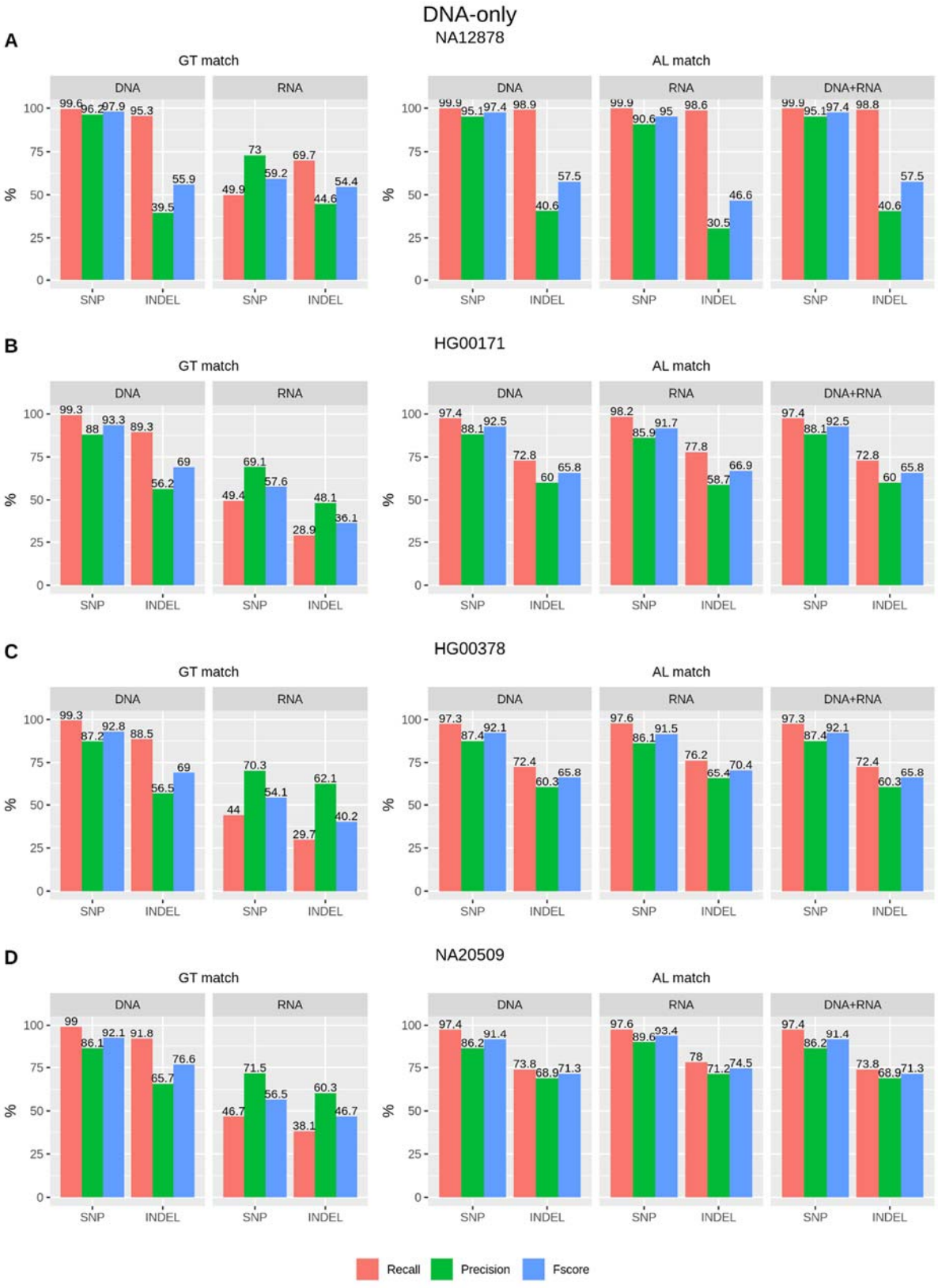

Figure 29. Benchmark results of DNA-only variants for different samples. Recall, precision and F-score in the identification of DNA-only variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.2.2.3 RNA-only variants

RNA-only variants have a DNA coverage lower than or equal to 6, therefore the alleles and genotype cannot be trusted from DNA data and ASE and RNA-editing cannot be identified. The presence of ASE variants can negatively affect the metrics when evaluating genotypes in RNA data since the truth dataset was derived from DNA data where the expression of the variants is unknown. The presence of RNA editing variants can also negatively affect the metrics when evaluating both genotype and alleles in RNA data since RNA editing occurs after DNA transcription and synthesis.

As expected, we observed a low recall, precision and F-score in genotyping from DNA data (Figure 30). Regarding genotyping of RNA data, there was good recall, precision and F-score for SNPs and indels but lower than that observed in Strong-evidence variants (Supplementary Table A.1). Genotyping of RNA-only SNPs from RNA data had a F-score greater than 94% in all samples (Figure 30). Moreover, we observed lower recall, precision and F-score in genotyping of RNA-only variants relative to total variants in RNA data, including Strong-evidence, RNA-only and RNA-rescue variants (Supplementary Table A.1).

Reliable metrics were observed in allele identification of RNA-only variants from RNA data and with the joint use of DNA and RNA data (Figure 30). There were also good metrics when using DNA data, but as expected, they were lower (Figure 30). In all the cases (DNA, RNA or DNA+RNA), alleles of RNA-only SNPs were identified with a F-score greater than 89% (Figure 30). We observed higher precision and F-score in detecting alleles for RNA-only SNPs than for total SNPs, including all variant types except RNA-editing variants, when using RNA data and RNA with DNA data, but a decrease when using DNA data (Supplementary Table A.1). However, there was a decrease in precision and F-score for allele identification in RNA-only in relation to the total indels in all cases (DNA, RNA or DNA+RNA) (Supplementary Table A.1).

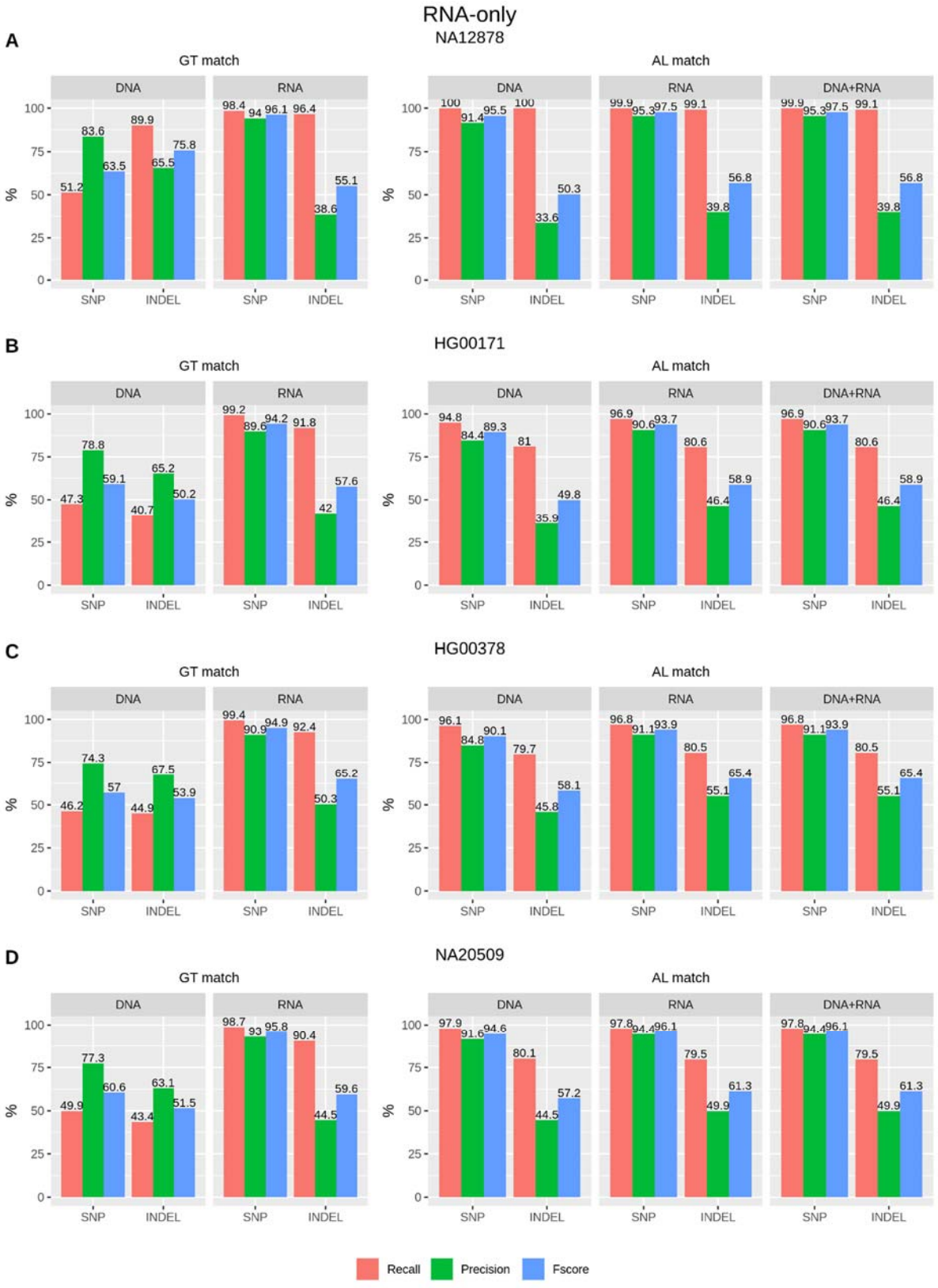

Figure 30. Benchmark results of RNA-only variants for different samples. Recall, precision and F-score in the identification of RNA-only variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.2.2.4 ASE variants

ASE variants are heterozygous for DNA but homozygous for RNA because only one of the alleles has been expressed. For this reason, recall, precision and F-score in the AL match were the same in the DNA and RNA data, but the metrics in the GT match changed (Figure 31). We observed an improvement in the metrics when using DNA data instead of RNA data (Figure 31), this is because the truth dataset had been obtained from DNA data so the expression of the variants was not considered. In any case, the identification of ASE variants had low recall, precision and F-score in all samples and cases except for SNPs identification in NA12878 where all metrics were greater than 70% (Figure 31).

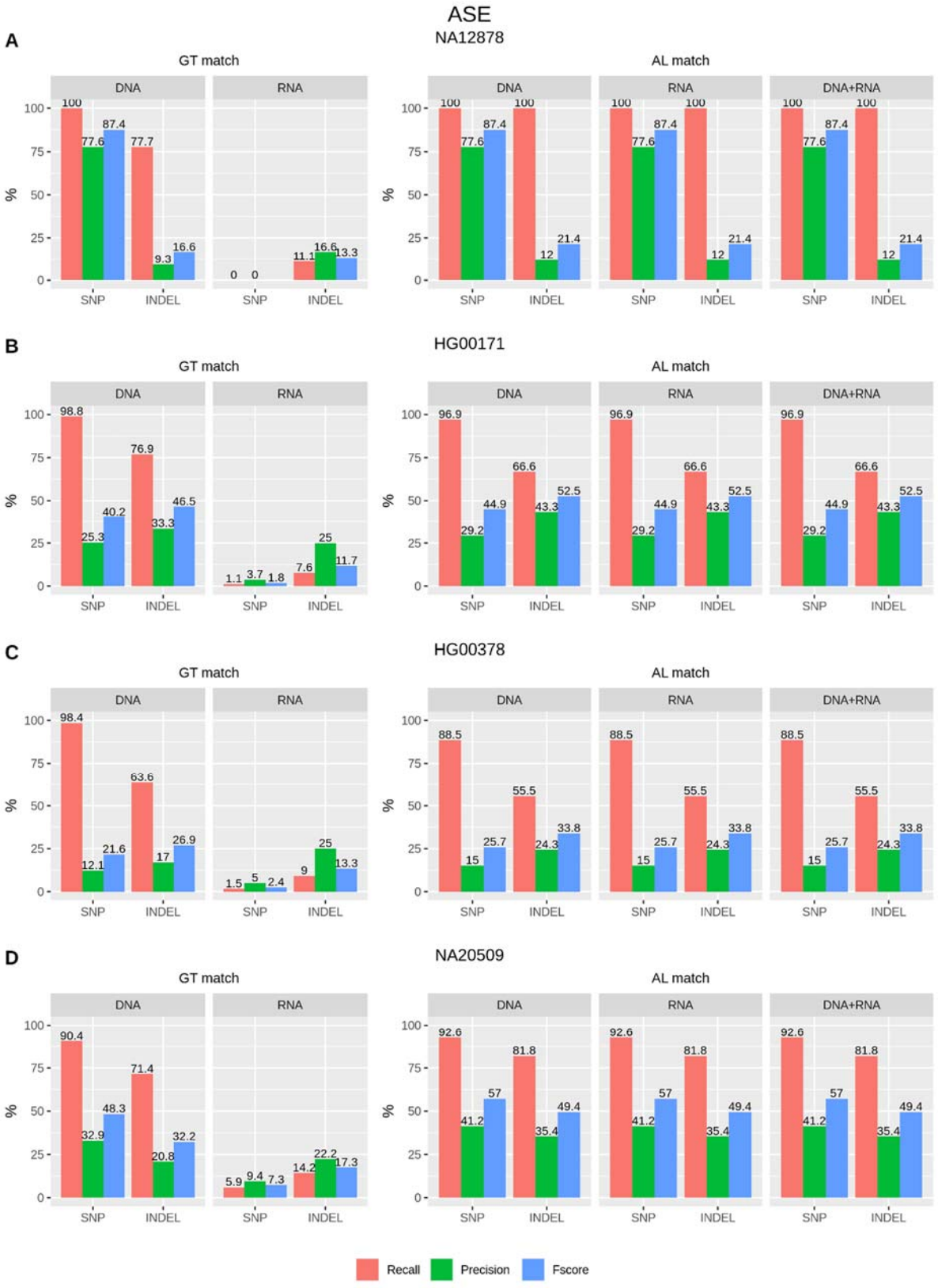

Figure 31. Benchmark results of ASE variants for different samples. Recall, precision and F-score in the identification of ASE variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.2.2.5 RNA-editing variants

RNA-editing variants are present in RNA but absent in DNA. They cannot be evaluated as our truth dataset was derived from DNA data where RNA-editing variants were not present. For this reason, we observed very low recall, precision and F-score in genotype and allele identification from RNA data (Figure 32).

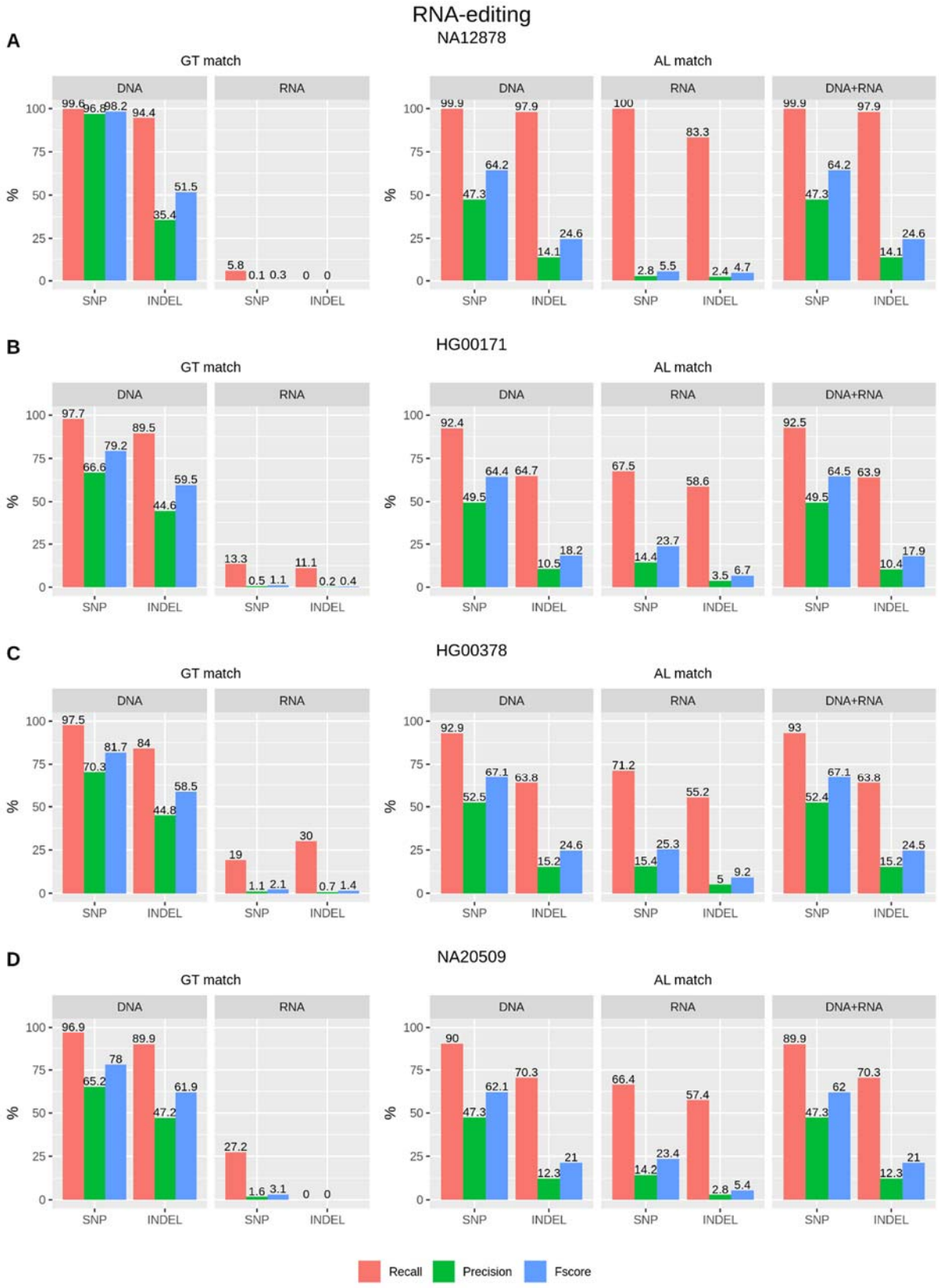

Figure 32. Benchmark results of RNA-editing variants for different samples. Recall, precision and F-score in the identification of RNA-editing variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.2.2.6 RNA-rescue variants

RNA-rescue variants meet the minimum thresholds of coverage and strand bias in DNA and RNA, have an RNA GQ greater than 50 but a DNA GQ lower than or equal to 20. Therefore, the genotype information from RNA data can be trusted but not the genotype information from DNA data. In addition, RNA-editing and ASE variants cannot be detected and will be present within the RNA-rescue variants, which may affect the evaluation.

We observed sufficient recall, precision and F-score in the genotype and allele identification of RNA-rescue SNPs in RNA data as the metrics were greater than or close to 70% in all samples (Figure 33). However, identification of indels in RNA data had very low metrics (Figure 33).

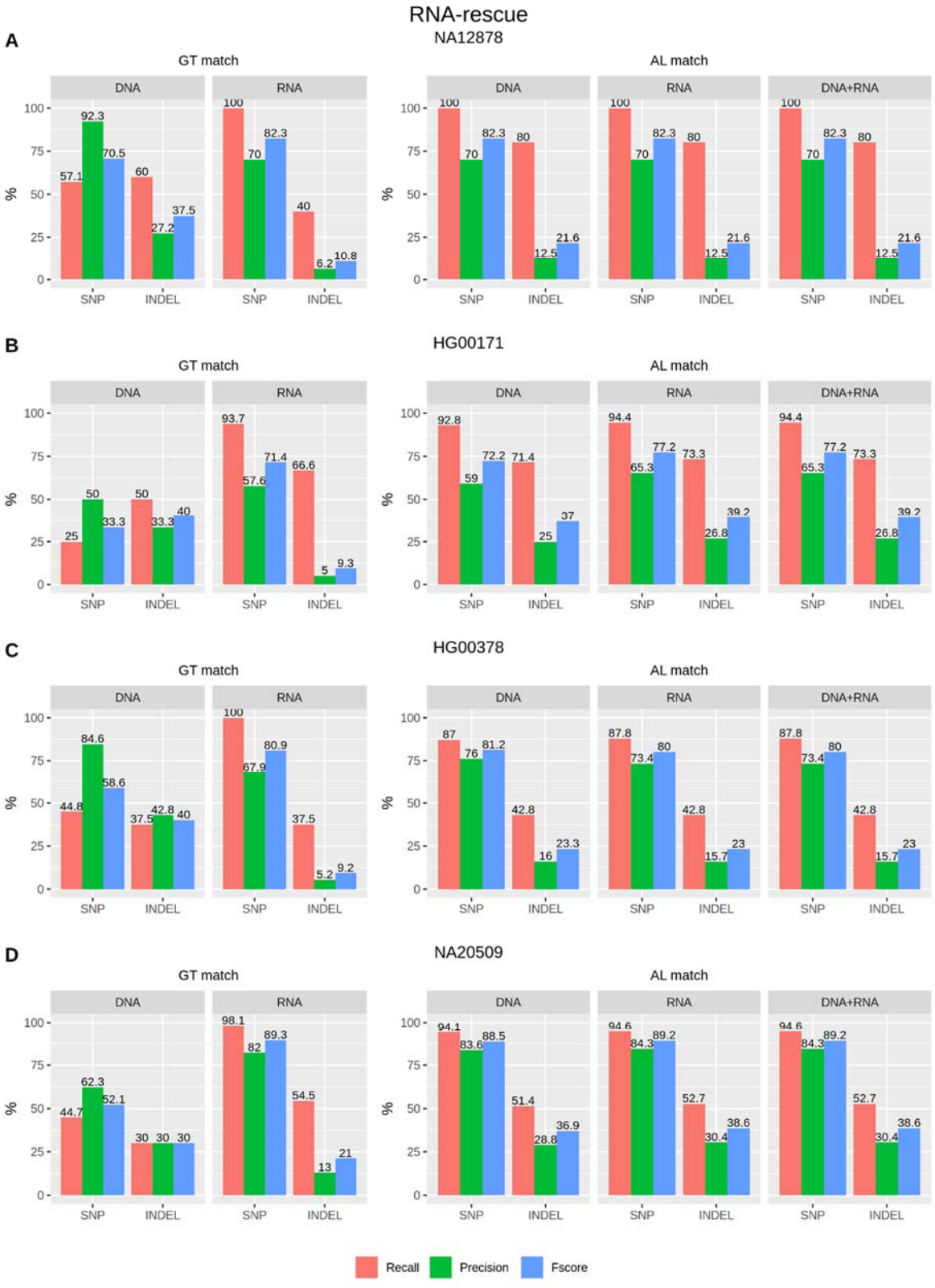

Figure 33. Benchmark results of RNA-rescue variants for different samples. Recall, precision and F-score in the identification of RNA-rescue variants with varRED for NA12878 (A), HG00171 (B), HG00378 (C) and NA20509 (D) considering genotype (GT) and allele (AL) match.

