Master's thesis

Master's Programme in Computer Science

Automated copy number variation concordance analysis

Purushottam Thapa Magar

June 14, 2021

Faculty of Science

University of Helsinki

Supervisor(s)

Prof. Veli Mäkinen, Mr. Jukka Matilainen, Dr. Christophe Roos

Examiner(s)

Prof. Veli Mäkinen, Dr. Christophe Roos

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

The virtues of science are skepticism and independence of thought.

Walter Gilbert

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Master's Programme in Computer Science

Tekijä — Författare — Author

Purushottam Thapa Magar

Työn nimi — Arbetets titel — Title

Automated copy number variation concordance analysis

Ohjaajat — Handledare — Supervisors

Prof. Veli Mäkinen, Mr. Jukka Matilainen, Dr. Christophe Roos

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	June 14, 2021	55 pages

Tiivistelmä — Referat — Abstract

Rapid growth and advancement of next generation sequencing (NGS) technologies have changed the landscape of genomic medicine. Today, clinical laboratories perform DNA sequencing on a regular basis, which is an error prone process. Erroneous data affects downstream analysis and produces fallacious result. Therefore, external quality assessment (EQA) of laboratories working with NGS data is crucial. Validation of variations such as single nucleotide polymorphism (SNP) and InDels (<50 bp) is fairly accurate these days. However, detection and quality assessment of large changes such as the copy number variation (CNV) continues to be a concern.

In this work, we aimed to study the feasibility of an automated CNV concordance analysis for the laboratory EQA services. We benchmarked variants reported by 25 laboratories against the highly curated gold standard for the son (HG002/NA24385) of the askenazim trio from the Personal Genome Project published by the Genome in a Bottle Consortium (GIAB). We employed two methods to conduct concordance of CNVs, the sequence based comparison with Truvari and the in-house exome-based comparison. For deletion calls of two whole genome sequencing (WGS) submissions, Truvari gained a value greater than 88% and 68% for precision and recall respectively. Conversely, the in-house method's precision and recall score peaked at 39% and 7.9% respectively for one WGS submission for both deletion and duplication calls. The results indicate that automated CNV concordance analysis of the deletion calls for the WGS-based callset might be feasible with Truvari. On the other hand, results for panel-based targeted sequencing for the deletion calls showed precision and recall rates ranging from 0-80% and 0-5.6% respectively with Truvari. The result suggests that automated concordance analysis of CNVs for targeted sequencing remains a challenge. In conclusion, CNV concordance analysis depends on how the sequence data is generated.

ACM Computing Classification System (CCS)
Applied computing → Life and medical sciences → Computational biology → Computational genomics
General and reference → Cross-computing tools and techniques → Validation

Avainsanat — Nyckelord — Keywords

next generation sequencing, copy number variation, structural variation, benchmarking NGS data

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Algorithms study track

Acknowledgements

In the course of writing this thesis, I have received plenty of help. First of all, I am extremely grateful to my supervisor, Professor Veli Mäkinen - your guidance was invaluable in the completion of this work.

I would like to express my deepest gratitude to my second supervisor Mr. Jukka Matilainen. Thank you for all those brilliant ideas - your expertise was instrumental throughout this work.

I would like to express my deepest appreciation to my third supervisor Dr. Christophe Roos. Thank you for always encouraging and nurturing me.

I would also like to thank my employer Euformatics Oy for allowing me to work on this project. Further, thank you all the colleagues for creating a supportive environment. Special thanks to Jussi Volanen for spontaneous ideas.

Many thanks to EMQN and GenQA. Thank you for collaborating and providing the data for the project.

I am deeply indebted to my friends, Corinna Hertweck and Jyoti Prabha Satta, along with my Rush. Thank you for reading my thesis and providing valuable feedback.

Finally, heartfelt thanks to my family for always being there and to my friends Anil, Ashika, Mangesh, Mina and Keshav for keeping me sane in this difficult times.