3.1.3 Conclusion

varRED provides an easy-to-use tool for germline short variant discovery from WES and RNA-seq data. By incorporating RNA-seq data, it increases the identified variants with virtually no performance impact and enables the identification of ASE and RNA-editing variants. It also improves prioritization and interpretation of variants by classifying them into six groups: Strong-evidence, DNA-only, RNA-only, ASE, RNA-editing and RNA-rescue variants.

Chapter 4 | Genomics data analysis platform

This chapter describes the development of GINO, an online platform for the visualization and interpretation of genomics variants. There, all germline variants identified through the workflows developed in the previous chapters (Chapter 2 and Chapter 3) are displayed to facilitate variant interpretation.

4.1 GINO: a platform for visualization and interpretation of variants

GINO is a platform for the visualization and interpretation of human genomic variation under license for use. The user's raw sequence data is analyzed on an internal server where germline variants are detected following the pipelines and workflows developed in this thesis (Chapter 2 and Chapter 3). Then, the identified variants are uploaded to GINO where the user can visualize and browse the results. GINO is available at the following link: https://cloud.gino.sequentiabiotech.com/. An article on GINO is scheduled to be submitted to bioinformatics journals and uploaded to bioRxiv until final publication, a free repository of unpublished preprints.

4.1.1 Implementation

The creation of GINO has required the development of: (i) a robust database for data storage and management and (ii) an easy-to-use graphical interface to dynamically display and browse genomics variants.

4.1.1.1 Database structure

The GINO database has been developed using the MySQL relational database v5.7.34 and is stored on an Ubuntu server v18.04. As the GINO data, consisting of a series of germline variants and their functional annotations, have a well-defined structure and are related to each other, the relational database has been our system of choice. Using a relational database provides us with a highly organized and structured system to ensure the validity of database transactions, safeguard data integrity, and reduce anomalies. In addition, it allows us to store both the data and the relationship between these data. However, it also reduces the level of flexibility compared to what we could achieve using a non-relational database, where it is not necessary to predefine the number of data types and variables to store [221]. Despite the increasing use of non-relational databases today [221]–[224], the MySQL relational database has proven its effectiveness in handling genomic data as it has been used in two large projects: the Ensemble project [225] and the University of California Santa Cruz (UCSC) Genome Browser [158].

GINO is designed to be run by multiple users who perform a series of experiments. Each of these experiments can consist of several samples that can be related to each other. Considering this, in order to structure the database, each of these information are stored in independent tables but related to each other

through their keys. Keys are very useful as they establish relationships between tables and also uniquely identify a row in a table. Specifically, GINO contains four tables to store information, which are called: (i) users, (ii) experiments, (iii) samples and, (iv) families (Figure 34).

Furthermore, the short variants and CNVs identified for each sample are stored using different sets of tables but with a similar structure. In both cases, the aim of our storage system is to minimize space usage on the basis that each sample has multiple variants that can be shared between samples. Regarding the storage of short variants, a table of variants and multiple tables of experiments were created (Figure 34), both table types are linked by a fingerprint. This fingerprint has been created from a combination of the position of the variant in the genome and the reference and alternative alleles, such as "chr1_13657_13658_AG_-". The table of variants contains sample-independent information, such as the annotations obtained using ANNOVAR (Table 4) which include the affected gene, the clinical significance of the variant or the population allele frequency. The tables of experiments contain information about variants that are unique to each sample, such as mapping quality (MQ) or genotype quality (GQ). Each user can have multiple tables of experiments which, in turn, contain information for multiple samples; this division of data avoids reaching the maximum number of rows in a table established by the SQL engine (approximately one billion) [226]. When a new sample is uploaded to the GINO database, if it contains a new variation, this is added to the table of variants and to the corresponding table of experiment, but if it is an already existing variant, simply the sample-related information (MQ, GQ, etc.) is added to the experiment table, thus avoiding duplicated information. Concerning the CNV storage in the GINO database, the same structure is used as for the short variants: a table of CNVs and multiple tables of experiments. However, the information stored in the table of CNVs corresponds to the AnnotSV annotations (Table 6) and the fingerprint is a combination of the CNV type (DEL or DUP) and the affected gene and region, such as "DUP_ABCC2_exon24_exon25".

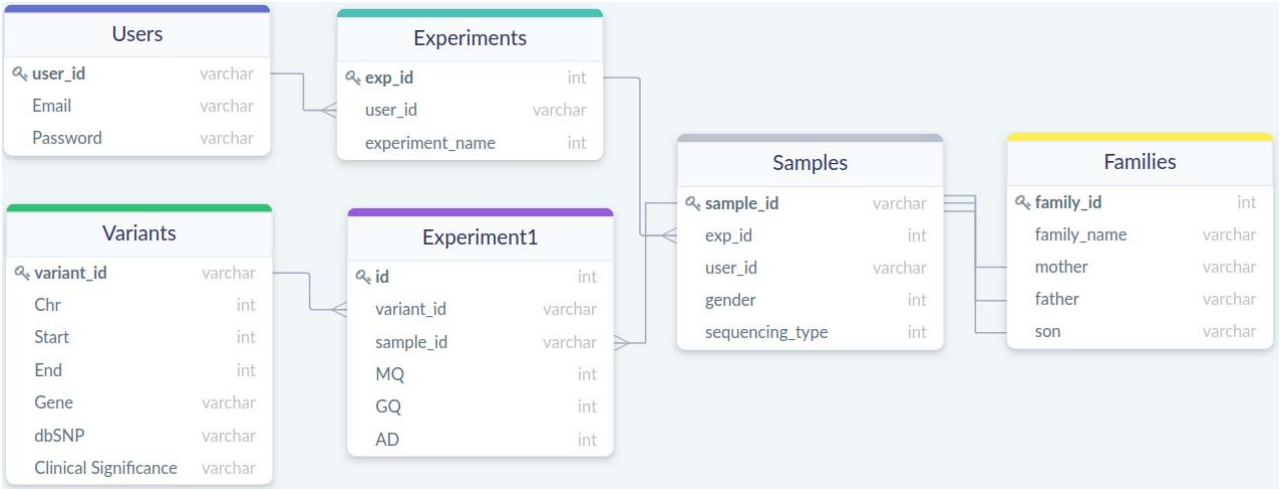

Figure 34. Toy example of GINO database structure.

4.1.1.2 Graphical interface

The graphical interface has been built using three different programming languages: HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and Javascript. This interface communicates with the database through the PHP v5.6 language. There are two main sections in the graphical interface of GINO: samples and analysis. The samples section provides a summary of all user samples, while the analysis section allows users to inspect all variants of a specific sample (short variants or CNVs) or a specific parent-child trio.

The samples section consists of a table of samples (Figure 35). There, users can search for a sample of interest via a search box and export the table of samples to an Excel file. The table of samples provides different columns with relevant information about each sample:

- Sample name
- Experiment name (Batch). By clicking on the name of the experiment, the user can get information about the experiment, such as a brief description, the creation date or the number of samples belonging to that experiment.
- Family name
- Metadata:
    - Sequencing platform (e.g., Illumina or Ion Torrent)
    - Reference genome
    - Gender
    - Age
    - Layout (single-end or paired-end)
    - DNA sequencing strategy (WGS, WES or TS)
    - Sequencing kit
- General metrics: percentage of the target at different coverages (10x and 20x).
- Detected variants: number of short variants detected in the sample. Clicking on this column will display the analysis section showing all the short variants in the sample.
- Copy Number Variants: number of CNVs identified in the sample. Clicking on this column will display the analysis section showing all the CNVs in the sample.
- Exports: The main samples files can be downloaded: VCF and bigWig. The VCF file contains all the short variants of the sample. The bigWig file stores the sample coverage information, that is the number of reads overlapping each base of the genome.

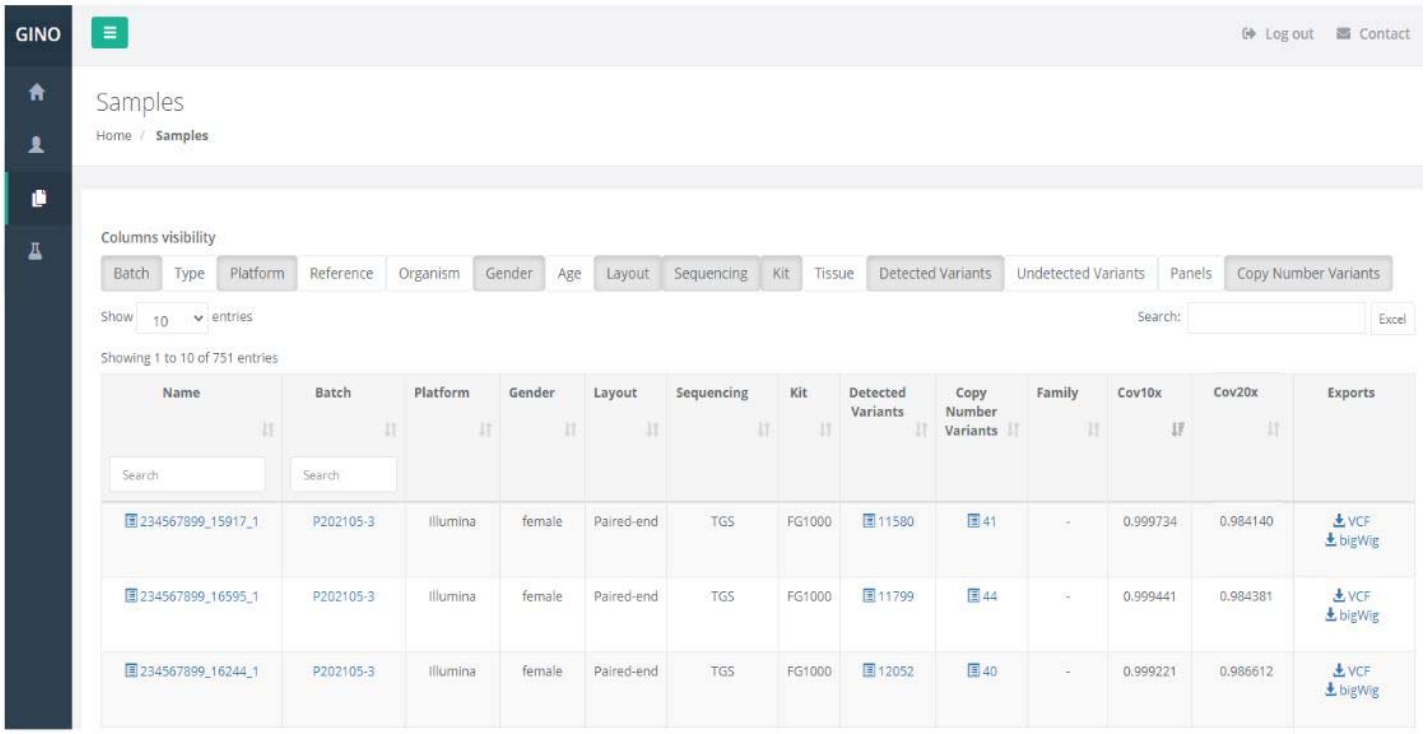

Figure 35. Screenshot of the samples section in GINO.

The analysis section is composed of three different visualization tools: the short variant browser, the parent-child trios displayer, and the copy number variant viewer.

The short variant browser focuses on the visualization of the short variations of a sample (SNPs and indels). The main core of this section is an interactive table showing all variations of a sample along with their functional annotation, complemented by the Integrative Genomics Viewer (IGV) tool [74], by a table of variants of interest selected by the user and by an automatic report creation tool.

Concerning the interactive table, it allows to explore genomic variations and filter them according to the user's needs. For example, variants can be filtered by chromosome, by genotype, by pathogenicity or by affected gene using the filtering boxes in the table header (Figure 36). Moreover, an additional button displays a pop-up for further filtering of variants regarding other features not shown in the table columns, such as mapping quality (MQ), genotype quality (GQ) or population frequency (Figure 37). In addition, potentially false positive variants are highlighted with a warning and can be filtered from the variant table (Figure 36). Two criteria are used to define these potentially false positives: (i) an odd allelic balance (AB) when the ratio of allelic depth is not ~100% or ~50% and (ii) a high homology (HOM) when the variant is inside a region with a percentage of similarity greater than 90% with another locus in the genome, according to the genomicSuperDups database [180].

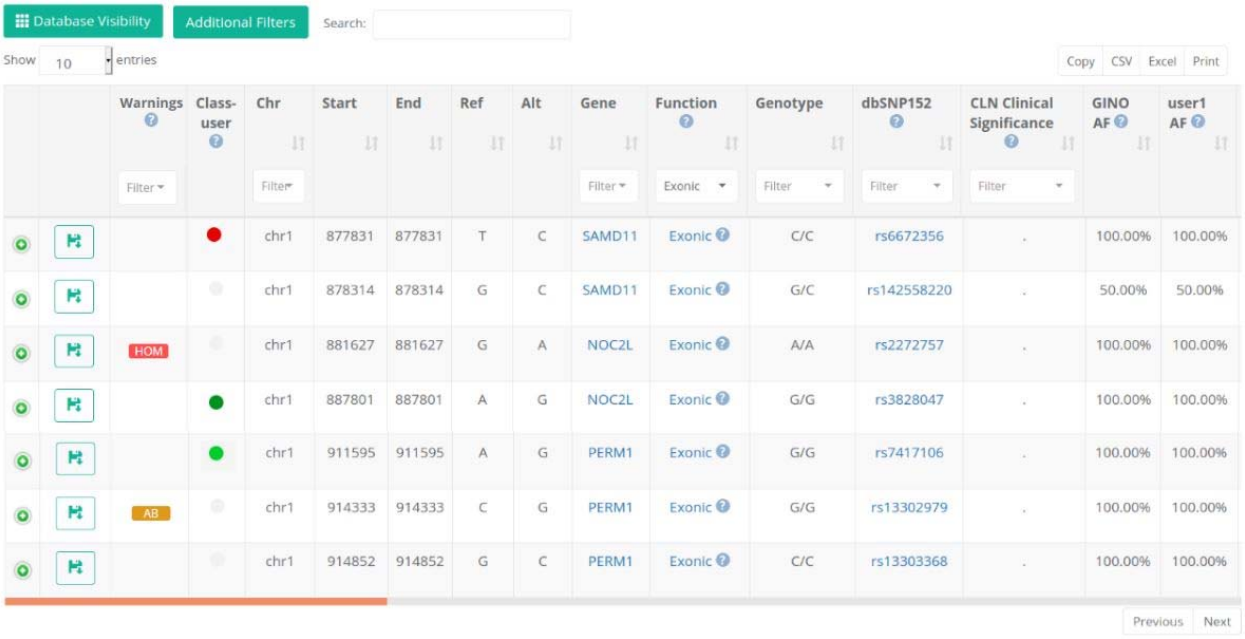

Figure 36. Interactive table for browsing and filtering variations in GINO's short variant browser.

Figure 37. Additional filters pop-up in GINO's short variant browser.

This core table shows summary information of each variant but if user clicks on the "plus" button, a child row appears with more information about the quality of the variant (Sample section), the gene affected (Gene section), its pathogenicity according to the ACMG guidelines [166], [167] (ACMG section), the allelic frequency of that variant in different populations (Population section) and the prediction of its functional consequences (Functional section) (Figure 38). Furthermore, users can customize the table to be adapted to their specific needs: the main information to be showed/hidden in the columns of the table can be switched using the database visibility button (Figure 36 and 38) and the column size and order can be

customized through a click and drag operation, letting the user move columns that they wish to compare next to each other for easier comparison. In addition, users can customize the number of variants displayed per page (Figure 36 and 38). This pagination is extremely important for the usability and speed of GINO, as it allows not to load all the variants of a sample at once, but only a custom number.

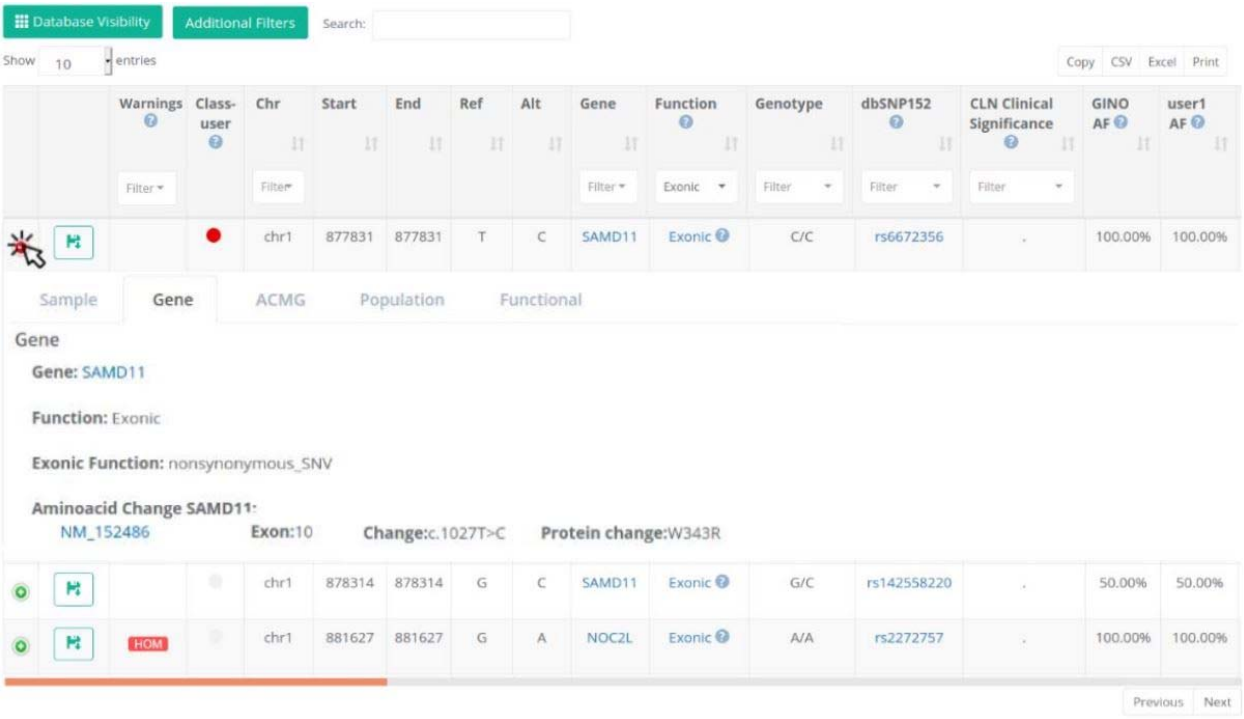

Figure 38. Additional information displayed in the child-rows of the interactive table of the GINO's short variant browser.

It is worth stressing two important features in the table wherein there is even a higher individualized experience: pathogenic user classification and GINO allele frequency (AF) (Figure 39). The pathogenic user classification is of great value for clinical diagnosis and allows the user to manually classify the variations according to their own criteria into eleven different categories based on their pathogenicity: benign, likely benign, uncertain significance, likely pathogenic, pathogenic, common artefact, drug response, disease association, risk factor, protective and phenotype association (Figure 39). The GINO allele frequency (AF) is based on an internal pipeline that estimates the alternative allele frequency for each variant in the GINO database (all samples in GINO) and in the user database (all samples of the user). The estimation of the allele frequency is of fundamental importance in population genetic analyses and in association mapping, so the GINO allele frequency adds an important additional value to GINO. It is important to note that only samples in which the same genomics regions are sequenced are used to calculate this AF value. Thus, we compute different AFs for the same variation depending on the type of sequencing (WGS, WES or TS) and the sequencing kit. The reasoning behind this approach is that when a genomic region is not present in the VCF file, it can be difficult to know the reason, it may be due to absence of variation, to the genomic region not being sequenced, to insufficient coverage or to technical errors. This

issue could be addressed through the usage of a GVCF file instead of a VCF file, since the GVCF has records for all sites. However, we have opted for using VCF files because they are much smaller than GVCF files, which saves us computational resources and storage space.

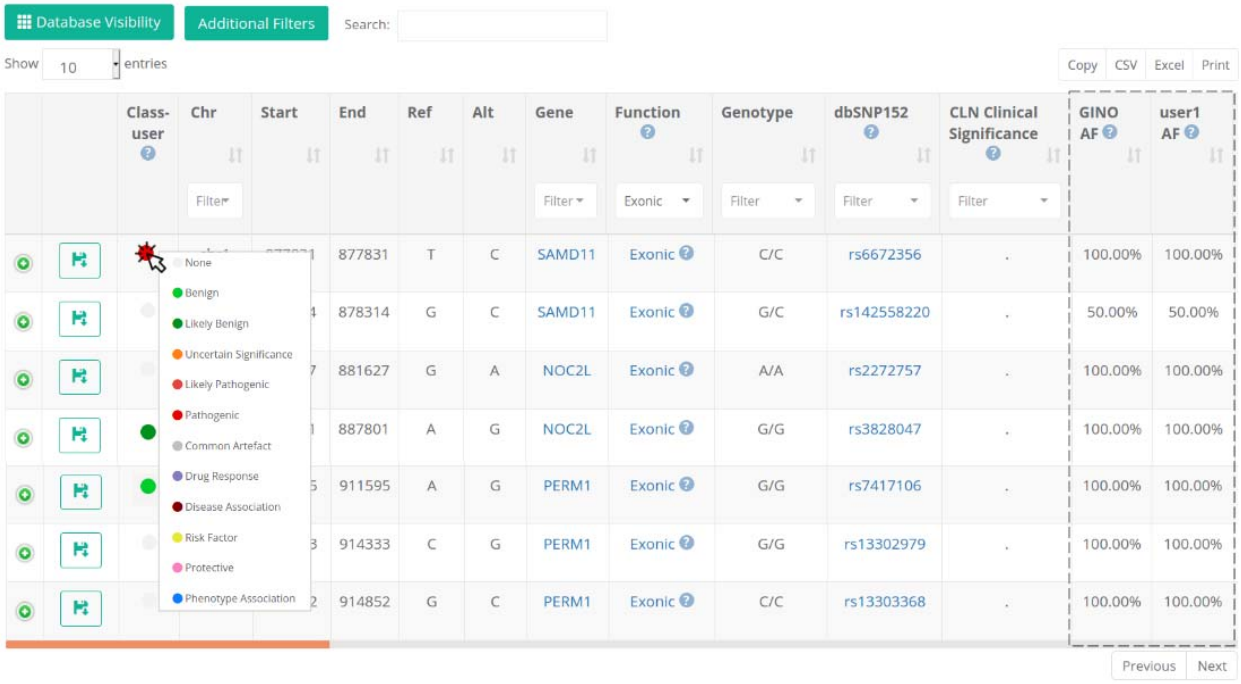

Figure 39. Pathogenic user classification and GINO allele frequency in the interactive table of the GINO's short variant browser.

Regarding IGV [74], it is a high-performance visualization tool that is embedded in our variant browser and allows interactive exploration of variants and genomic datasets. Using IGV, users can explore variations of a sample at their genomic locations, can compare them to the reference genome, and can check how many reads support the reference and the alternative allele or which gene the variant can affect (Figure 40).

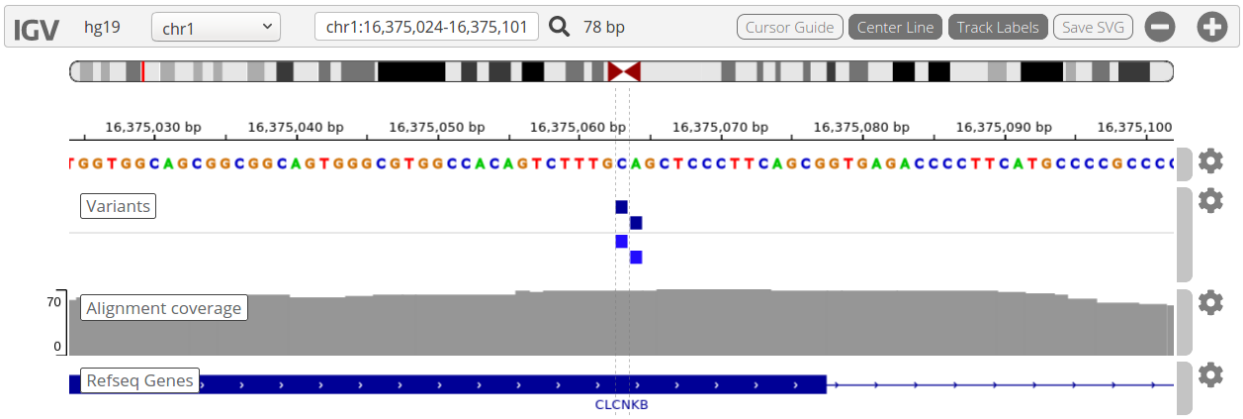

Figure 40. Visualization of variations by IGV tool in the GINO's short variant browser.

Additionally, in the short variant browser section, the most interesting variants for the user can be saved, extracted and displayed in an additional table (Figure 41) from which an automatic report can be obtained with information on these variations. This report is created using LaTex v3.14159265 and provides summary information on all variants of interest, as well as detailed information on each variant in separate sections (Figure 42).

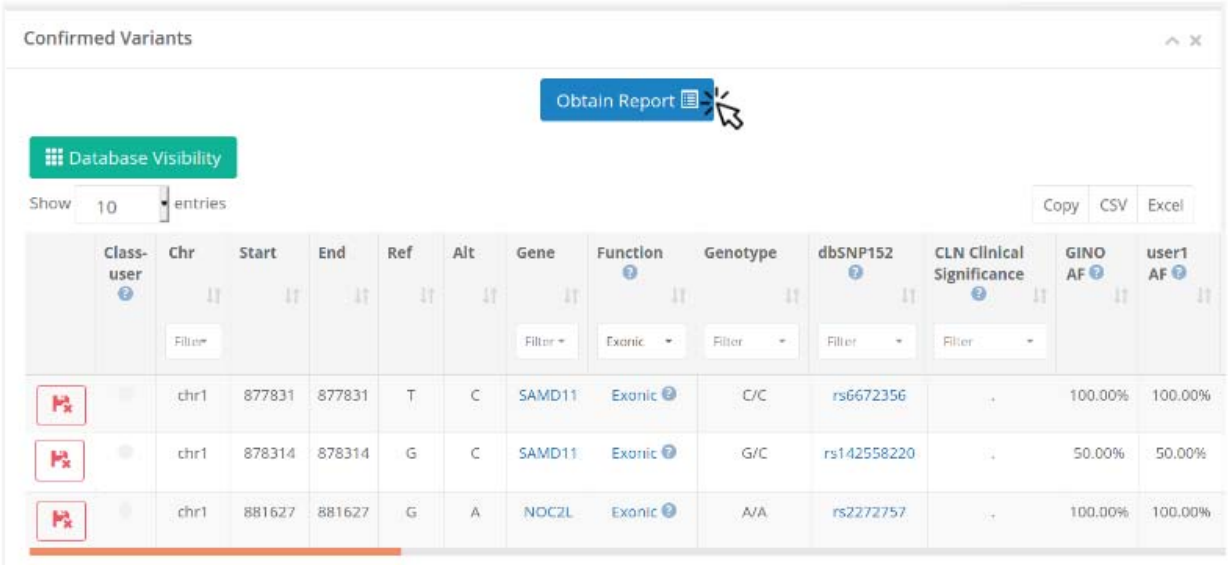

Figure 41. Table with the variations of interest in the GINO's short variant browser.

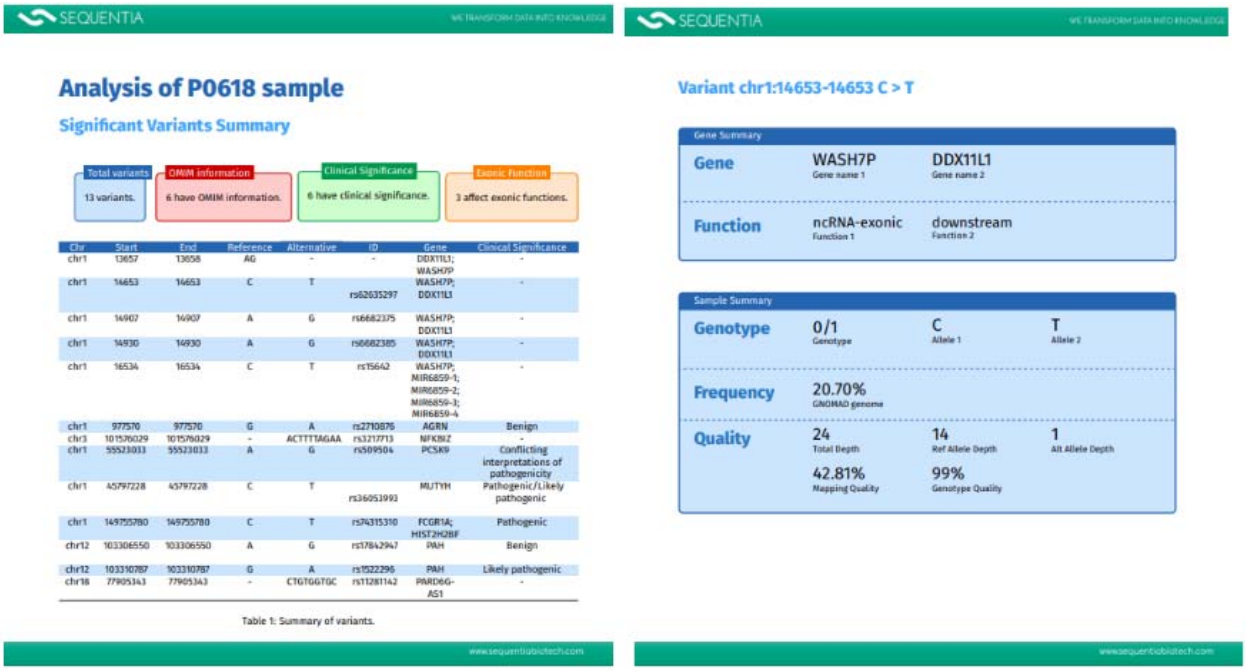

Figure 42. Example of the LaTex automatic report of GINO.

In the parent-child trios displayer, users can visualize all short variants in any of the family members in the same table, allowing easier genotype comparison to verify inheritance (Figure 43).

Furthermore, these variants can be explored through IGV, as in the short variant browser (Figure 40). The parent-child trios displayer can be used to investigate potential disease relevant variants, of which "accumulative" and *de novo* variants are the most interesting. The "accumulative" mutations are present in a heterozygous state in the parents but are homozygous in the child. The *de novo* variants are present in the child but absent in the parents.

Database Visibility | Additional Filters | Search:

Show 10 entries

Copy | CSV | Excel | Print

Family Name	Chromosome	Start	End	Reference	Alternative	dbSNP id	Mother			Father			Child 1		
							Genotype	AD	GQ	Genotype	AD	GQ	Genotype	AD	GQ
F_123_14	chr1	13657	13658	AG	-	rs1263393206	0/1	10,9	99	0/1	13,23	99	0/1	7,15	99
F_123_14	chr1	14907	14907	A	G	rs6682375	0/1	76,154	99	0/1	38,141	99	0/1	59,115	99
F_123_14	chr1	14930	14930	A	G	rs6682385	0/1	80,169	99	0/1	42,159	99	0/1	69,132	99
F_123_14	chr1	14933	14933	G	A	rs199856693	0/1	190,60	99	0/0	41,0	99	0/1	181,25	89
F_123_14	chr1	14976	14976	G	A	rs71252251	0/1	155,55	99	0/0	41,0	99	0/0	157,17	48
F_123_14	chr1	15118	15118	A	G	rs11580262	0/1	6,10	99	0/1	9,8	99	0/1	4,7	99
F_123_14	chr1	15211	15211	T	G	rs3982632	1/1	1,19	34	1/1	1,15	20	1/1	0,9	27
F_123_14	chr1	15274	15274	A	T	rs2758118	1/1	0,15	45	1/1	0,14	42	1/1	0,7	21
F_123_14	chr1	16949	16949	A	C	rs199745162	0/1	15,2	32	0/0	11,0	28	0/0	18,0	54
F_123_14	chr1	17385	17385	G	A	rs201535981	0/1	83,19	99	0/1	72,11	83	0/1	90,18	99

Previous | Next

Figure 43. Screenshot of the table of variants in the GINO's parent-child trios displayer.

The copy number variant viewer focuses on the visualization of CNVs. All CNVs in a sample and their functional annotations are displayed in an interactive table (Figure 44) with the same features and filters as the short variant browser. Here, the criteria to warn about potentially false positive CNVs are the following: (i) a median coverage in the region below 100, (ii) a minimum size less than 150bp, (iii) a precise copy number value greater than 4, or (iv) cover more than three genes. The IGV tool (Figure 40) is also present in the copy number variant viewer to explore the genomic location of the CNVs.

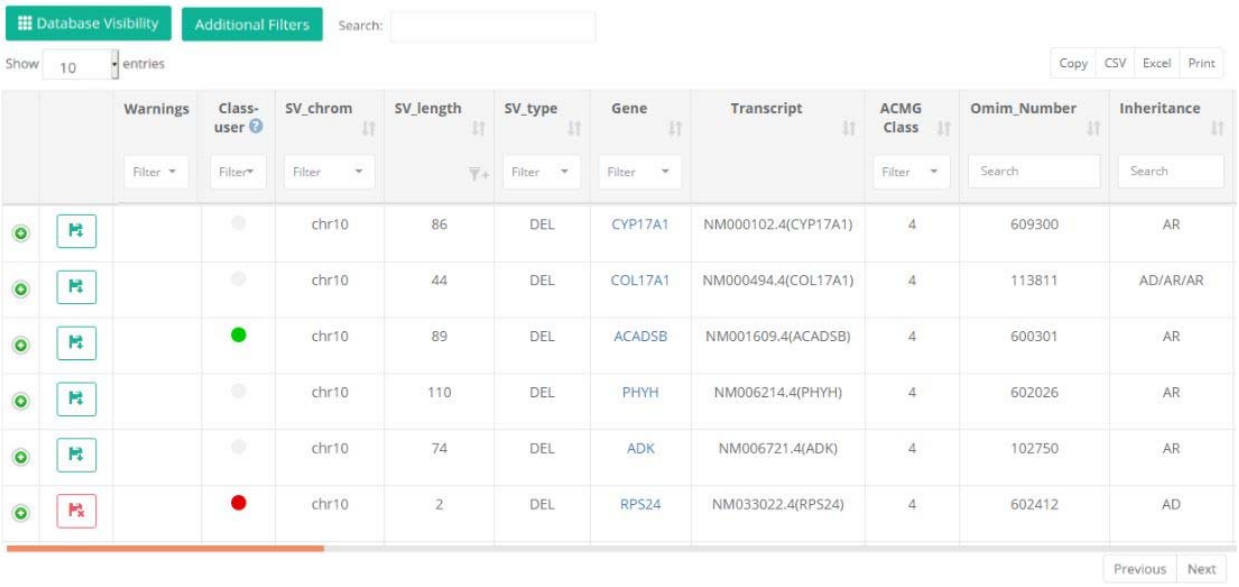

Figure 44. Screenshot of the table of variants in the GINO's copy number variant viewer.

4.1.2 Conclusion

GINO is a comprehensive platform for easy, fast, and affordable interpretation of human variation from NGS data. Its main target is users who want to focus on the interpretation of results without the need to worry about data processing. This data processing includes both the creation of workflows and the selection of algorithms, as well as data management. Using GINO allows to analyze genetic variants obtained with a strong scientific basis, since an exhaustive bibliographic review has been performed to select the best tools for each analysis along with the creation of new pipelines, such as isoCNV. Furthermore, it provides very complete annotations for each variant to facilitate the prioritization of them. It also allows users to store genomic data and interpret it in a simple and accessible way, which is a critical factor when dealing with big data, as it cannot be analyzed without powerful tools to quickly and dynamically store, visualize and filter it.

Chapter 5 | Discussion

Advances in NGS technologies in recent years have made it possible to sequence high-throughput human DNA quickly and affordably. In turn, the sequencing data obtained from NGS could be used for the identification of genomic variants. The detection of variants is important both in modern DNA for the study of its relationship with certain diseases and in ancient DNA to study past human evolution. However, this sequencing data is big data and therefore difficult to process and manage.

In this scenario, it is necessary to improve data accessibility to provide researchers with useful tools that facilitate the efficient obtaining of conclusions. For this reason, in this thesis different genomic tools have been created to automate the identification of short variants and structural variants, genomic data have been integrated with another type of omics data, transcriptomics, to increase the range of identified variants and, finally, a platform to facilitate the interpretation of these variants has been developed (GINO).

5.1 Genomics tools

5.1.1 Short variant discovery

Short variant discovery analyses are complex, they are multi-step processes composed of multiple software applications. Specifically, the entire short variant identification process requires seven different steps to obtain the final SNPs and indels fully annotated and ready for interpretation from the raw sequencing data. To carry out each of these steps, different tools are needed. These tools are often incredibly sophisticated and intricate in their statistical and algorithmic approaches. Although these algorithms do not need to be improved since it is already possible to identify short variants with high performance, it is necessary to invest time and resources in finding the best tools for each of the 7 steps and integrating them satisfactorily to achieve the most optimal results. In addition, given the significant data size and computation time requirements of these analyses, a good computing structure is required to carry them out, such as a server or a computer cluster.

On this basis, we have created different automated pipelines comprised of the best available tools to identify germline short variants in modern DNA in single samples and in parent-child trios for all types of DNA sequencing technology (WGS, WES and TS) as well as to identify short variants in ancient WGS samples. These tools make it possible to automate and democratize the identification of variants since they allow the entire detection process to be carried out from start to finish. They are also modular so that if a new tool or algorithm is developed or improved for any of the seven steps of short variant discovery, it can be easily implemented. It is important to note that these pipelines have been implemented on an internal server but, in the future and outside the context of this thesis, they will be in the cloud for integration in GINO.

It is true that there are already very detailed manuals, which explain the tools to use and how to execute each of the steps of the short variant discovery for modern DNA, such as the GATK guidelines [76] that include all steps except the variant annotation. However, each step must be executed one by one, which requires at least basic bioinformatics knowledge. Recently, GATK has developed FireCloud [227], also called Terra, an online bioinformatics platform to launch analytics in the cloud under user license. Nevertheless, it is not very intuitive and still requires minimal knowledge of scripting and bioinformatics to understand it. Also, the variant annotation step is not yet included. For this reason, our pipelines have been designed to perform the complete analysis until obtaining completely annotated and filtered variants, so the user will only have to worry about uploading the data and interpreting the results.

Regarding the analysis of ancient DNA variants, different protocols are available to carry it out [185], [228], [229], among which PALEOMIX [185] stands out. PALEOMIX is a pipeline to perform variant calling from ancient WGS data. However, its main drawback is that it requires intricate command-line instructions to operate and set analysis parameters, hampering its use by non-bioinformaticians. For this reason, we have created a fully automatic pipeline with the ancestral genome as the only input and also with some improvements compared to PALEOMIX. For example, we use Sentieon [115], the accelerated version of GATK, to perform the variant discovery instead of SAMtools [73], allowing us to improve both the speed and performance of the variant calling. In addition, we used GAIA v2.02 [186] for the microbial profile, which has proven to be one of the best performing tools for identifying microbial populations [186].

In brief, the creation of all these genomic tools has made it possible to improve data processing in terms of automation, thus bringing these analyses closer to non-bioinformaticians to potentially unlock new insights on important biomedical issues.

5.1.2 Structural variant discovery

Structural variant discovery follows the same seven steps as short variant discovery, but using different algorithms for the last three steps: SV discovery, filtering and annotation. Multiple tools have been developed for the discovery of germline structural variants from WGS data, each based on different approaches [134]–[137], [140], [141], [149]–[152]. In addition, several benchmarks have been performed to analyze these tools and they have been shown to detect structural variants from WGS data with high accuracy and performance [153], [154]. However, most of these algorithms start from alignment data rather than raw sequencing data and do not perform variant annotation. Thereby, we have implemented an automatic pipeline to perform the entire process, from raw sequence data to annotated variants. Annotating these variants against many annotation sources is extremely important as another challenge with these analyses lies in prioritizing variants to find meaningful results. Moreover, this pipeline has the same features

as those implemented for the short variant discovery pipelines, it is simple for ease of application, modular to easily implement updates, and scientifically robust.

There are also multiple tools for CNV discovery from WES or TS data [146], [147]. However, these tools can still be improved as CNV detection using WES or TS data is much more challenging than using WGS data. This is due to the fact that in the WES and TS data the breakpoint is sometimes not in the capture regions and due to the inefficiencies of the technique [44]–[46]. Among the different tools available is DECoN [148], which has shown high performance in the identification of CNV from the NGS panel data [156], [157]. Nevertheless, DECoN starts from aligned data rather than from raw sequence data and its parameters must be optimized for the highest possible detection sensitivity. There is already a tool to perform the parameter optimization, optimizer from the CNVbenchmarkeR framework [157]. One of the optimizer inputs are the validated CNVs with which to compare the results obtained by DECoN to perform parameter optimization. These validated CNVs are obtained from orthogonal techniques, such as MLPA or aCGH [196], which increase the cost and time of the analyses. Hence, we have developed isoCNV, a pipeline to optimize the parameters of DECoN using only sequencing data. Its effectiveness in increasing the sensitivity of DECoN has been demonstrated with a real dataset of WES samples and another of TS samples. This pipeline, like the previous ones, is automatic and allows to obtain analysis-ready CNVs from the raw sequencing data of a batch of samples. In addition to facilitating analysis due to its automatic nature, it reduces analysis time and cost as no orthogonal methods are required.

In this way, we have implemented workflows to analyze all types of germline variants with high throughput from any type of DNA sequencing data (WGS, WES or TS). All of these automated and modular workflows enable standardization, consistency, and reproducibility, which is critical in the clinical settings. In addition, they are the beginning of a democratization of the processing of sequencing data for non-bioinformaticians, which will culminate with their implementation in the cloud outside the context of this thesis, thus giving even greater accessibility and simplicity.