Contents

Abbreviations

AS	*de novo* assembly of genome
BAM	Binary alignment map
BED	Browser extensible data
BND	Breakend
CBS	Circular binary segmentation
CCS	Circular consensus sequencing
CFTR	Cystic fibrosis transmembrane conductance regulator
CNV	Copy number variation
DOC	Depth of coverage
DEL	Deletion
DUP	Duplication
EMQN	The European Molecular Genetics Quality Network
EQA	External Quality Assessment
FP	False positive
FN	False negative
GenQA	Genomic Quality Assessment
GIAB	Genome in a Bottle Consortium
HMM	Hidden Markov Model
IGV	Integrative Genomics Viewer
INS	Insertion
kb	Kilobase
NGS	Next generation sequencing
PacBio	Pacific Biosciences

PEM	Paired end mapping
PT	Proficiency testing
RD	Read depth
ROI	Region of interest
SAM	Sequence alignment map
SMRT	Single molecule real-time
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SR	Split read
SV	Structural variant
TP	True positive
TS	Targeted sequencing
VCF	Variant call format
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 Introduction

With the dawn of high-throughput next generation sequencing (NGS, also known as massively parallel sequencing), the field of genomics has produced more data than ever [6]. High speed deep sequencing capabilities of NGS technologies have been proven to detect more *de novo* and rare genetic variants than conventional Sanger sequencing [23]. NGS technologies have made existing analytics much cheaper both in cost and time and new advancements are rapidly widening the scope of applications. However, NGS has certain disadvantages, the fundamental technologies used for calling the DNA bases frequently make errors, and the type of errors depend on the technology [45, 81]. For example, some technologies make sequence-dependent errors, while others simply make random errors independent of the molecule to be sequenced. There are preventive measures such as reading the same genomic region multiple times to obtain the correct read through a majority vote system, but there are also other computational tricks to get more accurate reading. Nevertheless, these sequencing artifacts happen all the time. Moreover, it is not only the sequencing *per se* that introduces error, the sequence quality is heavily dependent on how the sample is handled in the wet-lab before sequencing using NGS technologies [81].

To guarantee the correctness of the end result, laboratories need to perform validation runs in which the concordance of the obtained result is evaluated with the high quality truth set. In particular, clinical diagnostic laboratories must perform well in their concordance analysis to have confidence in their result and to be allowed to provide their diagnostic services [43]. However, performing well in an internal assessment is not enough. It is a standard practice for laboratories to have an external quality assessment (EQA), also known as proficiency testing (PT) from an EQA service providers to ascertain the quality of their result. The EQA is a method where an external party examines the laboratory result without prejudice [75]. It is not a one-time process, but rather a recurring one, in which laboratories must have their results repeatedly verified by the EQA providers in certain intervals. The key purpose of EQA is to standardize laboratory practices to guarantee quality and uniformity of result among laboratories. [78, 75]

For genetic testing, EQA service providers would benefit from an automated system to conduct a concordance analysis because, typically, there are thousands of genomic varia-

tions to validate, which is not practical to do manually. Genomic (or genetic) variations are changes in a person's DNA. Variations can be enormously diverse. They can range from single nucleotide variants (or polymorphisms, SNV / SNP) in the DNA, to large structural variants (SV) affecting hundreds of base-pairs or more, or even whole section of a chromosome (millions of base-pairs). The large SVs are much more complex from both the detection and interpretation point of view. [13] Copy number variation (CNV) is one type of SV in which a piece of DNA longer than 50 bp is either added or deleted. CNVs are clinically important because they have been linked with several genetic diseases including cancer [46]. Automatic concordance analysis of changes such as SNPs and small (<50 bp) insertions or deletions (InDels) is fairly accurate these days [73, 69]. However, comparing sequences that are thousands of bases long is a computationally hard problem, thus making automated concordance analysis difficult for CNVs. In this work, we aim to study the practicality of automating CNV concordance analysis.

There are 8 chapters in this thesis. Chapter 2 reviews the relevant literature about NGS in general. The reviewing continues in Chapter 3 where the details of the CNV detection methods are explained. Then, Chapter 4 describes the datasets used in the study. Chapter 5 explains the two benchmarking methods (sequence based and in-house exome-based comparison) used, results obtained are then presented in Chapter 6. In Chapter 7, key findings of the study along with the acknowledgement of the limitations faced during the study is discussed. At last, conclusions and possible directions for the future work are presented.

2 Next Generation Sequencing (NGS)

Next generation sequencing is a successor to Sanger sequencing that ruled the DNA sequencing domain together with Maxam-Gilbert sequencing [1] for more than two decades. NGS refers to the collection of technologies that were developed after Sanger that can carry out deep parallel sequencing on a massive scale. Due to its high-throughput capabilities, the cost and time of genomic sequencing were tremendously reduced compared to its predecessors. [39] To put this improvement into perspective, the yearly throughput of NGS as of today is in petabases while Sanger sequencing would yield just a fraction of this [83]. The cost of sequencing has also dropped significantly over the last 20 years [44]. Reduced time and cost as well as a much more comprehensive understanding of the genome (human, but also many other species) has enabled almost any laboratory or research groups to use NGS in their routine work. Additionally, NGS users have developed new, innovative and creative ways to make use of NGS technologies, which have lead to an explosion of new scientific and analytical approaches in the field. Now, the NGS and technologies built around it are evolving quickly.