5.2 Data integration

The integration of genomic data with transcriptomic data offers new opportunities for variant discovery. These new opportunities range from improving the accuracy of variant identification in highly expressed genes to finding new variants only detectable from RNA-seq data, such as RNA-editing variants or ASE variants, as well as offering an additional method of variant validation. There are already tools that integrate these two types of data, such as RADIA [171] or VaDiR [172]. However, these tools are based on the detection of somatic short variants and do not detect germline short variants. The main objective of RADIA is to use the RNA-seq data to validate the somatic variants found in DNA [171]. All somatic variants called by RADIA are supported by DNA, and RNA-only variants are not called. In addition, it does not include

the detection of RNA-editing variants or ASE variants, so part of the benefits of using RNA-seq are not being exploited. VaDiR calls somatic variants only using RNA-seq data, DNA data is only used to filter germline variants [172]. Because of this, VaDiR misses mainly low-frequency RNA-seq variants and some of the potential of using both types of data is not exploited.

Taking into account all the above, varRED has been developed to identify germline short variants by integrating the variants identified by both DNA and RNA-seq, thus making the most of these two omics. To our knowledge, there is no other pipeline to date that allows the identification of germline variants by integrating both DNA and RNA-seq data. Specifically, we have focused on integrating RNA-seq with WES data to offer a cost-efficient alternative to using WGS data, as WES data is currently preferred in both clinical and research use [5], [42], [43] due to its lower price. By integrating RNA-seq data into WES data with varRED, more variants (RNA-only variants and RNA-rescue variants) can be identified without a great effect on performance and the variants identified by both types of data (Strong-evidence variants) can be validated. However, it is important to note that the RNA-editing variants and ASE variants identified by varRED could not be confirmed or evaluated due to the lack of availability of validated variants from RNA-seq. Finally, an additional benefit of this approach is that RNA-seq data can be used for further analyses on gene expression levels or gene fusions.

In the same way as the rest of the pipelines implemented in this thesis, varRED has been developed to be completely automatic and modular, thus facilitating its use and giving consistency and reproducibility to the analyses. This automation of variant discovery helps optimize time, improve productivity, and maintain infrastructure.

5.3 Genomics data analysis platform

With the increased use of sequencing and variant discovery, challenges have arisen around the ease of use of software, data management and reproducibility of results. Some of these challenges have been solved with the creation of the genomic tools that have already been discussed in the previous sections. Data storage is also another challenge because datasets can easily reach terabyte sizes per sequencing run. Furthermore, it is not only complex to store the sequencing data, but also to collect metadata and provide data to end users. Therefore, in this thesis a powerful database has been developed to store all this information and avoid users having to worry about data storage or the computational resources that this entails.

Another important challenge of these analyses lies in finding meaning and significance in the results of the vast variants identified. There is a need to improve simplicity in the prioritization and interpretation of results so that clinicians and researchers with no knowledge of bioinformatics can easily perform these analyses. Multiple platforms have been developed to facilitate the interpretation of results for non-

bioinformaticians, among them VarSome [169] stands out. Although VarSome is very comprehensive, it can still improve both in terms of the variant discovery algorithms, especially for structural variants, and in usability and simplicity. For example, VarSome applies ExomeDepth [201] to identify CNVs from WES or TS data when a better option might be to use DECoN [148] with optimized parameters.

In this thesis we have developed GINO, a platform for the visualization and interpretation of variants obtained using the best algorithms and software available. In GINO, users are not required to have computer or bioinformatics knowledge because it has been designed to be as user-friendly as possible. In addition to being simple, it shows the most complete information possible since the variants have been annotated using a large number of databases. These annotations are very important when prioritizing variants to easily find those that may affect the patient's health. In addition, GINO has a very important feature, the GINO allele frequency, which allows users to obtain the frequency of a variant in all their samples. This GINO allele frequency provides added value, especially in population genetic analyses and in association mapping.

GINO makes it possible to bring variant analysis closer to institutions without bioinformatics facilities or with little experience in bioinformatics and gives reproducibility and consistency to the analyses. Currently GINO is a platform for storage and visualization of variants but in the future, and outside the context of this thesis, all analyses will be carried out in the cloud within the platform.

Chapter 6 | Conclusions and Future Research

6.1 Conclusions

The identification of human germline variation is important given its implication in multiple diseases [25], [119], [120]. These analyses are already implemented in clinical care to improve both the prognosis and the diagnosis of several diseases in modern genomes [8]–[11]. Furthermore, the detection of variants in ancient human genomes is applied to gain insight into past human evolution [12]–[16]. However, the large number of tools available to perform these analyses, the different algorithms required by different types of variations and data, and the challenge of managing big data hinder its implementation. To solve these issues, different workflows based on the best available algorithms have been implemented in this thesis to identify all types of human germline variation from NGS data (Chapter 2 and 3). In addition, GINO, a platform to store and visualize these variants, has been developed to avoid users having to worry about data management when interpreting the results (Chapter 4).

Regarding the development of workflows, different pipelines have been implemented for the identification of short variants (SNPs and indels) and structural variants (CNVs and chromosomal rearrangements events) from data sequenced using different NGS strategies. A pipeline for identifying short variants in modern WGS, WES or TS data (Section 2.1.1) and a pipeline for identifying short variants from ancient WGS data (Section 2.1.2) have been developed. Both pipelines are based on existing tools but, to date, a complete and automated pipeline has not been built with them to obtain fully annotated, filtered, and ready-for-interpretation variants from the raw sequencing data. Additionally, the pipeline for the discovery of short variants in ancient genomes has been applied to a human mandible dated between 16980-16510 calibrated years before the present, allowing to separate between endogenous ancient DNA and modern DNA contamination. Although the identified variants could not be interpreted reliably due to the low coverage of the sample (0.28 x), the subsequent analyses carried out by our collaborating group using the isolated endogenous aDNA, allowed us to find new knowledge about the migration of ancient populations [184]. This work has been published in the journal Current Biology with the doctoral student as a co-author [184].

Furthermore, a workflow for structural variant discovery in modern WGS data (Section 2.2.1) and another workflow for *in silico* optimization of CNV detection using modern WES or TS data (Section 2.2.2) have been implemented. The pipeline for structural variant discovery from WGS data builds on existing tools that have shown high performance but have never been automated in a single pipeline. The *in silico* optimization pipeline for CNV detection from WES or TS data (isoCNV) is also based on existing tools, but these have been combined in a novel way. The isoCNV pipeline has shown to increase the sensitivity of CNV detection using only NGS data rather than orthogonal methods, such has MLPA or aCGH. An

article on this work has been submitted to the journal BMC bioinformatics with the doctoral student as first author and is currently under review.

The last pipeline implemented in this doctoral thesis is varRED (Chapter 3). The varRED pipeline relies on the integration of WES and RNA-seq data to increase the number of identified germline variants (such as ASE or RNA-editing variants) over those identified if only WES data were used. Moreover, this combined strategy is more cost-effective than the use of whole genome sequencing, since the availability of RNA-seq data allows for additional analyses, such as the identification of gene fusions. To our knowledge, there is no other pipeline to date that performs this type of integration.

Finally, this thesis has concluded with the development of GINO, a platform under license for use (Chapter 4). GINO is a powerful platform to store and interpret the SNPs, indels and CNVs obtained through the workflows developed in the context of this thesis. This platform solves two major obstacles in the variant discovery analysis, the informatic issue (data storage and management) and the scientific issue (data interpretation), given that it allows users to store fully annotated genetic variants in a database and visualize them in an easy-to-use graphical interface. It is important to note that at the moment this online platform only allows the visualization of results. Now, variant discovery is performed on an internal server, but in the future the entire process will run in the cloud.

This thesis has laid the foundations to create an online platform for the analysis and visualization of genomic data in the cloud in the near future. The workflows for the identification of variants and the visualization tool are now available. The last remaining part of the development process is the implementation of the workflows in the cloud and their union with the visualization platform to obtain totally online and automatic data processing.

6.2 Future work

Future work will mainly focus on the cloud implementation of the workflows developed in the context of this thesis and their integration into the GINO visualization platform. A cloud-based architecture will provide us with multiple benefits such as vertical and horizontal scalability. By properly enforcing scalability policies, the system will be able to respond to workload changes and thus provide a seamless user experience at all times.

Another key point of our cloud computing architecture will be reliability. The system will be designed to be fault tolerant: it will be able to retrieve and complete user analysis even if cloud resources are not available. In addition, a data loading system will be developed within the browser and without external dependencies, with integrity checks and resumable functions to allow users to upload large files reliably and quickly.

It is important to highlight that the pipelines already implemented in this thesis will continue to be reviewed, optimized and updated to be up to date and have the best variant discovery tools at all times. The updating of the pipelines will be possible and facilitated by the modular nature of the workflows already implemented, which will allow modifying only a part of the workflows without affecting the rest of the process. These continuous updates are important due to the constant evolution and improvement of the genomics field, which is possible thanks to the large investment of capital and resources in this field nowadays.

References

[1] R. Boland C, ‘Non-coding RNA: It’s Not Junk’, *Dig. Dis. Sci.*, vol. 62, no. 5, pp. 1107–1109, 2017, doi: 10.1007/s10620-017-4506-1.

[2] R. Nussbaum, R. McInnes, and H. Willard, *Thompson & Thompson Genetics in Medicine*, 8th ed. Elsevier, 2015.

[3] D. F. Gudbjartsson *et al.*, ‘Large-scale whole-genome sequencing of the Icelandic population’, *Nat. Genet.*, vol. 47, no. 5, pp. 435–444, 2015, doi: 10.1038/ng.3247.

[4] E. S. Lander, ‘Initial impact of the sequencing of the human genome’, *Nature*, vol. 470, no. 7333, pp. 187–197, 2011, doi: 10.1038/nature09792.

[5] F. E. Dewey *et al.*, ‘Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study’, *Science (80-.).*, vol. 354, no. 6319, Dec. 2016, doi: 10.1126/science.aaf6814.

[6] A. Auton *et al.*, ‘A global reference for human genetic variation’, *Nature*, vol. 526, no. 7571, pp. 68–74, Sep. 2015, doi: 10.1038/nature15393.

[7] C. Manzoni *et al.*, ‘Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences’, *Brief. Bioinform.*, vol. 19, no. 2, pp. 286–302, 2018, doi: 10.1093/BIB/BBW114.

[8] D. Carbonell *et al.*, ‘Next-generation sequencing improves diagnosis, prognosis and clinical management of myeloid neoplasms’, *Cancers (Basel).*, vol. 11, no. 9, Sep. 2019, doi: 10.3390/cancers11091364.

[9] H. Y. Liu *et al.*, ‘Diagnostic and clinical utility of whole genome sequencing in a cohort of undiagnosed Chinese families with rare diseases’, *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019, doi: 10.1038/s41598-019-55832-1.

[10] M. Vinkšel, K. Writzl, A. Maver, and B. Peterlin, ‘Improving diagnostics of rare genetic diseases with NGS approaches’, *J. Community Genet.*, vol. 12, no. 2, pp. 247–256, Jan. 2021, doi: 10.1007/s12687-020-00500-5.

[11] Z. Liu, L. Zhu, R. Roberts, and W. Tong, ‘Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We?’, *Trends in Genetics*, vol. 35, no. 11. Elsevier Ltd, pp. 852–867, Nov. 01, 2019, doi: 10.1016/j.tig.2019.08.006.

[12] I. Lazaridis *et al.*, ‘Ancient human genomes suggest three ancestral populations for present-day Europeans’, *Nature*, vol. 513, no. 7518, pp. 409–413, Sep. 2014, doi: 10.1038/nature13673.

[13] T. Günther and M. Jakobsson, ‘Genes mirror migrations and cultures in prehistoric Europe — a

population genomic perspective', *Current Opinion in Genetics and Development*, vol. 41. Elsevier Ltd, pp. 115–123, Dec. 01, 2016, doi: 10.1016/j.gde.2016.09.004.

[14] P. Skoglund and I. Mathieson, 'Ancient genomics of modern humans: The first decade', *Annual Review of Genomics and Human Genetics*, vol. 19. Annual Reviews Inc., pp. 381–404, Aug. 31, 2018, doi: 10.1146/annurev-genom-083117-021749.

[15] D. Toncheva, D. Serbezov, S. Karachanak-Yankova, and D. Nesheva, 'Ancient mitochondrial DNA pathogenic variants putatively associated with mitochondrial disease', *PLoS One*, vol. 15, no. 9 September, p. e0233666, Sep. 2020, doi: 10.1371/journal.pone.0233666.

[16] G. Kerner *et al.*, 'Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years', 2021, doi: 10.1016/j.ajhg.2021.02.009.

[17] M. Escala-Garcia *et al.*, 'Genome-wide association study of germline variants and breast cancer-specific mortality', *Br. J. Cancer*, vol. 120, no. 6, pp. 647–657, 2019, doi: 10.1038/s41416-019-0393-x.

[18] M. Cargill *et al.*, 'Characterization of single-nucleotide polymorphisms in coding regions of human genes', *Nat. Genet.*, vol. 22, no. 3, pp. 231–238, 1999, doi: 10.1038/10290.

[19] A. Gusev *et al.*, 'Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases', *Am. J. Hum. Genet.*, vol. 95, no. 5, pp. 535–552, 2014, doi: 10.1016/j.ajhg.2014.10.004.

[20] D. H. Spencer, B. Zhang, and J. Pfeifer, 'Chapter 8 - Single Nucleotide Variant Detection Using Next Generation Sequencing', in *Clinical Genomics*, S. Kulkarni and J. B. T.-C. G. Pfeifer, Eds. Boston: Academic Press, 2015, pp. 109–127.

[21] C. Gonzaga-Jauregui, J. R. Lupski, and R. A. Gibbs, 'Human genome sequencing in health and disease', *Annual Review of Medicine*, vol. 63. Annual Reviews , pp. 35–61, Jan. 16, 2012, doi: 10.1146/annurev-med-051010-162644.

[22] J. K. Sehn, 'Insertions and Deletions (Indels)', in *Clinical Genomics*, Elsevier Inc., 2015, pp. 129–150.

[23] J. L. Freeman *et al.*, 'Copy number variation: New insights in genome diversity', *Genome Research*, vol. 16, no. 8. Genome Res, pp. 949–961, 2006, doi: 10.1101/gr.3677206.

[24] S. Girirajan, C. D. Campbell, and E. E. Eichler, 'Human copy number variation and complex genetic disease', *Annu. Rev. Genet.*, vol. 45, pp. 203–226, 2011, doi: 10.1146/annurev-genet-102209-163544.

[25] C. Aouiche, X. Shang, and B. Chen, 'Copy number variation related disease genes', *Quantitative Biology*, vol. 6, no. 2. Higher Education Press, pp. 99–112, Jun. 01, 2018, doi: 10.1007/s40484-

018-0137-6.

[26] J. D. Rowley, ‘A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining’, *Nature*, vol. 243, no. 5405, pp. 290–293, 1973, doi: 10.1038/243290a0.

[27] J. Lejeune, M. Gautier, and R. Turpin, ‘“Etude des chromosomes somatiques de neuf enfants mongoliens” [Study of somatic chromosomes from 9 mongoloid children]’, *C. R. Hebd. Seances Acad. Sci.*, vol. 248, pp. 1721--1722, 1959.

[28] E. Gasperskaja and V. Kučinskas, ‘The most common technologies and tools for functional genome analysis’, *Acta medica Litu.*, vol. 24, no. 1, pp. 1–11, Apr. 2017, doi: 10.6001/actamedica.v24i1.3457.

[29] A. M. Maxam and W. Gilbert, ‘A new method for sequencing DNA’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 2, pp. 560–564, 1977, doi: 10.1073/pnas.74.2.560.

[30] F. Sanger and A. R. Coulson, ‘A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase’, *J. Mol. Biol.*, vol. 94, no. 3, pp. 441–448, May 1975, doi: 10.1016/0022-2836(75)90213-2.

[31] E. S. Lander *et al.*, ‘Initial sequencing and analysis of the human genome’, *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[32] E. R. Mardis, ‘Next-generation DNA sequencing methods’, *Annual Review of Genomics and Human Genetics*, vol. 9. Annu Rev Genomics Hum Genet, pp. 387–402, 2008, doi: 10.1146/annurev.genom.9.081307.164359.

[33] E. R. Mardis, ‘A decade’s perspective on DNA sequencing technology’, *Nature*, vol. 470, no. 7333. Nature, pp. 198–203, Feb. 10, 2011, doi: 10.1038/nature09796.

[34] M. L. Metzker, ‘Sequencing technologies the next generation’, *Nature Reviews Genetics*, vol. 11, no. 1. Nature Publishing Group, pp. 31–46, Jan. 08, 2010, doi: 10.1038/nrg2626.

[35] J. F. Thompson and P. M. Milos, ‘The properties and applications of single-molecule DNA sequencing’, *Genome Biology*, vol. 12, no. 2. BioMed Central, p. 217, Feb. 24, 2011, doi: 10.1186/gb-2011-12-2-217.

[36] M. Margulies *et al.*, ‘Genome sequencing in microfabricated high-density picolitre reactors’, *Nature*, vol. 437, no. 7057, pp. 376–380, Sep. 2005, doi: 10.1038/nature03959.

[37] D. M. Altshuler *et al.*, ‘An integrated map of genetic variation from 1,092 human genomes’, *Nature*, vol. 491, no. 7422, pp. 56–65, Nov. 2012, doi: 10.1038/nature11632.

[38] C. S. Ku and D. H. Roukos, ‘From next-generation sequencing to nanopore sequencing technology: Paving the way to personalized genomic medicine’, *Expert Review of Medical*

Devices, vol. 10, no. 1. Expert Rev Med Devices, pp. 1–6, Jan. 2013, doi: 10.1586/erd.12.63.

[39] M. Mignardi and M. Nilsson, 'Fourth-generation sequencing in the cell and the clinic', *Genome Med.*, vol. 6, no. 4, p. 31, Apr. 2014, doi: 10.1186/gm548.

[40] M. Seleman, R. Hoyos-Bachiloglu, R. S. Geha, and J. Chou, 'Uses of next-generation sequencing technologies for the diagnosis of primary immunodeficiencies', *Frontiers in Immunology*, vol. 8, no. JUL. Frontiers Media S.A., Jul. 24, 2017, doi: 10.3389/fimmu.2017.00847.

[41] A. Warr, C. Robert, D. Hume, A. Archibald, N. Deeb, and M. Watson, 'Exome sequencing: Current and future perspectives', *G3 Genes, Genomes, Genet.*, vol. 5, no. 8, pp. 1543–1550, 2015, doi: 10.1534/g3.115.018564.

[42] K. Walter *et al.*, 'The UK10K project identifies rare variants in health and disease', *Nature*, vol. 526, no. 7571, pp. 82–89, Oct. 2015, doi: 10.1038/nature14962.

[43] Y. Yang *et al.*, 'Molecular findings among patients referred for clinical whole-exome sequencing', *JAMA - J. Am. Med. Assoc.*, vol. 312, no. 18, pp. 1870–1879, Nov. 2014, doi: 10.1001/jama.2014.14601.

[44] L. Kadalayil *et al.*, 'Exome sequence read depth methods for identifying copy number changes', *Brief. Bioinform.*, vol. 16, no. 3, pp. 380–392, Aug. 2014, doi: 10.1093/bib/bbu027.

[45] N. Krumm *et al.*, 'Copy number variation detection and genotyping from exome sequence data', *Genome Res.*, vol. 22, no. 8, pp. 1525–1532, Aug. 2012, doi: 10.1101/gr.138115.112.

[46] J. M. Kebschull and A. M. Zador, 'Sources of PCR-induced distortions in high-throughput sequencing data sets', *Nucleic Acids Res.*, vol. 43, no. 21, Dec. 2015, doi: 10.1093/nar/gkv717.

[47] C. Xu, 'A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data', *Computational and Structural Biotechnology Journal*, vol. 16. Elsevier B.V., pp. 15–24, 2018, doi: 10.1016/j.csbj.2018.01.003.

[48] Grand View Research, 'Genomics Market Size, Share & Trends Analysis Report by Application and Technology (Functional Genomics, Pathway Analysis), by Deliverables (Products, Services), by End Use, by Region, and Segment Forecasts, 2020 - 2027', 2020.

[49] S. Pabinger *et al.*, 'A survey of tools for variant analysis of next-generation genome sequencing data', *Brief. Bioinform.*, vol. 15, no. 2, pp. 256–278, 2014, doi: 10.1093/bib/bbs086.

[50] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, 'Genotype and SNP calling from next-generation sequencing data', *Nature Reviews Genetics*, vol. 12, no. 6. Nature Publishing Group, pp. 443–451, Jun. 18, 2011, doi: 10.1038/nrg2986.

[51] A. Cornish and C. Guda, 'A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference', *Biomed Res. Int.*, vol. 2015, 2015, doi: 10.1155/2015/456479.

[52] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, Dec. 2009, doi: 10.1093/nar/gkp1137.

[53] B. Ewing, L. D. Hillier, M. C. Wendl, and P. Green, 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment', *Genome Res.*, vol. 8, no. 3, pp. 175–185, Mar. 1998, doi: 10.1101/gr.8.3.175.

[54] A. Conesa *et al.*, 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, vol. 17, no. 1. BioMed Central Ltd., pp. 1–19, Jan. 26, 2016, doi: 10.1186/s13059-016-0881-8.

[55] S. Andrews, 'FastQC: A Quality Control Tool for High Throughput Sequence Data'. 2014, [Online]. Available: www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[56] M. Shirley, 'fastqp: Simple FASTQ quality assessment using Python'. 2014, [Online]. Available: https://github.com/mdshw5/fastqp.

[57] V. A. Cantu, J. Sadural, and R. Edwards, 'PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets', *PeerJ Prepr.*, 2019, doi: 10.7287/peerj.preprints.27553v1.

[58] A. M. Bolger, M. Lohse, and B. Usadel, 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[59] M. Martin, 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.

[60] A. Criscuolo and S. Brisse, 'AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads', *Genomics*, vol. 102, no. 5–6, pp. 500–506, Nov. 2013, doi: 10.1016/j.ygeno.2013.07.011.

[61] H. Jiang, R. Lei, S. W. Ding, and S. Zhu, 'Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads', *BMC Bioinformatics*, vol. 15, no. 1, p. 182, Jun. 2014, doi: 10.1186/1471-2105-15-182.

[62] B. Bushnell, 'BBTools'. 2014, [Online]. Available: sourceforge.net/projects/bbmap/.

[63] M. Schubert, S. Lindgreen, and L. Orlando, 'AdapterRemoval v2: Rapid adapter trimming, identification, and read merging', *BMC Res. Notes*, vol. 9, no. 1, p. 88, Feb. 2016, doi: 10.1186/s13104-016-1900-2.

[64] H. Li, 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', 2013. Accessed: May 01, 2021. [Online]. Available: http://github.com/lh3/bwa.

[65] S. Thankaswamy-Kosalai, P. Sen, and I. Nookaew, 'Evaluation and assessment of read-mapping

by multiple next-generation sequencing aligners based on genome-wide characteristics', *Genomics*, vol. 109, no. 3–4, pp. 186–191, Jul. 2017, doi: 10.1016/j.ygeno.2017.03.001.

[66] T. J. Treangen and S. L. Salzberg, 'Repetitive DNA and next-generation sequencing: Computational challenges and solutions', *Nature Reviews Genetics*, vol. 13, no. 1. Nature Publishing Group, pp. 36–46, Jan. 29, 2012, doi: 10.1038/nrg3117.

[67] H. Fang *et al.*, 'Reducing INDEL calling errors in whole genome and exome sequencing data', *Genome Med.*, vol. 6, no. 10, p. 89, Dec. 2014, doi: 10.1186/s13073-014-0089-z.

[68] G. Narzisi *et al.*, 'Accurate de novo and transmitted indel detection in exome-capture data using microassembly', *Nat. Methods*, vol. 11, no. 10, pp. 1033–1036, Jan. 2014, doi: 10.1038/nmeth.3069.

[69] H. Li and R. Durbin, 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[70] B. Langmead and S. L. Salzberg, 'Fast gapped-read alignment with Bowtie 2', *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[71] 'Novoalign'. [Online]. Available: www.novocraft.com.

[72] G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin, 'MUMmer4: A fast and versatile genome alignment system', *PLOS Comput. Biol.*, vol. 14, no. 1, p. e1005944, Jan. 2018, doi: 10.1371/journal.pcbi.1005944.

[73] H. Li *et al.*, 'The Sequence Alignment/Map format and SAMtools', vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[74] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, 'Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration', *Brief. Bioinform.*, vol. 14, no. 2, pp. 178–192, Mar. 2013, doi: 10.1093/bib/bbs017.

[75] 'Picard toolkit'. Broad Institute, 2019, [Online]. Available: http://broadinstitute.github.io/picard/.

[76] G. A. Van der Auwera *et al.*, 'From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline', *Curr. Protoc. Bioinforma.*, vol. 11, no. SUPL.43, p. 11.10.1, 2013, doi: 10.1002/0471250953.bi1110s43.

[77] V. Bansal, 'A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments', *BMC Bioinformatics*, vol. 18, no. S3, p. 43, Mar. 2017, doi: 10.1186/s12859-017-1471-9.

[78] C. Der Sarkissian *et al.*, 'Shotgun microbial profiling of fossil remains', *Mol. Ecol.*, vol. 23, no. 7, pp. 1780–1798, Apr. 2014, doi: 10.1111/mec.12690.

[79] M. L. Sampietro *et al.*, ‘Tracking down human contamination in ancient human teeth’, *Mol. Biol. Evol.*, vol. 23, no. 9, pp. 1801–1807, Sep. 2006, doi: 10.1093/molbev/msl047.

[80] R. E. Green *et al.*, ‘Analysis of one million base pairs of Neanderthal DNA’, *Nature*, vol. 444, no. 7117, pp. 330–336, Nov. 2006, doi: 10.1038/nature05336.

[81] J. Krause *et al.*, ‘A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia’, *Curr. Biol.*, vol. 20, no. 3, pp. 231–236, Feb. 2010, doi: 10.1016/j.cub.2009.11.068.

[82] B. Llamas, G. Valverde, L. Fehren-Schmitz, L. S. Weyrich, A. Cooper, and W. Haak, ‘From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era’, *Science and Technology of Archaeological Research*, vol. 3, no. 1. Routledge, pp. 1–14, Jan. 01, 2017, doi: 10.1080/20548923.2016.1258824.

[83] A. W. Briggs *et al.*, ‘Patterns of damage in genomic DNA sequences from a Neandertal’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 37, pp. 14616–14621, Sep. 2007, doi: 10.1073/pnas.0704665104.

[84] M. García-Garcerà *et al.*, ‘Fragmentation of contaminant and endogenous dna in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics’, *PLoS One*, vol. 6, no. 8, 2011, doi: 10.1371/journal.pone.0024161.

[85] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, and S. Pääbo, ‘Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA’, *PLoS One*, vol. 7, no. 3, Mar. 2012, doi: 10.1371/journal.pone.0034131.

[86] M. Meyer *et al.*, ‘A high-coverage genome sequence from an archaic Denisovan individual’, *Science (80-.).*, vol. 338, no. 6104, pp. 222–226, Oct. 2012, doi: 10.1126/science.1224344.

[87] H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, and L. Orlando, ‘MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters’, in *Bioinformatics*, Jul. 2013, vol. 29, no. 13, pp. 1682–1684, doi: 10.1093/bioinformatics/btt193.

[88] J. Neukamm, A. Peltzer, and K. Nieselt, ‘DamageProfiler: Fast damage pattern calculation for ancient DNA’, doi: 10.1101/2020.10.01.322206.

[89] P. Skoglund *et al.*, ‘Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 6, pp. 2229–2234, Feb. 2014, doi: 10.1073/pnas.1318934111.

[90] G. Renaud, V. Slon, A. T. Duggan, and J. Kelso, ‘Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA’, *Genome Biol.*, vol. 16, no. 1, p. 224, Oct. 2015, doi: 10.1186/s13059-015-0776-0.

[91] N. Nakatsuka, É. Harney, S. Mallick, M. Mah, N. Patterson, and D. Reich, ‘ContamLD: Estimation

of ancient nuclear DNA contamination using breakdown of linkage disequilibrium', *Genome Biol.*, vol. 21, no. 1, pp. 1–22, Aug. 2020, doi: 10.1186/s13059-020-02111-2.

[92] S. Peyrégne and B. M. Peter, 'AuthentiCT: A model of ancient DNA damage to estimate the proportion of present-day DNA contamination', *Genome Biol.*, vol. 21, no. 1, pp. 1–16, Sep. 2020, doi: 10.1186/s13059-020-02123-y.

[93] Y. Lammers, P. D. Heintzman, and I. G. Alsos, 'Environmental palaeogenomic reconstruction of an Ice Age algal population', *Commun. Biol.*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s42003-021-01710-4.

[94] W. Xu *et al.*, 'An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole-genome sequencing data', *Ecol. Evol.*, vol. 11, no. 1, pp. 390–401, Jan. 2021, doi: 10.1002/ece3.7056.

[95] M. Lipson *et al.*, 'Three Phases of Ancient Migration Shaped the Ancestry of Human Populations in Vanuatu', *Curr. Biol.*, vol. 30, no. 24, pp. 4846-4856.e6, Dec. 2020, doi: 10.1016/j.cub.2020.09.035.

[96] V. Shinde *et al.*, 'An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers', *Cell*, vol. 179, no. 3, pp. 729-735.e10, Oct. 2019, doi: 10.1016/j.cell.2019.08.048.

[97] C. Posth *et al.*, 'Reconstructing the Deep Population History of Central and South America', *Cell*, vol. 175, no. 5, pp. 1185-1197.e22, Nov. 2018, doi: 10.1016/j.cell.2018.10.027.

[98] E. Garrison and G. Marth, 'Haplotype-based variant detection from short-read sequencing', Jul. 2012, Accessed: May 03, 2021. [Online]. Available: https://arxiv.org/abs/1207.3907.

[99] R. Poplin *et al.*, 'A universal snp and small-indel variant caller using deep neural networks', *Nat. Biotechnol.*, vol. 36, no. 10, p. 983, Nov. 2018, doi: 10.1038/nbt.4235.

[100] M. A. Depristo *et al.*, 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nat. Genet.*, vol. 43, no. 5, pp. 491–501, May 2011, doi: 10.1038/ng.806.

[101] D. C. Koboldt, 'Best practices for variant calling in clinical sequencing', *Genome Medicine*, vol. 12, no. 1. BioMed Central Ltd, Dec. 01, 2020, doi: 10.1186/s13073-020-00791-w.

[102] A. McKenna *et al.*, 'The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, doi: 10.1101/gr.107524.110.

[103] A. Rimmer *et al.*, 'Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications', *Nat. Genet.*, vol. 46, no. 8, pp. 912–918, 2014, doi: 10.1038/ng.3036.

[104] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, ‘Dindel: Accurate indel calls from short-read data’, *Genome Res.*, vol. 21, no. 6, pp. 961–973, Jun. 2011, doi: 10.1101/gr.112326.110.

[105] A. Wilm *et al.*, ‘LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets’, *Nucleic Acids Res.*, vol. 40, no. 22, pp. 11189–11201, Dec. 2012, doi: 10.1093/nar/gks918.

[106] Z. Wei, W. Wang, P. Hu, G. J. Lyon, and H. Hakonarson, ‘SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data’, *Nucleic Acids Res.*, vol. 39, no. 19, p. e132, Oct. 2011, doi: 10.1093/nar/gkr599.

[107] D. C. Koboldt *et al.*, ‘VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing’, *Genome Res.*, vol. 22, no. 3, pp. 568–576, Mar. 2012, doi: 10.1101/gr.129684.111.

[108] X. Liu, S. Han, Z. Wang, J. Gelernter, and B. Z. Yang, ‘Variant Callers for Next-Generation Sequencing Data: A Comparison Study’, *PLoS One*, vol. 8, no. 9, p. 75619, Sep. 2013, doi: 10.1371/journal.pone.0075619.

[109] M. Pirooznia *et al.*, ‘Validation and assessment of variant calling pipelines for next-generation sequencing’, *Hum. Genomics*, vol. 8, no. 1, p. 14, Jul. 2014, doi: 10.1186/1479-7364-8-14.

[110] B.-Y. Kim, J. H. Park, H.-Y. Jo, S. K. Koo, and M.-H. Park, ‘Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data’, *PLoS One*, vol. 12, no. 8, p. e0182272, Aug. 2017, doi: 10.1371/journal.pone.0182272.

[111] M. S. Hasan, X. Wu, and L. Zhang, ‘Performance evaluation of indel calling tools using real short-read data’, *Hum. Genomics*, vol. 9, no. 1, p. 20, Dec. 2015, doi: 10.1186/s40246-015-0042-2.

[112] J. A. Neuman, O. Isakov, and N. Shomron, ‘Analysis of insertion-deletion from deep-sequencing data: Software evaluation for optimal detection’, *Brief. Bioinform.*, vol. 14, no. 1, pp. 46–55, Jan. 2013, doi: 10.1093/bib/bbs013.

[113] H. Li, J. Ruan, and R. Durbin, ‘Mapping short DNA sequencing reads and calling variants using mapping quality scores’, *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008, doi: 10.1101/gr.078212.108.

[114] S. Pei, T. Liu, X. Ren, W. Li, C. Chen, and Z. Xie, ‘Benchmarking variant callers in next-generation and third-generation sequencing analysis’, *Brief. Bioinform.*, Jul. 2020, doi: 10.1093/bib/bbaa148.

[115] D. Freed, R. Aldana, J. A. Weber, and J. S. Edwards, ‘The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data’, *bioRxiv*. bioRxiv, p. 115717, Mar. 10, 2017, doi: 10.1101/115717.

[116] U.S. Food and Drug Administration, ‘PrecisionFDA Truth Challenge V1’, 2016. https://precision.fda.gov/challenges/truth (accessed May 04, 2021).

[117] N. D. Olson *et al.*, ‘precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map regions’, *bioRxiv*, vol. 5. bioRxiv, p. 21, Nov. 15, 2020, doi: 10.1101/2020.11.13.380741.

[118] International HapMap Consortium, ‘The international HapMap project’, *Nature*, vol. 426, no. 6968, pp. 789–796, Dec. 2003, doi: 10.1038/nature02168.

[119] L. Malcovati *et al.*, ‘SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts’, *Blood*, vol. 126, no. 2, pp. 233–241, Jul. 2015, doi: 10.1182/blood-2015-03-633537.

[120] X. Wu *et al.*, ‘Identification of novel SNPs associated with coronary artery disease and birth weight using a pleiotropic cFDR method’, *Aging (Albany. NY)*., vol. 13, no. 3, pp. 3618–3632, Feb. 2021, doi: 10.18632/aging.202322.

[121] P. Danecek *et al.*, ‘The variant call format and VCFtools’, *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011, doi: 10.1093/bioinformatics/btr330.

[122] S. Friedman, L. Gauthier, Y. Farjoun, and E. Banks, ‘Lean and deep models for more accurate filtering of SNP and INDEL variant calls’, *Bioinformatics*, vol. 36, no. 7, pp. 2060–2067, Apr. 2020, doi: 10.1093/bioinformatics/btz901.

[123] M. Lek *et al.*, ‘Analysis of protein-coding genetic variation in 60,706 humans’, *Nature*, vol. 536, no. 7616, pp. 285–291, Aug. 2016, doi: 10.1038/nature19057.

[124] S. T. Sherry *et al.*, ‘dbSNP: The NCBI database of genetic variation’, *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.

[125] K. J. Karczewski *et al.*, ‘The mutational constraint spectrum quantified from variation in 141,456 humans’, *Nature*, vol. 581, no. 7809, pp. 434–443, May 2020, doi: 10.1038/s41586-020-2308-7.

[126] M. J. Landrum *et al.*, ‘ClinVar: Public archive of relationships among sequence variation and human phenotype’, *Nucleic Acids Res.*, vol. 42, no. D1, p. D980, Jan. 2014, doi: 10.1093/nar/gkt1113.

[127] N. A. O’Leary *et al.*, ‘Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation’, *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, 2016, doi: 10.1093/nar/gkv1189.

[128] K. L. Howe *et al.*, ‘Ensembl 2021’, *Nucleic Acids Res.*, vol. 49, no. D1, pp. D884–D891, Jan. 2021, doi: 10.1093/nar/gkaa942.

[129] D. J. McCarthy *et al.*, ‘Choice of transcripts and software has a large effect on variant annotation’,

Genome Med., vol. 6, no. 3, pp. 1–16, Mar. 2014, doi: 10.1186/gm543.

[130] K. Wang, M. Li, and H. Hakonarson, 'ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Res.*, vol. 38, no. 16, Jul. 2010, doi: 10.1093/nar/gkq603.

[131] P. Cingolani *et al.*, 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3', *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, 2012, doi: 10.4161/fly.19695.

[132] W. McLaren *et al.*, 'The Ensembl Variant Effect Predictor', *Genome Biol.*, vol. 17, no. 1, pp. 1–14, Jun. 2016, doi: 10.1186/s13059-016-0974-4.

[133] G. Escaramís, E. Docampo, and R. Rabionet, 'A decade of structural variants: Description, history and methods to detect structural variation', *Brief. Funct. Genomics*, vol. 14, no. 5, pp. 305–314, Sep. 2015, doi: 10.1093/bfgp/elv014.

[134] K. Chen *et al.*, 'BreakDancer: An algorithm for high-resolution mapping of genomic structural variation', *Nat. Methods*, vol. 6, no. 9, pp. 677–681, Aug. 2009, doi: 10.1038/nmeth.1363.

[135] T. Rausch, T. Zichner, A. Schlattl, A. M. Stütz, V. Benes, and J. O. Korbel, 'DELLY: Structural variant discovery by integrated paired-end and split-read analysis', *Bioinformatics*, vol. 28, no. 18, p. i333, Sep. 2012, doi: 10.1093/bioinformatics/bts378.

[136] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, 'LUMPY: A probabilistic framework for structural variant discovery', *Genome Biol.*, vol. 15, no. 6, pp. 1–19, Jun. 2014, doi: 10.1186/gb-2014-15-6-r84.

[137] Y. Jiang, Y. Wang, and M. Brudno, 'PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants', *Bioinformatics*, vol. 28, no. 20, pp. 2576–2583, Oct. 2012, doi: 10.1093/bioinformatics/bts484.

[138] Z. D. Zhang *et al.*, 'Identification of genomic indels and structural variations using split reads', *BMC Genomics*, vol. 12, p. 375, Jul. 2011, doi: 10.1186/1471-2164-12-375.

[139] K. Ye, G. Hall, and Z. Ning, 'Structural Variation Detection from Next Generation Sequencing', 2016, doi: 10.4172/2469-9853.S1-007.

[140] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, 'Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads', *Bioinformatics*, vol. 25, no. 21, pp. 2865–2871, Nov. 2009, doi: 10.1093/bioinformatics/btp394.

[141] K. Trappe, A. K. Emde, H. C. Ehrlich, and K. Reinert, 'Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone', *Bioinformatics*, vol. 30, no. 24, pp. 3484–3490, Dec. 2014, doi: 10.1093/bioinformatics/btu431.

[142] J. Wang *et al.*, ‘CREST maps somatic structural variation in cancer genomes with base-pair resolution’, *Nat. Methods*, vol. 8, no. 8, pp. 652–654, Aug. 2011, doi: 10.1038/nmeth.1628.

[143] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, ‘Statistical challenges associated with detecting copy number variations with next-generation sequencing’, *Bioinformatics*, vol. 28, no. 21. Oxford Academic, pp. 2711–2718, Nov. 01, 2012, doi: 10.1093/bioinformatics/bts535.

[144] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, ‘Sensitive and accurate detection of copy number variants using read depth of coverage’, *Genome Res.*, vol. 19, no. 9, pp. 1586–1592, Sep. 2009, doi: 10.1101/gr.092981.109.

[145] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, ‘CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing’, *Genome Res.*, vol. 21, no. 6, pp. 974–984, Jun. 2011, doi: 10.1101/gr.114876.110.

[146] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian, ‘CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing’, *PLoS Comput. Biol.*, vol. 12, no. 4, Apr. 2016, doi: 10.1371/journal.pcbi.1004873.

[147] G. Povysil *et al.*, ‘panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics’, *Hum. Mutat.*, vol. 38, no. 7, pp. 889–897, Jul. 2017, doi: 10.1002/humu.23237.

[148] A. Fowler *et al.*, ‘Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN’, *Wellcome Open Res.*, vol. 1, 2016, doi: 10.12688/wellcomeopenres.10069.1.

[149] J. F. Nijkamp, M. A. Van Den Broek, J. M. A. Geertman, M. J. T. Reinders, J. M. G. Daran, and D. De Ridder, ‘De novo detection of copy number variation by co-assembly’, *Bioinformatics*, vol. 28, no. 24, pp. 3195–3202, Dec. 2012, doi: 10.1093/bioinformatics/bts601.

[150] H. Li, ‘FermiKit: Assembly-based variant calling for Illumina resequencing data’, *Bioinformatics*, vol. 31, no. 22, pp. 3694–3696, May 2015, doi: 10.1093/bioinformatics/btv440.

[151] X. Chen *et al.*, ‘Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications’, *Bioinformatics*, vol. 32, no. 8, pp. 1220–1222, Apr. 2016, doi: 10.1093/bioinformatics/btv710.

[152] D. L. Cameron *et al.*, ‘GRIDSS: Sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly’, *Genome Res.*, vol. 27, no. 12, pp. 2050–2060, Dec. 2017, doi: 10.1101/gr.222109.117.

[153] S. Kosugi, Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani, ‘Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing’, *Genome Biol.*, vol. 20, no. 1, pp. 8–11, Jun. 2019, doi: 10.1186/s13059-019-1720-5.

[154] D. L. Cameron, L. Di Stefano, and A. T. Papenfuss, ‘Comprehensive evaluation and

characterisation of short read general-purpose structural variant calling software', *Nat. Commun.*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-11146-4.

[155] M. Pinese *et al.*, 'The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly', *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-019-14079-0.

[156] I. Roca, L. González-Castro, H. Fernández, M. L. Couce, and A. Fernández-Marmiesse, 'Free-access copy-number variant detection tools for targeted next-generation sequencing data', *Mutat. Res. - Rev. Mutat. Res.*, vol. 779, no. April 2018, pp. 114–125, Jan. 2019, doi: 10.1016/j.mrrev.2019.02.005.

[157] J. M. Moreno-Cabrera *et al.*, 'Evaluation of CNV detection tools for NGS panel data in genetic diagnostics', *Eur. J. Hum. Genet.*, vol. 28, no. 12, pp. 1645–1655, Dec. 2020, doi: 10.1038/s41431-020-0675-z.

[158] W. J. Kent *et al.*, 'The Human Genome Browser at UCSC', *Genome Res.*, vol. 12, no. 6, pp. 996–1006, May 2002, doi: 10.1101/gr.229102.

[159] A. R. Quinlan and I. M. Hall, 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Jan. 2010, doi: 10.1093/bioinformatics/btq033.

[160] Y. Liu, Y. Huang, G. Wang, and Y. Wang, 'A deep learning approach for filtering structural variants in short read sequencing data', *Brief. Bioinform.*, Dec. 2020, doi: 10.1093/bib/bbaa370.

[161] J. R. Belyeu *et al.*, 'Samplot: A platform for structural varian visual validation and automated filtering', *bioRxiv*. bioRxiv, p. 2020.09.23.310110, Sep. 25, 2020, doi: 10.1101/2020.09.23.310110.