Currently, NGS is used extensively in whole genome sequencing (WGS), which is improving our understanding of the genome. Additionally, such improved comprehensive knowledge will bring more crucial biological understanding and interpretation of variations. Apart from WGS, NGS is used for targeted and whole exome sequencing (WES), which are even faster and cheaper than WGS and can still provide invaluable information on targeted regions of the genome. [61] For instance, Iossifov et al. [24] detected CNVs, *de novo* missense variants and *de novo* gene-disrupting variants responsible for autism spectrum disorder (ASD) with WES.

Today, NGS is not just a high-tech innovation that only a few select people have access to. Its application has spread in domains that we could not even imagine a few years ago. It is used in forensic sciences but also in veterinary sciences, agriculture, and most recently for tracking the evolution of the SARS-CoV-2 virus and for following the Covid-19 pandemic [80]. Specially in clinical genomics, its use is rather a necessity now and the technology is very widespread. One of the key clinical areas where NGS can shine is in the field of precision medicine, where precise information about the patient genome can help instruct the development of tailored treatments [42]. Clinicians are also looking into

the possibilities of how NGS can be incorporated to obtain a necessary biological insight to develop a new clinical practice [10].

2.1 NGS platforms

DNA sequencing is a technically challenging task. Currently, several different NGS platforms for sequencing exist. Following the arrival of Illumina in 2006 [30], several other technologies with different sequencing methods have emerged. However, Illumina remains the most popular sequencing technology among the sequencing service providers [48, 7]. Methodologies used by some popular NGS platforms to call DNA bases are discussed here.

There are some fundamental differences in how different sequencing platforms work. Firstly, all of these technologies require platform-specific library preparation [48]. Secondly, different sequencing technologies determine nucleotides in a sequence by applying different methods. For example, Illumina is an imaging-based technology that uses fluorescence signals to determine bases because each DNA nucleotide has a unique fluorescence signature. On the other hand, Ion Torrent technology is based on a semiconductor chip that detects the change in the pH value resulting from a release of H^+ ion during the incorporation of a DNA nucleotide into a single strand of DNA. [18]

Pacific Biosciences (PacBio) technology's single molecule real-time (SMRT) sequencing has become popular in recent years. Similar to Illumina, it also uses a fluorescence signal of DNA nucleotides to call bases. However, PacBio's SMRT chips measure emitted fluorescence signal in real time, which gives it the capability to detect a single molecule. [18] Oxford Nanopore sequencing is another sequencing platform that is gaining ground in recent years. It measures the change in ionic current resulting from the passing of different DNA bases through a protein nanopore. The electrical signal changes because the size of each DNA base is different. [51]

2.1.1 Short vs long-read sequencing

Apart from how different NGS platforms perform base calling, they can be broadly classified into short-read and long-read based sequencing platforms. The long-read technologies are newer than the short-read ones and are sometimes referred to as the third generation sequencing technology. The main difference between the short-read and the long-read technologies is the length of reads produced by them. As the name suggests, the short-read

and the long-read platforms produce shorter and longer reads respectively. The Illumina and the Ion Torrent platforms are short-read based, whereas the PacBio and the Oxford Nanopore are long-read based sequencing platforms. [8]

The read length produced by the short-read technologies is limited by the short length of the DNA fragments used during the sequencing process, which is usually in a range of 75 to 400 base pairs [8]. Conversely, long-read technologies are capable of sequencing thousands of contiguous bases because they have the ability to identify a single molecule of DNA in real time. The long-read platforms have some clear advantages over the short-read platforms. For instance, the long-read technologies are able to sequence repetitive regions, which have been troublesome for short-read technologies. Furthermore, they can resolve sequences in structural variants better than the short-read technologies. Similarly, long reads are more useful than short reads in the *de novo* genome assembly process because they align better with the reference genome. However, long-read technologies are considerably more expensive than the short-read ones. Moreover, accuracy per read of short-read platforms are higher than that of the long-read platforms. [8] Currently, compared to the long-read sequencing, the existing bioinformatics infrastructure is more favorable to the short-read sequencing platforms [65]. However, this could change in the future when the use of the more advanced long-read technologies are widely adopted.

2.2 Types of NGS sequencing

Based on the genomic region of interest (ROI) to sequence, NGS sequencing can be divided into two main types, namely targeted sequencing (TS) and whole genome sequencing (WGS). We discuss the key ideas behind each strategy in the following section.