[162] V. Makarov, T. O'Grady, G. Cai, J. Lihm, J. D. Buxbaum, and S. Yoon, 'Anntools: A comprehensive and versatile annotation toolkit for genomic variants', *Bioinformatics*, vol. 28, no. 5, pp. 724–725, Mar. 2012, doi: 10.1093/bioinformatics/bts032.

[163] Y. Zhang *et al.*, 'DeAnnCNV: A tool for online detection and annotation of copy number variations from whole-exome sequencing data', *Nucleic Acids Res.*, vol. 43, no. W1, pp. W289–W294, 2015, doi: 10.1093/nar/gkv556.

[164] M. Zhao and Z. Zhao, 'CNVannotator: A comprehensive annotation server for copy number variation in the human genome', *PLoS One*, vol. 8, no. 11, p. 80170, Nov. 2013, doi: 10.1371/journal.pone.0080170.

[165] V. Geoffroy *et al.*, 'AnnotSV: An integrated tool for structural variations annotation', *Bioinformatics*, vol. 34, no. 20, pp. 3572–3574, Oct. 2018, doi: 10.1093/bioinformatics/bty304.

[166] S. Richards *et al.*, 'Standards and guidelines for the interpretation of sequence variants: A joint

consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genet. Med.*, vol. 17, no. 5, pp. 405–424, May 2015, doi: 10.1038/gim.2015.30.

[167] E. R. Riggs *et al.*, 'Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)', *Genet. Med.*, vol. 22, no. 2, pp. 245–257, Feb. 2020, doi: 10.1038/s41436-019-0686-8.

[168] N. A. Miller *et al.*, 'A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases', *Genome Med.*, vol. 7, no. 1, pp. 1–16, Sep. 2015, doi: 10.1186/s13073-015-0221-8.

[169] C. Kopanos *et al.*, 'VarSome: the human genomic variant search engine', *Bioinformatics*, vol. 35, no. 11, pp. 1978–1980, Jun. 2019, doi: 10.1093/bioinformatics/bty897.

[170] B. B. Misra, C. Langefeld, M. Olivier, and L. A. Cox, 'Integrated omics: Tools, advances and future approaches', *Journal of Molecular Endocrinology*, vol. 62, no. 1. BioScientifica Ltd., pp. R21–R45, Jan. 01, 2019, doi: 10.1530/JME-18-0055.

[171] A. J. Radenbaugh *et al.*, 'RADIA: RNA and DNA integrated analysis for somatic mutation detection', *PLoS One*, vol. 9, no. 11, Nov. 2014, doi: 10.1371/journal.pone.0111516.

[172] L. Neums *et al.*, 'VaDiR: An integrated approach to Variant Detection in RNA', *Gigascience*, vol. 7, no. 2, pp. 1–13, Feb. 2018, doi: 10.1093/gigascience/gix122.

[173] R. E. Mills *et al.*, 'An initial map of insertion and deletion (INDEL) variation in the human genome', *Genome Res.*, vol. 16, no. 9, pp. 1182–1190, 2006, doi: 10.1101/gr.4565806.

[174] A. Tan, G. R. Abecasis, and H. M. Kang, 'Unified representation of genetic variants', *Bioinformatics*, vol. 31, no. 13, pp. 2202–2204, Jul. 2015, doi: 10.1093/bioinformatics/btv112.

[175] H. Li, 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, Nov. 2011, doi: 10.1093/bioinformatics/btr509.

[176] P. Danecek *et al.*, 'Twelve years of SAMtools and BCFtools', *Gigascience*, vol. 10, no. 2, Feb. 2021, doi: 10.1093/gigascience/giab008.

[177] X. Liu, C. Wu, C. Li, and E. Boerwinkle, 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs', *Hum. Mutat.*, vol. 37, no. 3, pp. 235–241, Mar. 2016, doi: 10.1002/humu.22932.

[178] X. Jian, E. Boerwinkle, and X. Liu, 'In silico prediction of splice-altering single nucleotide variants in the human genome', *Nucleic Acids Res.*, vol. 42, no. 22, pp. 13534–13544, Dec. 2014, doi:

10.1093/nar/gku1206.

[179] J. A. Tennessen *et al.*, ‘Evolution and functional impact of rare coding variation from deep sequencing of human exomes’, *Science (80-.).*, vol. 336, no. 6090, pp. 64–69, Jul. 2012, doi: 10.1126/science.1219240.

[180] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler, ‘Segmental duplications: Organization and impact within the current human genome project assembly’, *Genome Res.*, vol. 11, no. 6, pp. 1005–1017, 2001, doi: 10.1101/gr.GR-1871R.

[181] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, ‘OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders’, *Nucleic Acids Res.*, vol. 43, no. D1, pp. D789–D798, Jan. 2015, doi: 10.1093/nar/gku1205.

[182] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, ‘OMIM.org: Leveraging knowledge across phenotype-gene relationships’, *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1038–D1043, Jan. 2019, doi: 10.1093/nar/gky1151.

[183] A. Xavier, R. J. Scott, and B. A. Talseth-Palmer, ‘TAPES: A tool for assessment and prioritisation in exome studies’, *PLoS Comput. Biol.*, vol. 15, no. 10, pp. 1–9, 2019, doi: 10.1371/journal.pcbi.1007453.

[184] E. Bortolini *et al.*, ‘Early Alpine occupation backdates westward human migration in Late Glacial Europe’, *Curr. Biol.*, vol. 31, no. 11, pp. 2484-2493.e7, Apr. 2021, doi: 10.1016/j.cub.2021.03.078.

[185] M. Schubert *et al.*, ‘Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX’, *Nat. Protoc.*, vol. 9, no. 5, pp. 1056–1082, Apr. 2014, doi: 10.1038/nprot.2014.063.

[186] A. Paytuví, E. Battista, F. Scippacercola, R. Aiese Cigliano, and W. Sanseverino, ‘GAIA: an integrated metagenomics suite’, *bioRxiv*, p. 804690, Jan. 2019, doi: 10.1101/804690.

[187] P. Skoglund, J. Storå, A. Götherström, and M. Jakobsson, ‘Accurate sex identification of ancient human remains using DNA shotgun sequencing’, *J. Archaeol. Sci.*, vol. 40, no. 12, pp. 4477–4482, 2013, doi: 10.1016/j.jas.2013.07.004.

[188] H. L. Rehm *et al.*, ‘ClinGen — The Clinical Genome Resource’, *N. Engl. J. Med.*, vol. 372, no. 23, pp. 2235–2242, Jun. 2015, doi: 10.1056/nejmsr1406261.

[189] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, ‘The Database of Genomic Variants: A curated collection of structural variation in the human genome’, *Nucleic Acids Res.*, vol. 42, no. D1, Jan. 2014, doi: 10.1093/nar/gkt958.

[190] I. Lappalainen *et al.*, ‘DbVar and DGVa: Public archives for genomic structural variation’, *Nucleic*

Acids Res., vol. 41, no. D1, p. D936, Jan. 2013, doi: 10.1093/nar/gks1213.

[191] H. V. Firth and C. F. Wright, 'The Deciphering Developmental Disorders (DDD) study', *Developmental Medicine and Child Neurology*, vol. 53, no. 8. Dev Med Child Neurol, pp. 702–703, Aug. 2011, doi: 10.1111/j.1469-8749.2011.04032.x.

[192] H. V. Firth *et al.*, 'DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources', *Am. J. Hum. Genet.*, vol. 84, no. 4, pp. 524–533, Apr. 2009, doi: 10.1016/j.ajhg.2009.03.010.

[193] R. L. Collins *et al.*, 'A structural variation reference for medical and population genetics', *Nature*, vol. 581, no. 7809, pp. 444–451, May 2020, doi: 10.1038/s41586-020-2287-8.

[194] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, 'Characterising and predicting haploinsufficiency in the human genome', *PLoS Genet.*, vol. 6, no. 10, pp. 1–11, Oct. 2010, doi: 10.1371/journal.pgen.1001154.

[195] H. J. Abel *et al.*, 'Mapping and characterization of structural variation in 17,795 human genomes', *Nature*, vol. 583, no. 7814, pp. 83–89, Jul. 2020, doi: 10.1038/s41586-020-2371-0.

[196] J. Kerkhof *et al.*, 'Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels', *J. Mol. Diagnostics*, vol. 19, no. 6, pp. 905–920, 2017, doi: 10.1016/j.jmoldx.2017.07.004.

[197] S. Mahamdallie *et al.*, 'The ICR96 exon CNV validation series: A resource for orthogonal assessment of exon CNV calling in NGS data', *Wellcome Open Res.*, vol. 2, no. 0, pp. 1–9, 2017, doi: 10.12688/wellcomeopenres.11689.1.

[198] S. J. Sanders *et al.*, 'De novo mutations revealed by whole-exome sequencing are strongly associated with autism', *Nature*, vol. 485, no. 7397, pp. 237–241, May 2012, doi: 10.1038/nature10945.

[199] N. Krumm *et al.*, 'Excess of rare, inherited truncating mutations in autism', *Nat. Genet.*, vol. 47, no. 6, pp. 582–588, May 2015, doi: 10.1038/ng.3303.

[200] R. Leinonen, H. Sugawara, and M. Shumway, 'The sequence read archive', *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, p. D19, Jan. 2011, doi: 10.1093/nar/gkq1019.

[201] V. Plagnol *et al.*, 'A robust model for read count data in exome sequencing experiments and implications for copy number variant calling', *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, Nov. 2012, doi: 10.1093/bioinformatics/bts526.

[202] R. K. Dale, B. S. Pedersen, and A. R. Quinlan, 'Pybedtools: A flexible Python library for manipulating genomic datasets and annotations', *Bioinformatics*, vol. 27, no. 24, pp. 3423–3424, Dec. 2011, doi: 10.1093/bioinformatics/btr539.

[203] R. Piskol, G. Ramaswami, and J. B. Li, ‘Reliable identification of genomic variants from RNA-seq data’, *Am. J. Hum. Genet.*, vol. 93, no. 4, pp. 641–651, Oct. 2013, doi: 10.1016/j.ajhg.2013.08.008.

[204] S. Lam *et al.*, ‘Development and comparison of RNA-sequencing pipelines for more accurate SNP identification: Practical example of functional SNP detection associated with feed efficiency in Nellore beef cattle’, *BMC Genomics*, vol. 21, no. 1, pp. 1–17, Oct. 2020, doi: 10.1186/s12864-020-07107-7.

[205] J. H. Lee, J. K. Ang, and X. Xiao, ‘Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants’, *RNA*, vol. 19, no. 6. Cold Spring Harbor Laboratory Press, pp. 725–732, Jun. 2013, doi: 10.1261/rna.037903.112.

[206] N. Paz-Yaacov *et al.*, ‘Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors’, *Cell Rep.*, vol. 13, no. 2, pp. 267–276, Oct. 2015, doi: 10.1016/j.celrep.2015.08.080.

[207] L. Han *et al.*, ‘The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers’, *Cancer Cell*, vol. 28, no. 4, pp. 515–528, Oct. 2015, doi: 10.1016/j.ccell.2015.08.013.

[208] C. P. Kung, L. B. Maggi, and J. D. Weber, ‘The Role of RNA Editing in Cancer Development and Metabolic Disorders’, *Frontiers in Endocrinology*, vol. 9. Frontiers Media S.A., p. 762, Dec. 18, 2018, doi: 10.3389/fendo.2018.00762.

[209] H. Krestel and J. C. Meier, ‘RNA editing and retrotransposons in neurology’, *Frontiers in Molecular Neuroscience*, vol. 11. Frontiers Media S.A., p. 163, May 23, 2018, doi: 10.3389/fnmol.2018.00163.

[210] C. Wang *et al.*, ‘RVboost: RNA-seq variants prioritization using a boosting method’, *Bioinformatics*, vol. 30, no. 23, pp. 3414–3416, Dec. 2014, doi: 10.1093/bioinformatics/btu577.

[211] X. Tang *et al.*, ‘The eSNV-detect: A computational system to identify expressed single nucleotide variants from transcriptome sequencing data’, *Nucleic Acids Res.*, vol. 42, no. 22, p. e172, Dec. 2014, doi: 10.1093/nar/gku1005.

[212] J. M. Zook *et al.*, ‘Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls’, *Nat. Biotechnol.*, vol. 32, no. 3, pp. 246–251, Feb. 2014, doi: 10.1038/nbt.2835.

[213] T. Lappalainen *et al.*, ‘Transcriptome and genome sequencing uncovers functional variation in humans’, *Nature*, vol. 501, no. 7468, pp. 506–511, Sep. 2013, doi: 10.1038/nature12531.

[214] A. Dobin *et al.*, ‘STAR: Ultrafast universal RNA-seq aligner’, *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[215] B. A. Veeneman, S. Shukla, S. M. Dhanasekaran, A. M. Chinnaiyan, and A. I. Nesvizhskii, 'Two-pass alignment improves novel splice junction quantification', *Bioinformatics*, vol. 32, no. 1, pp. 43–49, Jan. 2016, doi: 10.1093/bioinformatics/btv642.

[216] Y. Guo *et al.*, 'The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation', *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.*, vol. 744, no. 2, pp. 154–160, May 2012, doi: 10.1016/j.mrgentox.2012.02.006.

[217] H. M. Amemiya, A. Kundaje, and A. P. Boyle, 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Sci. Rep.*, vol. 9, no. 1, pp. 1–5, Dec. 2019, doi: 10.1038/s41598-019-45839-z.

[218] P. Krusche *et al.*, 'Best practices for benchmarking germline small-variant calls in human genomes', *Nat. Biotechnol.*, vol. 37, no. 5, pp. 555–560, May 2019, doi: 10.1038/s41587-019-0054-x.

[219] P. Krusche, 'Haplotype comparison tools / hap.py'. Accessed: May 07, 2021. [Online]. Available: https://github.com/illumina/hap.py.

[220] J. Cleary *et al.*, 'Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines', *bioRxiv*, p. 023754, Aug. 2015, doi: 10.1101/023754.

[221] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. S. Durant, and R. Torres, 'Evaluation of relational and NoSQL database architectures to manage genomic annotations', *Journal of Biomedical Informatics*, vol. 64. pp. 288–295, 2016, doi: 10.1016/j.jbi.2016.10.015.

[222] R. Aniceto *et al.*, 'Evaluating the cassandra NoSQL database approach for genomic data persistency', *Int. J. Genomics*, vol. 2015, 2015, doi: 10.1155/2015/502795.

[223] P. Wercelens *et al.*, 'Bioinformatics Workflows With NoSQL Database in Cloud Computing', *Evol. Bioinforma.*, vol. 15, 2019, doi: 10.1177/1176934319889974.

[224] H. Samra, A. Li, and B. Soh, 'Gene2d: A nosql integrated data repository of genetic disorders data', *Healthc.*, vol. 8, no. 3, 2020, doi: 10.3390/healthcare8030257.

[225] M. Ruffier *et al.*, 'Ensembl core software resources: Storage and programmatic access for DNA sequence and genome annotation', *Database*, vol. 2017, no. 1, 2017, doi: 10.1093/database/bax020.

[226] 'MySQL - Restrictions and Limitations'. https://dev.mysql.com/doc/mysql-reslimits-excerpt/5.7/en/limits.html (accessed Jun. 22, 2021).

[227] G. A. Van der Auwera and B. D. O'Connor, *Genomics in the cloud : using Docker, GATK, and WDL in Terra*, 1st Editio. O'Reilly Media, 2020.

[228] K. Prüfer *et al.*, 'The complete genome sequence of a Neanderthal from the Altai Mountains',

Nature, vol. 505, no. 7481, pp. 43–49, 2014, doi: 10.1038/nature12886.

[229] T. Lan *et al.*, 'Improving Species Identification of Ancient Mammals Based on Next-Generation Sequencing Data', *Genes (Basel).*, vol. 10, no. 7, p. 509, 2019, doi: 10.3390/genes10070509.

Appendix A | Supplementary data

A.1 Supplementary tables

Supplementary Table A. 1. Benchmark results for variants identified with varRED and with the short variant discovery.

Sample	Data type	Variant type	Match	Variants			Analyzed variant	Recall	Precision	F-score
				Total	SNPs	Indels				
NA12878	DNA	ALL	GT	260535	224987	35548	SNP	0.9965	0.9652	0.9806
NA12878	DNA	ALL	GT	260535	224987	35548	INDEL	0.9567	0.3987	0.5628
NA12878	DNA	ALL	AL	266317	229854	36463	SNP	0.9998	0.9534	0.9761
NA12878	DNA	ALL	AL	266317	229854	36463	INDEL	0.9901	0.4053	0.5751
NA12878	RNA	ALL	GT	47288	42359	4929	SNP	0.9916	0.9614	0.9763
NA12878	RNA	ALL	GT	47288	42359	4929	INDEL	0.9744	0.4027	0.5698
NA12878	RNA	ALL	AL	68881	60859	8022	SNP	0.9998	0.9492	0.9739
NA12878	RNA	ALL	AL	68881	60859	8022	INDEL	0.9929	0.3735	0.5428
NA12878	DNA+RNA	ALL	AL	287333	248646	38687	SNP	0.9998	0.9541	0.9764
NA12878	DNA+RNA	ALL	AL	287333	248646	38687	INDEL	0.9899	0.4058	0.5756
NA12878	DNA	Strong-evidence	GT	21748	19671	2077	SNP	0.9999	0.9863	0.9931
NA12878	DNA	Strong-evidence	GT	21748	19671	2077	INDEL	0.9908	0.437	0.6066
NA12878	DNA	Strong-evidence	AL	21748	19671	2077	SNP	1	0.9864	0.9931
NA12878	DNA	Strong-evidence	AL	21748	19671	2077	INDEL	0.9981	0.4403	0.611
NA12878	RNA	Strong-evidence	GT	21748	19671	2077	SNP	0.9999	0.9863	0.9931
NA12878	RNA	Strong-evidence	GT	21748	19671	2077	INDEL	0.9908	0.437	0.6066
NA12878	RNA	Strong-evidence	AL	21748	19671	2077	SNP	1	0.9864	0.9931
NA12878	RNA	Strong-evidence	AL	21748	19671	2077	INDEL	0.9981	0.4403	0.611
NA12878	DNA+RNA	Strong-evidence	AL	21748	19671	2077	SNP	1	0.9864	0.9931
NA12878	DNA+RNA	Strong-evidence	AL	21748	19671	2077	INDEL	0.9981	0.4403	0.611
NA12878	DNA	DNA-only	GT	238787	205316	33471	SNP	0.9961	0.9627	0.9791
NA12878	DNA	DNA-only	GT	238787	205316	33471	INDEL	0.9534	0.3954	0.559
NA12878	DNA	DNA-only	AL	238787	205316	33471	SNP	0.9997	0.951	0.9748
NA12878	DNA	DNA-only	AL	238787	205316	33471	INDEL	0.989	0.4061	0.5758
NA12878	RNA	DNA-only	GT	20348	17535	2813	SNP	0.4992	0.73	0.5929
NA12878	RNA	DNA-only	GT	20348	17535	2813	INDEL	0.6979	0.4465	0.5446
NA12878	RNA	DNA-only	AL	20348	17535	2813	SNP	0.9998	0.9061	0.9506
NA12878	RNA	DNA-only	AL	20348	17535	2813	INDEL	0.9863	0.3056	0.4667
NA12878	DNA+RNA	DNA-only	AL	238795	205320	33475	SNP	0.9997	0.951	0.9748

NA12878	DNA+RNA	DNA-only	AL	238795	205320	33475	INDEL	0.9888	0.4061	0.5757
NA12878	DNA	RNA-only	GT	4339	3822	517	SNP	0.5122	0.8362	0.6353
NA12878	DNA	RNA-only	GT	4339	3822	517	INDEL	0.899	0.6556	0.7582
NA12878	DNA	RNA-only	AL	4339	3822	517	SNP	1	0.914	0.9551
NA12878	DNA	RNA-only	AL	4339	3822	517	INDEL	1	0.3363	0.5034
NA12878	RNA	RNA-only	GT	25344	22607	2737	SNP	0.9841	0.94	0.9616
NA12878	RNA	RNA-only	GT	25344	22607	2737	INDEL	0.9643	0.3864	0.5517
NA12878	RNA	RNA-only	AL	25344	22607	2737	SNP	0.9998	0.953	0.9758
NA12878	RNA	RNA-only	AL	25344	22607	2737	INDEL	0.9914	0.3985	0.5685
NA12878	DNA+RNA	RNA-only	AL	25344	22607	2737	SNP	0.9998	0.953	0.9758
NA12878	DNA+RNA	RNA-only	AL	25344	22607	2737	INDEL	0.9914	0.3985	0.5685
NA12878	DNA	ASE	GT	1250	967	283	SNP	1	0.7768	0.8744
NA12878	DNA	ASE	GT	1250	967	283	INDEL	0.7777	0.0933	0.1666
NA12878	DNA	ASE	AL	1250	967	283	SNP	1	0.7768	0.8744
NA12878	DNA	ASE	AL	1250	967	283	INDEL	1	0.12	0.2142
NA12878	RNA	ASE	GT	1245	965	280	SNP	0	0	0
NA12878	RNA	ASE	GT	1245	965	280	INDEL	0.1111	0.1666	0.1333
NA12878	RNA	ASE	AL	1245	965	280	SNP	1	0.7768	0.8744
NA12878	RNA	ASE	AL	1245	965	280	INDEL	1	0.12	0.2142
NA12878	DNA+RNA	ASE	AL	1250	967	283	SNP	1	0.7768	0.8744
NA12878	DNA+RNA	ASE	AL	1250	967	283	INDEL	1	0.12	0.2142
NA12878	DNA	RNA-editing	GT	13257	10852	2405	SNP	0.9968	0.9681	0.9823
NA12878	DNA	RNA-editing	GT	13257	10852	2405	INDEL	0.9448	0.3547	0.5158
NA12878	DNA	RNA-editing	AL	13257	10852	2405	SNP	0.9996	0.4737	0.6428
NA12878	DNA	RNA-editing	AL	13257	10852	2405	INDEL	0.9795	0.141	0.2465
NA12878	RNA	RNA-editing	GT	2125	1183	942	SNP	0.0588	0.0016	0.0032
NA12878	RNA	RNA-editing	GT	2125	1183	942	INDEL	0	0	0
NA12878	RNA	RNA-editing	AL	2125	1183	942	SNP	1	0.0283	0.0551
NA12878	RNA	RNA-editing	AL	2125	1183	942	INDEL	0.8333	0.0245	0.0476
NA12878	DNA+RNA	RNA-editing	AL	13264	10857	2407	SNP	0.9996	0.4736	0.6427
NA12878	DNA+RNA	RNA-editing	AL	13264	10857	2407	INDEL	0.9795	0.141	0.2465
NA12878	DNA	RNA-rescue	GT	193	78	115	SNP	0.5714	0.923	0.7058

NA12878	DNA	RNA-rescue	GT	193	78	115	INDEL	0.6	0.2727	0.375
NA12878	DNA	RNA-rescue	AL	193	78	115	SNP	1	0.7	0.8235
NA12878	DNA	RNA-rescue	AL	193	78	115	INDEL	0.8	0.125	0.2162
NA12878	RNA	RNA-rescue	GT	196	81	115	SNP	1	0.7	0.8235
NA12878	RNA	RNA-rescue	GT	196	81	115	INDEL	0.4	0.0625	0.1081
NA12878	RNA	RNA-rescue	AL	196	81	115	SNP	1	0.7	0.8235
NA12878	RNA	RNA-rescue	AL	196	81	115	INDEL	0.8	0.125	0.2162
NA12878	DNA+RNA	RNA-rescue	AL	196	81	115	SNP	1	0.7	0.8235
NA12878	DNA+RNA	RNA-rescue	AL	196	81	115	INDEL	0.8	0.125	0.2162
HG00171	DNA	ALL	GT	131453	113844	17609	SNP	0.9937	0.885	0.9362
HG00171	DNA	ALL	GT	131453	113844	17609	INDEL	0.895	0.5655	0.6931
HG00171	DNA	ALL	AL	134158	116030	18128	SNP	0.9744	0.8839	0.9269
HG00171	DNA	ALL	AL	134158	116030	18128	INDEL	0.7306	0.5953	0.656
HG00171	RNA	ALL	GT	13295	11425	1870	SNP	0.9954	0.9345	0.9639
HG00171	RNA	ALL	GT	13295	11425	1870	INDEL	0.9285	0.4487	0.605
HG00171	RNA	ALL	AL	19922	17403	2519	SNP	0.9782	0.9019	0.9385
HG00171	RNA	ALL	AL	19922	17403	2519	INDEL	0.8	0.515	0.6266
HG00171	DNA+RNA	ALL	AL	140257	120943	19314	SNP	0.9745	0.8858	0.928
HG00171	DNA+RNA	ALL	AL	140257	120943	19314	INDEL	0.734	0.5896	0.6539
HG00171	DNA	Strong-evidence	GT	4863	4667	196	SNP	0.9993	0.9907	0.995
HG00171	DNA	Strong-evidence	GT	4863	4667	196	INDEL	0.9806	0.7835	0.871
HG00171	DNA	Strong-evidence	AL	4863	4667	196	SNP	0.9855	0.9912	0.9883
HG00171	DNA	Strong-evidence	AL	4863	4667	196	INDEL	0.8437	0.7886	0.8152
HG00171	RNA	Strong-evidence	GT	4863	4667	196	SNP	0.9993	0.9907	0.995
HG00171	RNA	Strong-evidence	GT	4863	4667	196	INDEL	0.9806	0.7835	0.871
HG00171	RNA	Strong-evidence	AL	4863	4667	196	SNP	0.9855	0.9912	0.9883
HG00171	RNA	Strong-evidence	AL	4863	4667	196	INDEL	0.8437	0.7886	0.8152
HG00171	DNA+RNA	Strong-evidence	AL	4863	4667	196	SNP	0.9855	0.9912	0.9883
HG00171	DNA+RNA	Strong-evidence	AL	4863	4667	196	INDEL	0.8437	0.7886	0.8152
HG00171	DNA	DNA-only	GT	126590	109177	17413	SNP	0.9934	0.8803	0.9335
HG00171	DNA	DNA-only	GT	126590	109177	17413	INDEL	0.8938	0.5629	0.6908
HG00171	DNA	DNA-only	AL	126590	109177	17413	SNP	0.9743	0.8818	0.9258

HG00171	DNA	DNA-only	AL	126590	109177	17413	INDEL	0.7282	0.6005	0.6582
HG00171	RNA	DNA-only	GT	6256	5638	618	SNP	0.4941	0.6912	0.5762
HG00171	RNA	DNA-only	GT	6256	5638	618	INDEL	0.2891	0.4819	0.3614
HG00171	RNA	DNA-only	AL	6256	5638	618	SNP	0.9826	0.8595	0.917
HG00171	RNA	DNA-only	AL	6256	5638	618	INDEL	0.7786	0.5875	0.6697
HG00171	DNA+RNA	DNA-only	AL	126590	109177	17413	SNP	0.9743	0.8818	0.9258
HG00171	DNA+RNA	DNA-only	AL	126590	109177	17413	INDEL	0.7282	0.6005	0.6582
HG00171	DNA	RNA-only	GT	2263	1821	442	SNP	0.4733	0.7883	0.5915
HG00171	DNA	RNA-only	GT	2263	1821	442	INDEL	0.4078	0.6526	0.502
HG00171	DNA	RNA-only	AL	2263	1821	442	SNP	0.9487	0.8445	0.8936
HG00171	DNA	RNA-only	AL	2263	1821	442	INDEL	0.81	0.3597	0.4982
HG00171	RNA	RNA-only	GT	8357	6730	1627	SNP	0.9925	0.8966	0.9421
HG00171	RNA	RNA-only	GT	8357	6730	1627	INDEL	0.9186	0.42	0.5765
HG00171	RNA	RNA-only	AL	8357	6730	1627	SNP	0.9696	0.9069	0.9372
HG00171	RNA	RNA-only	AL	8357	6730	1627	INDEL	0.8061	0.4643	0.5893
HG00171	DNA+RNA	RNA-only	AL	8357	6730	1627	SNP	0.9696	0.9069	0.9372
HG00171	DNA+RNA	RNA-only	AL	8357	6730	1627	INDEL	0.8061	0.4643	0.5893
HG00171	DNA	ASE	GT	372	341	31	SNP	0.988	0.253	0.4029
HG00171	DNA	ASE	GT	372	341	31	INDEL	0.7692	0.3333	0.4651
HG00171	DNA	ASE	AL	372	341	31	SNP	0.9696	0.2926	0.4496
HG00171	DNA	ASE	AL	372	341	31	INDEL	0.6666	0.4333	0.5252
HG00171	RNA	ASE	GT	371	340	31	SNP	0.0119	0.037	0.018
HG00171	RNA	ASE	GT	371	340	31	INDEL	0.0769	0.25	0.1176
HG00171	RNA	ASE	AL	371	340	31	SNP	0.9696	0.2926	0.4496
HG00171	RNA	ASE	AL	371	340	31	INDEL	0.6666	0.4333	0.5252
HG00171	DNA+RNA	ASE	AL	372	341	31	SNP	0.9696	0.2926	0.4496
HG00171	DNA+RNA	ASE	AL	372	341	31	INDEL	0.6666	0.4333	0.5252
HG00171	DNA	RNA-editing	GT	2501	1817	684	SNP	0.9773	0.6666	0.7926
HG00171	DNA	RNA-editing	GT	2501	1817	684	INDEL	0.8955	0.4461	0.5955
HG00171	DNA	RNA-editing	AL	2501	1817	684	SNP	0.9242	0.495	0.6447
HG00171	DNA	RNA-editing	AL	2501	1817	684	INDEL	0.6475	0.1059	0.1821
HG00171	RNA	RNA-editing	GT	835	362	473	SNP	0.1333	0.0057	0.011

HG00171	RNA	RNA-editing	GT	835	362	473	INDEL	0.1111	0.0021	0.0041
HG00171	RNA	RNA-editing	AL	835	362	473	SNP	0.6753	0.144	0.2374
HG00171	RNA	RNA-editing	AL	835	362	473	INDEL	0.5862	0.0359	0.0677
HG00171	DNA+RNA	RNA-editing	AL	2503	1818	685	SNP	0.9253	0.495	0.645
HG00171	DNA+RNA	RNA-editing	AL	2503	1818	685	INDEL	0.6393	0.1044	0.1796
HG00171	DNA	RNA-rescue	GT	70	24	46	SNP	0.25	0.5	0.3333
HG00171	DNA	RNA-rescue	GT	70	24	46	INDEL	0.5	0.3333	0.4
HG00171	DNA	RNA-rescue	AL	70	24	46	SNP	0.9285	0.5909	0.7222
HG00171	DNA	RNA-rescue	AL	70	24	46	INDEL	0.7142	0.25	0.3703
HG00171	RNA	RNA-rescue	GT	75	28	47	SNP	0.9375	0.5769	0.7142
HG00171	RNA	RNA-rescue	GT	75	28	47	INDEL	0.6666	0.05	0.093
HG00171	RNA	RNA-rescue	AL	75	28	47	SNP	0.9444	0.6538	0.7727
HG00171	RNA	RNA-rescue	AL	75	28	47	INDEL	0.7333	0.2682	0.3928
HG00171	DNA+RNA	RNA-rescue	AL	75	28	47	SNP	0.9444	0.6538	0.7727
HG00171	DNA+RNA	RNA-rescue	AL	75	28	47	INDEL	0.7333	0.2682	0.3928
HG00378	DNA	ALL	GT	134219	116216	18003	SNP	0.9936	0.8797	0.9332
HG00378	DNA	ALL	GT	134219	116216	18003	INDEL	0.8869	0.568	0.6925
HG00378	DNA	ALL	AL	138657	120023	18634	SNP	0.9742	0.8774	0.9232
HG00378	DNA	ALL	AL	138657	120023	18634	INDEL	0.7271	0.5993	0.657
HG00378	RNA	ALL	GT	19761	17454	2307	SNP	0.9965	0.9425	0.9687
HG00378	RNA	ALL	GT	19761	17454	2307	INDEL	0.9272	0.5179	0.6646
HG00378	RNA	ALL	AL	27187	24154	3033	SNP	0.9753	0.9044	0.9385
HG00378	RNA	ALL	AL	27187	24154	3033	INDEL	0.7897	0.5792	0.6683
HG00378	DNA+RNA	ALL	AL	147339	127262	20077	SNP	0.974	0.8809	0.9251
HG00378	DNA+RNA	ALL	AL	147339	127262	20077	INDEL	0.732	0.5983	0.6584
HG00378	DNA	Strong-evidence	GT	7260	6985	275	SNP	0.9995	0.9941	0.9968
HG00378	DNA	Strong-evidence	GT	7260	6985	275	INDEL	0.9613	0.7252	0.8267
HG00378	DNA	Strong-evidence	AL	7260	6985	275	SNP	0.9859	0.9945	0.9902
HG00378	DNA	Strong-evidence	AL	7260	6985	275	INDEL	0.806	0.7435	0.7735
HG00378	RNA	Strong-evidence	GT	7260	6985	275	SNP	0.9995	0.9941	0.9968
HG00378	RNA	Strong-evidence	GT	7260	6985	275	INDEL	0.9613	0.7252	0.8267
HG00378	RNA	Strong-evidence	AL	7260	6985	275	SNP	0.9859	0.9945	0.9902

HG00378	RNA	Strong-evidence	AL	7260	6985	275	INDEL	0.806	0.7435	0.7735
HG00378	DNA+RNA	Strong-evidence	AL	7260	6985	275	SNP	0.9859	0.9945	0.9902
HG00378	DNA+RNA	Strong-evidence	AL	7260	6985	275	INDEL	0.806	0.7435	0.7735
HG00378	DNA	DNA-only	GT	126959	109231	17728	SNP	0.9931	0.8722	0.9287
HG00378	DNA	DNA-only	GT	126959	109231	17728	INDEL	0.8856	0.5655	0.6903
HG00378	DNA	DNA-only	AL	126959	109231	17728	SNP	0.9738	0.8745	0.9215
HG00378	DNA	DNA-only	AL	126959	109231	17728	INDEL	0.7248	0.6038	0.6588
HG00378	RNA	DNA-only	GT	6808	6124	684	SNP	0.4401	0.7032	0.5414
HG00378	RNA	DNA-only	GT	6808	6124	684	INDEL	0.2974	0.6216	0.4023
HG00378	RNA	DNA-only	AL	6808	6124	684	SNP	0.9761	0.8615	0.9153
HG00378	RNA	DNA-only	AL	6808	6124	684	INDEL	0.7629	0.6543	0.7044
HG00378	DNA+RNA	DNA-only	AL	126960	109232	17728	SNP	0.9738	0.8745	0.9215
HG00378	DNA+RNA	DNA-only	AL	126960	109232	17728	INDEL	0.7248	0.6038	0.6588
HG00378	DNA	RNA-only	GT	3683	3157	526	SNP	0.4624	0.7438	0.5703
HG00378	DNA	RNA-only	GT	3683	3157	526	INDEL	0.4493	0.6754	0.5396
HG00378	DNA	RNA-only	AL	3683	3157	526	SNP	0.9616	0.8482	0.9013
HG00378	DNA	RNA-only	AL	3683	3157	526	INDEL	0.7973	0.458	0.5818
HG00378	RNA	RNA-only	GT	12357	10388	1969	SNP	0.9943	0.909	0.9497
HG00378	RNA	RNA-only	GT	12357	10388	1969	INDEL	0.9247	0.5038	0.6522
HG00378	RNA	RNA-only	AL	12357	10388	1969	SNP	0.9686	0.9115	0.9392
HG00378	RNA	RNA-only	AL	12357	10388	1969	INDEL	0.8059	0.5513	0.6547
HG00378	DNA+RNA	RNA-only	AL	12357	10388	1969	SNP	0.9686	0.9115	0.9392
HG00378	DNA+RNA	RNA-only	AL	12357	10388	1969	INDEL	0.8059	0.5513	0.6547
HG00378	DNA	ASE	GT	618	576	42	SNP	0.9841	0.1213	0.216
HG00378	DNA	ASE	GT	618	576	42	INDEL	0.6363	0.1707	0.2692
HG00378	DNA	ASE	AL	618	576	42	SNP	0.885	0.1506	0.2575
HG00378	DNA	ASE	AL	618	576	42	INDEL	0.5555	0.2439	0.3389
HG00378	RNA	ASE	GT	618	576	42	SNP	0.0158	0.05	0.024
HG00378	RNA	ASE	GT	618	576	42	INDEL	0.0909	0.25	0.1333
HG00378	RNA	ASE	AL	618	576	42	SNP	0.885	0.1506	0.2575
HG00378	RNA	ASE	AL	618	576	42	INDEL	0.5555	0.2439	0.3389
HG00378	DNA+RNA	ASE	AL	618	576	42	SNP	0.885	0.1506	0.2575

HG00378	DNA+RNA	ASE	AL	618	576	42	INDEL	0.5555	0.2439	0.3389
HG00378	DNA	RNA-editing	GT	3024	2343	681	SNP	0.9757	0.7036	0.8176
HG00378	DNA	RNA-editing	GT	3024	2343	681	INDEL	0.8404	0.4488	0.5851
HG00378	DNA	RNA-editing	AL	3024	2343	681	SNP	0.9294	0.5252	0.6712
HG00378	DNA	RNA-editing	AL	3024	2343	681	INDEL	0.6385	0.1527	0.2465
HG00378	RNA	RNA-editing	GT	797	383	414	SNP	0.1904	0.0113	0.0213
HG00378	RNA	RNA-editing	GT	797	383	414	INDEL	0.3	0.0075	0.0148
HG00378	RNA	RNA-editing	AL	797	383	414	SNP	0.7125	0.154	0.2533
HG00378	RNA	RNA-editing	AL	797	383	414	INDEL	0.5526	0.0502	0.0921
HG00378	DNA+RNA	RNA-editing	AL	3031	2348	683	SNP	0.9302	0.5248	0.671
HG00378	DNA+RNA	RNA-editing	AL	3031	2348	683	INDEL	0.6385	0.1523	0.2459
HG00378	DNA	RNA-rescue	GT	137	74	63	SNP	0.4489	0.8461	0.5866
HG00378	DNA	RNA-rescue	GT	137	74	63	INDEL	0.375	0.4285	0.4
HG00378	DNA	RNA-rescue	AL	137	74	63	SNP	0.8709	0.7605	0.812
HG00378	DNA	RNA-rescue	AL	137	74	63	INDEL	0.4285	0.1607	0.2337
HG00378	RNA	RNA-rescue	GT	144	81	63	SNP	1	0.6794	0.8091
HG00378	RNA	RNA-rescue	GT	144	81	63	INDEL	0.375	0.0526	0.0923
HG00378	RNA	RNA-rescue	AL	144	81	63	SNP	0.8787	0.7341	0.8
HG00378	RNA	RNA-rescue	AL	144	81	63	INDEL	0.4285	0.1578	0.2307
HG00378	DNA+RNA	RNA-rescue	AL	144	81	63	SNP	0.8787	0.7341	0.8
HG00378	DNA+RNA	RNA-rescue	AL	144	81	63	INDEL	0.4285	0.1578	0.2307
HG00145	DNA	ALL	GT	179773	161639	18134	SNP	0.991	0.775	0.8698
HG00145	DNA	ALL	GT	179773	161639	18134	INDEL	0.9014	0.5819	0.7072
HG00145	DNA	ALL	AL	187864	168683	19181	SNP	0.9737	0.7842	0.8687
HG00145	DNA	ALL	AL	187864	168683	19181	INDEL	0.7403	0.6165	0.6727
HG00145	RNA	ALL	GT	20229	18203	2026	SNP	0.9953	0.971	0.983
HG00145	RNA	ALL	GT	20229	18203	2026	INDEL	0.9245	0.6591	0.7696
HG00145	RNA	ALL	AL	30022	27278	2744	SNP	0.981	0.8899	0.9332
HG00145	RNA	ALL	AL	30022	27278	2744	INDEL	0.8203	0.6919	0.7506
HG00145	DNA+RNA	ALL	AL	192533	172609	19924	SNP	0.9739	0.7889	0.8717
HG00145	DNA+RNA	ALL	AL	192533	172609	19924	INDEL	0.7437	0.6211	0.6769
HG00145	DNA	Strong-evidence	GT	8486	8208	278	SNP	0.9987	0.9887	0.9937

HG00145	DNA	Strong-evidence	GT	8486	8208	278	INDEL	0.9699	0.8129	0.8845
HG00145	DNA	Strong-evidence	AL	8486	8208	278	SNP	0.9864	0.9903	0.9884
HG00145	DNA	Strong-evidence	AL	8486	8208	278	INDEL	0.8297	0.8273	0.8285
HG00145	RNA	Strong-evidence	GT	8486	8208	278	SNP	0.9987	0.9887	0.9937
HG00145	RNA	Strong-evidence	GT	8486	8208	278	INDEL	0.9699	0.8129	0.8845
HG00145	RNA	Strong-evidence	AL	8486	8208	278	SNP	0.9864	0.9903	0.9884
HG00145	RNA	Strong-evidence	AL	8486	8208	278	INDEL	0.8297	0.8273	0.8285
HG00145	DNA+RNA	Strong-evidence	AL	8486	8208	278	SNP	0.9864	0.9903	0.9884
HG00145	DNA+RNA	Strong-evidence	AL	8486	8208	278	INDEL	0.8297	0.8273	0.8285
HG00145	DNA	DNA-only	GT	171287	153431	17856	SNP	0.9904	0.7624	0.8616
HG00145	DNA	DNA-only	GT	171287	153431	17856	INDEL	0.8999	0.5781	0.7039
HG00145	DNA	DNA-only	AL	171287	153431	17856	SNP	0.9724	0.7711	0.8601
HG00145	DNA	DNA-only	AL	171287	153431	17856	INDEL	0.7339	0.6123	0.6676
HG00145	RNA	DNA-only	GT	8779	8102	677	SNP	0.4539	0.6653	0.5396
HG00145	RNA	DNA-only	GT	8779	8102	677	INDEL	0.2897	0.5428	0.3778
HG00145	RNA	DNA-only	AL	8779	8102	677	SNP	0.9754	0.8175	0.8895
HG00145	RNA	DNA-only	AL	8779	8102	677	INDEL	0.7953	0.6838	0.7354
HG00145	DNA+RNA	DNA-only	AL	171288	153431	17857	SNP	0.9724	0.7711	0.8601
HG00145	DNA+RNA	DNA-only	AL	171288	153431	17857	INDEL	0.7338	0.6122	0.6675
HG00145	DNA	RNA-only	GT	6921	5961	960	SNP	0.5547	0.7734	0.646
HG00145	DNA	RNA-only	GT	6921	5961	960	INDEL	0.469	0.7017	0.5622
HG00145	DNA	RNA-only	AL	6921	5961	960	SNP	0.9816	0.9174	0.9484
HG00145	DNA	RNA-only	AL	6921	5961	960	INDEL	0.8339	0.6474	0.729
HG00145	RNA	RNA-only	GT	11583	9883	1700	SNP	0.9926	0.959	0.9755
HG00145	RNA	RNA-only	GT	11583	9883	1700	INDEL	0.9189	0.6454	0.7582
HG00145	RNA	RNA-only	AL	11583	9883	1700	SNP	0.9819	0.9398	0.9604
HG00145	RNA	RNA-only	AL	11583	9883	1700	INDEL	0.8309	0.6867	0.752
HG00145	DNA+RNA	RNA-only	AL	11583	9883	1700	SNP	0.9819	0.9398	0.9604
HG00145	DNA+RNA	RNA-only	AL	11583	9883	1700	INDEL	0.8309	0.6867	0.752
HG00145	DNA	ASE	GT	1016	975	41	SNP	0.9902	0.1079	0.1946
HG00145	DNA	ASE	GT	1016	975	41	INDEL	0.923	0.3333	0.4897
HG00145	DNA	ASE	AL	1016	975	41	SNP	0.8702	0.1207	0.212