2.2.1 Targeted sequencing

Targeted sequencing is a technique in which the sequencing of the DNA is restricted to the predefined regions of the genome. The decision on ROIs for sequencing is based on the genetic testing being carried out. [4] For instance, BRAF and EGFR genes are frequently targeted because they are known to contain mutation hotspots for various cancer types [14]. The sequencing cost of TS is much cheaper than WGS. Moreover, greater sequencing depth achieved in TS allows the detection of more genomic variation. [4, 14] TS can be further divided into two types: (i) gene panel based sequencing, and (ii) whole exome

sequencing [4].

Gene panel based sequencing

In gene panel based sequencing, the handful of genes that are known to have clinical significance relating to patient's phenotype or indicated by some other medical procedures are sequenced. The genes or ROIs under study are enriched before the sequencing. The list of selected genes for testing are termed as a 'gene panel'. These gene panels can be either purchased or custom made. The volume of data produced by gene panel sequencing is small, which saves computational resources during analysis. [4, 14]

Whole Exome Sequencing

Exomes or protein-coding regions cover less than 2 percent of the human genome [39]. However, the majority of the Mendelian diseases are linked to mutations in the protein sequence [4]. Therefore, it makes sense for the clinical laboratories to conduct WES for disease diagnosis. In WES, all the known exons in the genome are sequenced, which is roughly around 22,000 protein coding genes [4].

2.2.2 Whole Genome Sequencing

WGS involves the sequencing of the entire DNA of the organism. For the human genome, WGS means the sequencing of 23 pairs of chromosomes along with the mitochondrial DNA. WGS produces lots of genomic data, which require intensive computational resources for analysis. Furthermore, the cost of WGS is higher compared to TS, which limits its use. [4]

2.3 Variant Calling

DNA sequencing is a complicated task that comprises multiple steps between sample collection and generation of a list of variants in a digital file such as the variant call format (VCF) [84]. In a typical clinical setting, a doctor examines a patient and orders a gene analysis (sequencing) job. Then, a laboratory takes a sample from the patient, for example, some drops of blood, and stores it appropriately. After that, DNA is extracted from the sample. The DNA is then used to prepare a library of short molecules to analyse with the sequencer, and the library is run through a high-throughput NGS sequencer. Depending upon the NGS platform being used, the most 'raw' data produced by the sequencer is either a set of images (Illumina platform) or electric signals (Ion Torrent, Oxford Nanopore) [31]. During the process of base-calling, that raw data is converted into a human as well as computer readable sequence data, usually produced as a FASTQ file. The FASTQ file is then forwarded for computational analysis. This part of the data processing is referred as the secondary analysis, while the sequencing *per se* was the primary analysis. The secondary analysis consists of two major steps: (i) aligning the obtained sequence to the reference genome and (ii) calling the variants, i.e. the positions in the DNA sequence that differ from the reference. After this secondary computational processing, the output is a list of variants as compared to the reference genome used. VCF files are normally used for reporting variants, which are then given back to a pathologist or molecular biologist for interpretation (tertiary analysis). The interpretation is then passed on to the doctor who ordered the gene test. Finally, the doctor uses the provided

interpretation of observed variants to take clinical actions accordingly. [84] Figure 2.1 depicts the typical variant calling process.