HG00145	DNA	ASE	AL	1016	975	41	INDEL	0.8333	0.4166	0.5555
HG00145	RNA	ASE	GT	1014	973	41	SNP	0.0097	0.0142	0.0115
HG00145	RNA	ASE	GT	1014	973	41	INDEL	0	0	0
HG00145	RNA	ASE	AL	1014	973	41	SNP	0.8625	0.1207	0.2118
HG00145	RNA	ASE	AL	1014	973	41	INDEL	0.8333	0.4166	0.5555
HG00145	DNA+RNA	ASE	AL	1016	975	41	SNP	0.8702	0.1207	0.212
HG00145	DNA+RNA	ASE	AL	1016	975	41	INDEL	0.8333	0.4166	0.5555
HG00145	DNA	RNA-editing	GT	3938	3435	503	SNP	0.9775	0.6083	0.7499
HG00145	DNA	RNA-editing	GT	3938	3435	503	INDEL	0.8061	0.4876	0.6076
HG00145	DNA	RNA-editing	AL	3938	3435	503	SNP	0.8806	0.4539	0.5991
HG00145	DNA	RNA-editing	AL	3938	3435	503	INDEL	0.6666	0.2114	0.321
HG00145	RNA	RNA-editing	GT	819	573	246	SNP	0.0952	0.0049	0.0094
HG00145	RNA	RNA-editing	GT	819	573	246	INDEL	0.3	0.0182	0.0344
HG00145	RNA	RNA-editing	AL	819	573	246	SNP	0.6507	0.1459	0.2383
HG00145	RNA	RNA-editing	AL	819	573	246	INDEL	0.3333	0.0391	0.07
HG00145	DNA+RNA	RNA-editing	AL	3940	3435	505	SNP	0.8806	0.4539	0.5991
HG00145	DNA+RNA	RNA-editing	AL	3940	3435	505	INDEL	0.6598	0.2092	0.3177
HG00145	DNA	RNA-rescue	GT	154	108	46	SNP	0.3472	0.862	0.495
HG00145	DNA	RNA-rescue	GT	154	108	46	INDEL	0.2307	0.3	0.2608
HG00145	DNA	RNA-rescue	AL	154	108	46	SNP	0.9605	0.6886	0.8021
HG00145	DNA	RNA-rescue	AL	154	108	46	INDEL	0.7037	0.4047	0.5139
HG00145	RNA	RNA-rescue	GT	160	112	48	SNP	0.9605	0.6759	0.7934
HG00145	RNA	RNA-rescue	GT	160	112	48	INDEL	0.6428	0.2142	0.3214
HG00145	RNA	RNA-rescue	AL	160	112	48	SNP	0.9625	0.7	0.8105
HG00145	RNA	RNA-rescue	AL	160	112	48	INDEL	0.7142	0.409	0.5202
HG00145	DNA+RNA	RNA-rescue	AL	160	112	48	SNP	0.9625	0.7	0.8105
HG00145	DNA+RNA	RNA-rescue	AL	160	112	48	INDEL	0.7142	0.409	0.5202
NA20509	DNA	ALL	GT	96406	87806	8600	SNP	0.9917	0.8763	0.9304
NA20509	DNA	ALL	GT	96406	87806	8600	INDEL	0.9203	0.6641	0.7715
NA20509	DNA	ALL	AL	101784	92342	9442	SNP	0.9756	0.876	0.9232
NA20509	DNA	ALL	AL	101784	92342	9442	INDEL	0.7442	0.6716	0.706
NA20509	RNA	ALL	GT	32702	28109	4593	SNP	0.9917	0.9497	0.9702

NA20509	RNA	ALL	GT	32702	28109	4593	INDEL	0.9097	0.4701	0.6199
NA20509	RNA	ALL	AL	42550	37076	5474	SNP	0.9797	0.9386	0.9587
NA20509	RNA	ALL	AL	42550	37076	5474	INDEL	0.7924	0.5501	0.6494
NA20509	DNA+RNA	ALL	AL	119228	106337	12891	SNP	0.976	0.8862	0.929
NA20509	DNA+RNA	ALL	AL	119228	106337	12891	INDEL	0.7541	0.6285	0.6856
NA20509	DNA	Strong-evidence	GT	10393	10040	353	SNP	0.9982	0.9852	0.9916
NA20509	DNA	Strong-evidence	GT	10393	10040	353	INDEL	0.9559	0.808	0.8757
NA20509	DNA	Strong-evidence	AL	10393	10040	353	SNP	0.9855	0.989	0.9872
NA20509	DNA	Strong-evidence	AL	10393	10040	353	INDEL	0.8149	0.828	0.8214
NA20509	RNA	Strong-evidence	GT	10393	10040	353	SNP	0.9982	0.9852	0.9916
NA20509	RNA	Strong-evidence	GT	10393	10040	353	INDEL	0.9559	0.808	0.8757
NA20509	RNA	Strong-evidence	AL	10393	10040	353	SNP	0.9855	0.989	0.9872
NA20509	RNA	Strong-evidence	AL	10393	10040	353	INDEL	0.8149	0.828	0.8214
NA20509	DNA+RNA	Strong-evidence	AL	10393	10040	353	SNP	0.9855	0.989	0.9872
NA20509	DNA+RNA	Strong-evidence	AL	10393	10040	353	INDEL	0.8149	0.828	0.8214
NA20509	DNA	DNA-only	GT	86013	77766	8247	SNP	0.9906	0.8618	0.9217
NA20509	DNA	DNA-only	GT	86013	77766	8247	INDEL	0.9185	0.6576	0.7665
NA20509	DNA	DNA-only	AL	86013	77766	8247	SNP	0.9742	0.862	0.9147
NA20509	DNA	DNA-only	AL	86013	77766	8247	INDEL	0.7387	0.6893	0.7131
NA20509	RNA	DNA-only	GT	9337	8506	831	SNP	0.4672	0.7151	0.5652
NA20509	RNA	DNA-only	GT	9337	8506	831	INDEL	0.3814	0.6036	0.4674
NA20509	RNA	DNA-only	AL	9337	8506	831	SNP	0.9764	0.8965	0.9347
NA20509	RNA	DNA-only	AL	9337	8506	831	INDEL	0.7808	0.7123	0.745
NA20509	DNA+RNA	DNA-only	AL	86014	77767	8247	SNP	0.9742	0.862	0.9147
NA20509	DNA+RNA	DNA-only	AL	86014	77767	8247	INDEL	0.7387	0.6893	0.7131
NA20509	DNA	RNA-only	GT	4705	3959	746	SNP	0.4991	0.7734	0.6067
NA20509	DNA	RNA-only	GT	4705	3959	746	INDEL	0.4349	0.6313	0.515
NA20509	DNA	RNA-only	AL	4705	3959	746	SNP	0.9793	0.9161	0.9467
NA20509	DNA	RNA-only	AL	4705	3959	746	INDEL	0.8014	0.4453	0.5725
NA20509	RNA	RNA-only	GT	22135	17941	4194	SNP	0.9879	0.9306	0.9584
NA20509	RNA	RNA-only	GT	22135	17941	4194	INDEL	0.9049	0.4453	0.5969
NA20509	RNA	RNA-only	AL	22135	17941	4194	SNP	0.9786	0.9443	0.9612

NA20509	RNA	RNA-only	AL	22135	17941	4194	INDEL	0.7955	0.499	0.6133
NA20509	DNA+RNA	RNA-only	AL	22135	17941	4194	SNP	0.9786	0.9443	0.9612
NA20509	DNA+RNA	RNA-only	AL	22135	17941	4194	INDEL	0.7955	0.499	0.6133
NA20509	DNA	ASE	GT	512	461	51	SNP	0.9041	0.3296	0.4832
NA20509	DNA	ASE	GT	512	461	51	INDEL	0.7142	0.2083	0.3225
NA20509	DNA	ASE	AL	512	461	51	SNP	0.9264	0.4126	0.5709
NA20509	DNA	ASE	AL	512	461	51	INDEL	0.8181	0.3541	0.4943
NA20509	RNA	ASE	GT	511	461	50	SNP	0.0598	0.0943	0.0732
NA20509	RNA	ASE	GT	511	461	50	INDEL	0.1428	0.2222	0.1739
NA20509	RNA	ASE	AL	511	461	50	SNP	0.9264	0.4126	0.5709
NA20509	RNA	ASE	AL	511	461	50	INDEL	0.8181	0.3541	0.4943
NA20509	DNA+RNA	ASE	AL	512	461	51	SNP	0.9264	0.4126	0.5709
NA20509	DNA+RNA	ASE	AL	512	461	51	INDEL	0.8181	0.3541	0.4943
NA20509	DNA	RNA-editing	GT	5278	4013	1265	SNP	0.9697	0.6527	0.7803
NA20509	DNA	RNA-editing	GT	5278	4013	1265	INDEL	0.8992	0.472	0.619
NA20509	DNA	RNA-editing	AL	5278	4013	1265	SNP	0.9009	0.4739	0.6211
NA20509	DNA	RNA-editing	AL	5278	4013	1265	INDEL	0.7031	0.1238	0.2106
NA20509	RNA	RNA-editing	GT	1580	734	846	SNP	0.2727	0.0166	0.0314
NA20509	RNA	RNA-editing	GT	1580	734	846	INDEL	0	0	0
NA20509	RNA	RNA-editing	AL	1580	734	846	SNP	0.6645	0.1426	0.2348
NA20509	RNA	RNA-editing	AL	1580	734	846	INDEL	0.5744	0.0284	0.0541
NA20509	DNA+RNA	RNA-editing	AL	5284	4018	1266	SNP	0.8999	0.4731	0.6202
NA20509	DNA+RNA	RNA-editing	AL	5284	4018	1266	INDEL	0.7031	0.1238	0.2106
NA20509	DNA	RNA-rescue	GT	161	116	45	SNP	0.4479	0.6231	0.5212
NA20509	DNA	RNA-rescue	GT	161	116	45	INDEL	0.3	0.3	0.3
NA20509	DNA	RNA-rescue	AL	161	116	45	SNP	0.9411	0.8362	0.8855
NA20509	DNA	RNA-rescue	AL	161	116	45	INDEL	0.5142	0.2888	0.3699
NA20509	RNA	RNA-rescue	GT	174	128	46	SNP	0.9813	0.8203	0.8936
NA20509	RNA	RNA-rescue	GT	174	128	46	INDEL	0.5454	0.1304	0.2105
NA20509	RNA	RNA-rescue	AL	174	128	46	SNP	0.9469	0.8437	0.8923
NA20509	RNA	RNA-rescue	AL	174	128	46	INDEL	0.5277	0.3043	0.386
NA20509	DNA+RNA	RNA-rescue	AL	174	128	46	SNP	0.9469	0.8437	0.8923

NA20509	DNA+RNA	RNA-rescue	AL	174	128	46	INDEL	0.5277	0.3043	0.386
HG00342	DNA	ALL	GT	129642	112838	16804	SNP	0.9935	0.8723	0.929
HG00342	DNA	ALL	GT	129642	112838	16804	INDEL	0.8995	0.5826	0.7072
HG00342	DNA	ALL	AL	134542	117182	17360	SNP	0.9742	0.869	0.9186
HG00342	DNA	ALL	AL	134542	117182	17360	INDEL	0.7324	0.617	0.6698
HG00342	RNA	ALL	GT	22708	20502	2206	SNP	0.9942	0.952	0.9726
HG00342	RNA	ALL	GT	22708	20502	2206	INDEL	0.9285	0.6485	0.7636
HG00342	RNA	ALL	AL	31556	28478	3078	SNP	0.9795	0.9102	0.9436
HG00342	RNA	ALL	AL	31556	28478	3078	INDEL	0.8149	0.6772	0.7397
HG00342	DNA+RNA	ALL	AL	144971	126229	18742	SNP	0.9748	0.875	0.9222
HG00342	DNA+RNA	ALL	AL	144971	126229	18742	INDEL	0.7397	0.6229	0.6763
HG00342	DNA	Strong-evidence	GT	8184	7865	319	SNP	0.9987	0.9905	0.9946
HG00342	DNA	Strong-evidence	GT	8184	7865	319	INDEL	0.9691	0.7911	0.8711
HG00342	DNA	Strong-evidence	AL	8184	7865	319	SNP	0.9839	0.9917	0.9878
HG00342	DNA	Strong-evidence	AL	8184	7865	319	INDEL	0.8493	0.8132	0.8309
HG00342	RNA	Strong-evidence	GT	8184	7865	319	SNP	0.9987	0.9905	0.9946
HG00342	RNA	Strong-evidence	GT	8184	7865	319	INDEL	0.9691	0.7911	0.8711
HG00342	RNA	Strong-evidence	AL	8184	7865	319	SNP	0.9839	0.9917	0.9878
HG00342	RNA	Strong-evidence	AL	8184	7865	319	INDEL	0.8493	0.8132	0.8309
HG00342	DNA+RNA	Strong-evidence	AL	8184	7865	319	SNP	0.9839	0.9917	0.9878
HG00342	DNA+RNA	Strong-evidence	AL	8184	7865	319	INDEL	0.8493	0.8132	0.8309
HG00342	DNA	DNA-only	GT	121458	104973	16485	SNP	0.9931	0.8632	0.9236
HG00342	DNA	DNA-only	GT	121458	104973	16485	INDEL	0.8976	0.5785	0.7036
HG00342	DNA	DNA-only	AL	121458	104973	16485	SNP	0.9735	0.8642	0.9156
HG00342	DNA	DNA-only	AL	121458	104973	16485	INDEL	0.728	0.614	0.6662
HG00342	RNA	DNA-only	GT	8051	7227	824	SNP	0.4447	0.7027	0.5447
HG00342	RNA	DNA-only	GT	8051	7227	824	INDEL	0.2977	0.574	0.3921
HG00342	RNA	DNA-only	AL	8051	7227	824	SNP	0.9766	0.8579	0.9134
HG00342	RNA	DNA-only	AL	8051	7227	824	INDEL	0.7745	0.6531	0.7086
HG00342	DNA+RNA	DNA-only	AL	121465	104978	16487	SNP	0.9735	0.8642	0.9156
HG00342	DNA+RNA	DNA-only	AL	121465	104978	16487	INDEL	0.728	0.6139	0.6661
HG00342	DNA	RNA-only	GT	3988	3519	469	SNP	0.4738	0.7649	0.5851

HG00342	DNA	RNA-only	GT	3988	3519	469	INDEL	0.4041	0.6666	0.5031
HG00342	DNA	RNA-only	AL	3988	3519	469	SNP	0.9707	0.8822	0.9244
HG00342	DNA	RNA-only	AL	3988	3519	469	INDEL	0.8111	0.6381	0.7143
HG00342	RNA	RNA-only	GT	14404	12557	1847	SNP	0.9914	0.9285	0.9589
HG00342	RNA	RNA-only	GT	14404	12557	1847	INDEL	0.922	0.6344	0.7516
HG00342	RNA	RNA-only	AL	14404	12557	1847	SNP	0.9785	0.9324	0.9549
HG00342	RNA	RNA-only	AL	14404	12557	1847	INDEL	0.8308	0.6818	0.749
HG00342	DNA+RNA	RNA-only	AL	14405	12557	1848	SNP	0.9785	0.9324	0.9549
HG00342	DNA+RNA	RNA-only	AL	14405	12557	1848	INDEL	0.8308	0.6818	0.749
HG00342	DNA	ASE	GT	797	749	48	SNP	0.9823	0.1541	0.2665
HG00342	DNA	ASE	GT	797	749	48	INDEL	0.7692	0.2272	0.3508
HG00342	DNA	ASE	AL	797	749	48	SNP	0.9492	0.1833	0.3073
HG00342	DNA	ASE	AL	797	749	48	INDEL	0.6785	0.3409	0.4538
HG00342	RNA	ASE	GT	797	749	48	SNP	0.0176	0.0416	0.0248
HG00342	RNA	ASE	GT	797	749	48	INDEL	0.0769	0.25	0.1176
HG00342	RNA	ASE	AL	797	749	48	SNP	0.9492	0.1833	0.3073
HG00342	RNA	ASE	AL	797	749	48	INDEL	0.6785	0.3409	0.4538
HG00342	DNA+RNA	ASE	AL	797	749	48	SNP	0.9492	0.1833	0.3073
HG00342	DNA+RNA	ASE	AL	797	749	48	INDEL	0.6785	0.3409	0.4538
HG00342	DNA	RNA-editing	GT	3186	2752	434	SNP	0.9779	0.6605	0.7885
HG00342	DNA	RNA-editing	GT	3186	2752	434	INDEL	0.8666	0.423	0.5685
HG00342	DNA	RNA-editing	AL	3186	2752	434	SNP	0.9412	0.5044	0.6568
HG00342	DNA	RNA-editing	AL	3186	2752	434	INDEL	0.5755	0.194	0.2902
HG00342	RNA	RNA-editing	GT	606	404	202	SNP	0.1153	0.0128	0.0231
HG00342	RNA	RNA-editing	GT	606	404	202	INDEL	0	0	0
HG00342	RNA	RNA-editing	AL	606	404	202	SNP	0.8028	0.1548	0.2596
HG00342	RNA	RNA-editing	AL	606	404	202	INDEL	0.4	0.034	0.0628
HG00342	DNA+RNA	RNA-editing	AL	3191	2757	434	SNP	0.9412	0.5034	0.656
HG00342	DNA+RNA	RNA-editing	AL	3191	2757	434	INDEL	0.5755	0.194	0.2902
HG00342	DNA	RNA-rescue	GT	115	76	39	SNP	0.3396	0.72	0.4615
HG00342	DNA	RNA-rescue	GT	115	76	39	INDEL	0.625	0.625	0.625
HG00342	DNA	RNA-rescue	AL	115	76	39	SNP	0.9636	0.7794	0.8617

HG00342	DNA	RNA-rescue	AL	115	76	39	INDEL	0.5294	0.2571	0.3461
HG00342	RNA	RNA-rescue	GT	120	80	40	SNP	0.9473	0.7605	0.8437
HG00342	RNA	RNA-rescue	GT	120	80	40	INDEL	0.6666	0.1621	0.2608
HG00342	RNA	RNA-rescue	AL	120	80	40	SNP	0.9661	0.76	0.8507
HG00342	RNA	RNA-rescue	AL	120	80	40	INDEL	0.5555	0.2702	0.3636
HG00342	DNA+RNA	RNA-rescue	AL	120	80	40	SNP	0.9661	0.76	0.8507
HG00342	DNA+RNA	RNA-rescue	AL	120	80	40	INDEL	0.5555	0.2702	0.3636
NA12878	WES calling	ALL	GT	259692	224840	34852	SNP	0.9966	0.9659	0.981
NA12878	WES calling	ALL	GT	259692	224840	34852	INDEL	0.9569	0.4182	0.582
NA12878	WES calling	ALL	AL	259692	224840	34852	SNP	0.9997	0.9692	0.9842
NA12878	WES calling	ALL	AL	259692	224840	34852	INDEL	0.9895	0.4338	0.6032
HG00171	WES calling	ALL	GT	130068	113298	16770	SNP	0.9936	0.8845	0.9359
HG00171	WES calling	ALL	GT	130068	113298	16770	INDEL	0.896	0.5843	0.7073
HG00171	WES calling	ALL	AL	130068	113298	16770	SNP	0.9752	0.8931	0.9323
HG00171	WES calling	ALL	AL	130068	113298	16770	INDEL	0.7329	0.6266	0.6756
HG00378	WES calling	ALL	GT	133523	116342	17181	SNP	0.9934	0.8772	0.9317
HG00378	WES calling	ALL	GT	133523	116342	17181	INDEL	0.8899	0.588	0.7081
HG00378	WES calling	ALL	AL	133523	116342	17181	SNP	0.9748	0.886	0.9283
HG00378	WES calling	ALL	AL	133523	116342	17181	INDEL	0.7306	0.6276	0.6752
HG00145	WES calling	ALL	GT	176389	158968	17421	SNP	0.9908	0.7747	0.8695
HG00145	WES calling	ALL	GT	176389	158968	17421	INDEL	0.9021	0.5965	0.7181
HG00145	WES calling	ALL	AL	176389	158968	17421	SNP	0.9734	0.7893	0.8717
HG00145	WES calling	ALL	AL	176389	158968	17421	INDEL	0.7385	0.6352	0.683
NA20509	WES calling	ALL	GT	97676	89030	8646	SNP	0.9913	0.8722	0.9279
NA20509	WES calling	ALL	GT	97676	89030	8646	INDEL	0.919	0.6616	0.7693
NA20509	WES calling	ALL	AL	97676	89030	8646	SNP	0.9757	0.8856	0.9285
NA20509	WES calling	ALL	AL	97676	89030	8646	INDEL	0.7438	0.7027	0.7227
HG00342	WES calling	ALL	GT	129270	113203	16067	SNP	0.9934	0.8681	0.9265
HG00342	WES calling	ALL	GT	129270	113203	16067	INDEL	0.9031	0.6015	0.722
HG00342	WES calling	ALL	AL	129270	113203	16067	SNP	0.9745	0.8774	0.9234
HG00342	WES calling	ALL	AL	129270	113203	16067	INDEL	0.7347	0.64	0.684