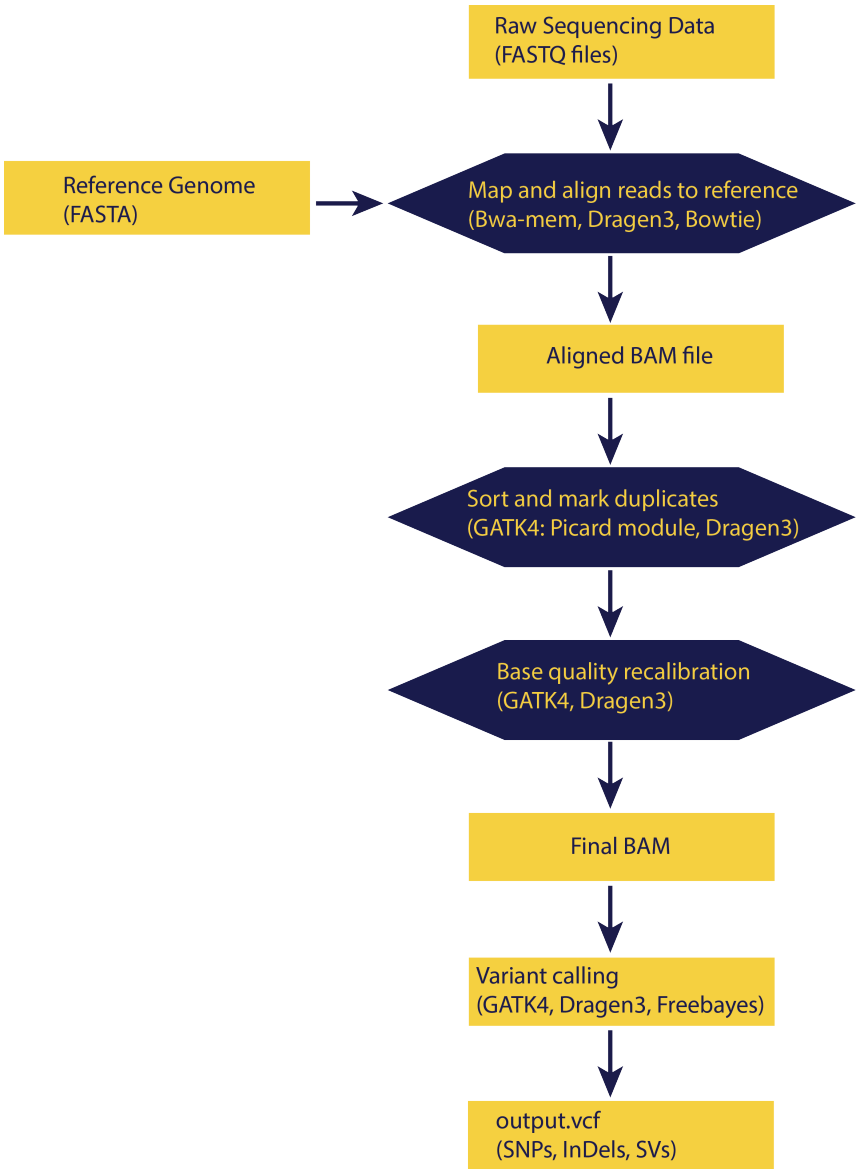

Figure 2.1: A typical workflow of variant calling pipeline.(Adapted from [63, 38])

In a clinical context, biological interpretation of the calls produced by the variant calling pipeline during the secondary analysis provides the most important piece of information to the medical practitioner. Hence, the accuracy of these calls is very important because the patient's healthcare plan will be affected by it. [41]

2.4 Causes of erroneous NGS data

The success of any genomic analysis depends on the correctness of the variants produced by the variant calling pipelines. Recent advances in NGS platforms have improved error rates significantly. Nonetheless, the size of the human genome is about 3 billion base pairs [25], while working with the data of this magnitude, even a small error during variant calling can cause serious damage to the robustness of the variants. Such spurious technical errors require difficult and costly computational techniques to minimize the number of erroneous variants. This becomes crucial when incorrect variant resembles rare and deleterious mutations. Furthermore, the errors will have serious negative implications on downstream analysis and may result into producing wrong interpretation if not identified and resolved in time. [37]

Since DNA sequencing is a multi-step process, an error can happen in any of these steps. Frequently, the correctness of variants depends on how things are done in a wet-lab. For example, simple human errors such as mislabelling of a sample during sample preparation would completely jeopardize the outcome. In addition, low DNA amount or quality would also adversely affect the result of sequencing. Moreover, contamination of samples, machine failures, faulty workflow protocol, strand biases due to molecular amplification and working conditions amongst others are also the major cause of experimental errors during laboratory work. [37]

There are many strategies that laboratories can employ to mitigate the challenges mentioned above. Increasing the read depth, which is the frequency of a base sequenced at an individual position contributing to form the consensus sequence is the simplest method to improve accuracy. Nevertheless, problems caused by batch effects or the DNA sequence complexity itself, are immune to this approach. In addition to improving read depth, laboratories can redo the whole experimental workflow on the same sample, which can also improve the accuracy of the result (duplicate experiment). [37] Additionally, integration of data from different NGS platforms is known to improve sequence quality immensely since some errors are typical for one platform and absent from others [58]. In particular, combining short and long reads is now gaining a lot of interest [34]. Since certain kinds of errors are typical to some NGS platforms, data integration is a very good mitigation approach that the laboratories can use. However, it can be very costly because it requires more than one NGS platform. [37]

Even with all the mitigation techniques mentioned above, some errors such as faulty vari-

ants due to misalignment of short-reads in repetitive regions or the presence of very similar pseudo-genes in the genome need to be resolved by computational methods. Furthermore, the quality and completeness of the reference genome used has a bearing on the outcome. Human reference genome build GRCh37 is not complete and has gaps. Mutations falling into those gaps are always missed. However, with the arrival of the new build GRCh38, this problem has been addressed, albeit not completely. [37]

3 Variation in the genome

The human genomes of any two individuals are 99.9% similar to each other [40]. The remaining 0.1% difference is a key in diagnosing any genetic disorders. Before the era of today's advanced sequencing platforms, such genetic differences were first observed cytologically using microscopic techniques. They were very large changes of size ≥3 Mb and were mostly rare variants. They caused a visible change to the structure and quantity of chromosomes. These types of large changes are termed as microscopic structural variants. [40] With the advancement of sequencing technologies, smaller changes such as SNP/SNVs and InDels also became accessible. Nowadays, the convention is to separate small changes, *i.e.* SNP/SNVs and InDels, from larger ones, which are referred to as SV. Although the small events account for the most common occurrence of mutations in the human genome [77], SVs are responsible for most per base changes [71]. Recent findings suggest that each human genome comprises more than 20,000 SVs. Notably, the majority of them are not detectable by short-read sequencing platforms. [71] Among all the variants including SNPs, InDels and SVs, SVs larger than 2 kilobase (kb) are relatively the most unexplored. Their obscurity is chiefly due to the limitation in our capabilities to detect them. [49]

3.1 Structural variation

Originally, genomic alterations larger than 1 kb were referred to as SV [40]. However, the definition of SV has been widened. Now, any changes greater than 50 bp are called SV [71]. Thus, SV is a super class of different kinds of large mutational changes. CNV, translocation and inversion are a few examples of SVs. [77, 71, 40]

3.1.1 Copy Number Variation (CNV)

Copy number variation also known as copy number polymorphism is a type of SV where the copy number of a portion of a DNA is changed with respect to the reference genome [50]. Redon et al. [57] described CNVs as changes larger than 1 kb but now, any deletion and duplication ≥50 bp are defined as CNVs. CNVs are extremely common throughout the human genome. Current estimates suggest that CNVs cover about 12% of the human genome, making them as important as SNPs [54]. Moreover, Sebat et al. [32] found that 50 percent of all reported CNVs touch coding regions to some extent, further highlighting their importance for functional understanding of the genome. Evidence also suggests that CNVs play a crucial role in driving the evolutionary process. [54] Figure 3.1 depicts the CNV event on a chromosome.

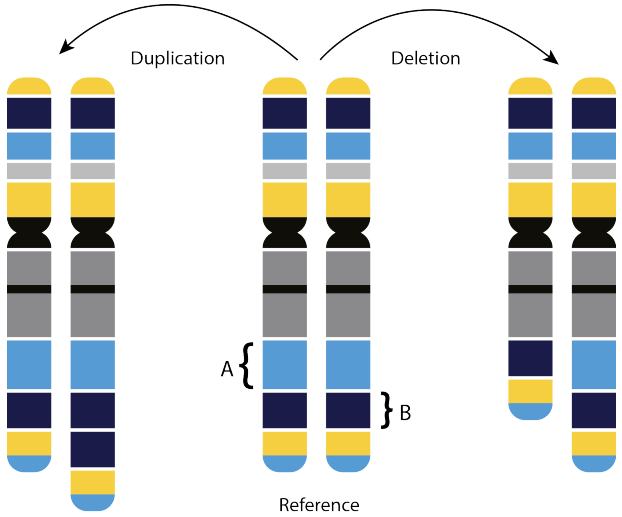

Figure 3.1: Illustration of the CNV event on a chromosome. The chunk 'A' from the reference genome has been removed causing a deletion event. Likewise, the chunk 'B' has been copied causing a duplication event. The figure is based on reference [17].

As shown in Figure 3.1, deletion and duplication are the two major types of CNVs. However, duplication can be of different types such as segmental duplications, tandem duplications and higher-grade amplifications. [77] Both somatic and meiotic mutations are known to cause copy number changes. CNVs can be of no effect or they might cause increased or decreased expression of a gene. Furthermore, it might also be that one copy retains the original functionality and the new copy develops entirely different functionality. At

the initial stages of copy number variation research, CNVs were positively received because researchers found that their presence in the gene responsible for olfaction resulted in the improvement of the sense of smell. [54] Nevertheless, CNVs are not perceived as advantageous any more. Although a majority of them are harmless and contribute to a natural process of an organism's evolution, some of them have been definitively linked to effects on biological processes, such as observable phenotype changes but also development, progression and aggressiveness of cancer. [54]

3.2 CNV detection methods

Today, variant calling is almost exclusively performed on NGS data in the clinical disease diagnostics context. Moreover, SNP and InDel calling has become a routine laboratory procedure. The large number of computational tools have been developed to mine genomic variations from short reads produced by NGS technologies. Many are even designed to detect copy number variations specifically. However, the detection of CNVs with reasonable sensitivity still remains a challenge. Also, breakpoint (start coordinate of the variant or range if position is unknown) detection of insertion variants is difficult because of the repetitive nature of the human genome [79]. Detection of CNV breakpoint is particularly difficult in targeted sequencing because sequences tend to discontinue [19].

Despite having a surplus of tools to detect CNVs, none is capable to find all of the changes present in the sample. All of these tools have their strengths and weaknesses as a result of trade-offs made during the development. There is a lot of literature available comparing the performance of these tools [33, 50, 9]. Users should choose wisely as per their needs. For instance, tools designed for WGS and WES data would not produce an expected result for data generated with panel-based targeted sequencing. Regardless of different names and tricks used by these tools, they mostly fall into one of the four general methods used

in CNV identification: (i) paired-end mapping (PEM), (ii) split read (SR), (iii) read depth (RD), and (iv) *de novo* assembly of the genome (AS). [47, 50] Figure 3.2 demonstrates the workflow of each method.

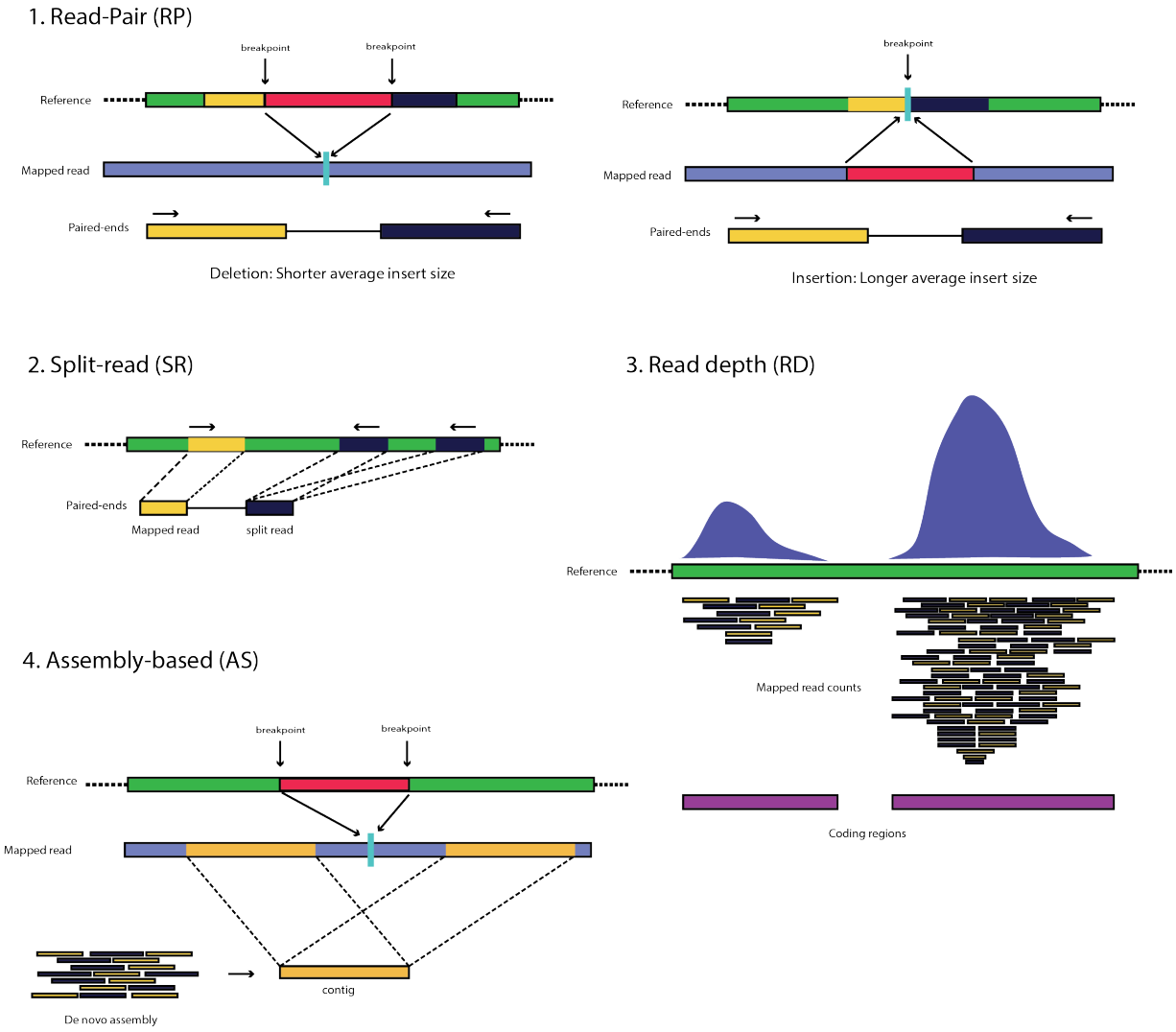

Figure 3.2: Four main strategies used for identifying CNVs from NGS data. However, these methods are not limited to CNV detection and are also used in general SV detection as well. 1. PEM-based approach looks into the average insert size to detect CNVs. 2. SR strategies report CNVs by making use of completely or partially unmapped reads. 3. RD-based approach looks into coverage of a genomic region to determine CNVs. 4. AS approach compare *de novo* contigs built from NGS data against reference genome to detect CNVs. The figure is based on references [50, 47].

Most read-depth based tools are designed for CNV detection. On the other hand, three other approaches are used to detect structural variants in general, which includes CNVs as well. Tools based on RD often take sequence alignment map (SAM) or binary alignment map (BAM) files as an input because they require read depth details of each sequenced

nucleotide. On the contrary, PEM and SR based tools accept FASTQ files because they only require positional information of a sequence. [77, 50]

Paired-end mapping

Paired-end mapping is a very popular technique employed by many tools used for CNV detection. It also happens to be the very first method used for calling SVs/CNVs from NGS data [29]. As the name suggests, the PEM algorithm requires paired reads to function. Paired reads are obtained by sequencing a DNA fragment from both ends. Consequently, paired-end sequencing produces twice the amount of data than single-read sequencing. Nonetheless, the data generated with this method is highly reliable and allows quality alignment. [22] In paired-end sequencing, the average insert size is determined during the sequencing process. Therefore, the average insert size is expected to follow a particular distribution. [29, 47, 50] PEM based methods use this phenomenon along with read pair orientation to identify SVs/CNVs. Discordant read pairs that are located too far signal deletion. On the contrary, insertion is reported if read pairs are located too close. [77] PEM is a very capable method that can effectively detect inversions, tandem duplications and mobile element insertions. Breakdancer, PEMer and Ulysses are some popular tools that use the PEM algorithm. [50, 47]

Despite being a very powerful technique at our disposal, PEM is not without limitations. For instance, PEM is not capable to identify insertions larger than the average insert size and copy number estimation reported by read pair is not very accurate [77]. Furthermore, the PEM method fails to detect CNVs falling in low-complexity regions of the genome [50]. Increasing the average insert size has some benefits such as better coverage and more variant detection power. However, a smaller insert size is very effective in identifying smaller variants with high confidence. PEM is inapplicable to single read data because at the time of development of this method, paired-end sequencing was a standard sequencing technique. Later, newer sequencing platforms started to produce single read data, which made this approach a bit obsolete. [77]

Split read

Similar to paired-end mapping, the split read method is also incapable to handle single read data and thus, requires read pairs. In SR methods, one read aligns accurately and the other one is either partially aligned or not aligned at all to the reference genome. [50] Mills

et al. first demonstrated this method to identify InDels in the human genome [60, 77]. The unmapped or partially mapped reads are further split into multiple pieces. The resulting pieces are then remapped independently against the reference genome. Consequently, SR based method is very effective in resolving breakpoint (down to single base pair level) of SVs/CNVs. Even though, the potential to resolve breakpoint accurately gives SR an edge over other methods, its application is limited to the unique region of the genome. In addition, local mapping used by the SR method massively relies on read length and is not suitable for large variants. Tools such as Pindel, Gustaf and PRISM apply SR algorithm. [50, 77, 35, 47]

Read depth

The emergence of read depth based methods is a direct consequence of the availability of a considerable amount of high-coverage NGS data. Methods based on read depth are relatively new. The use of depth of coverage (DOC) for CNV identification is based on the idea that more reads are detected if a duplication event has occurred and conversely, the number of reads would be significantly lower if a deletion event has happened. [67] Teo et al. [67] explained that read intensity is central to the RD-based CNV detection methods. The hypothesis mentioned above expects sequencing data to have uniform coverage. Moreover, aligned reads are expected to adhere to Poisson distribution. [77]

The RD-based approach follows a four-stage linear workflow. First of all, reads are aligned against the reference genome and the read depth is computed in each window locally. In the second stage, read depth is corrected for GC content and repeat region biases. In the third stage, normalized read depth is used to determine the precise copy number of gain or loss of DNA content. Finally, in the segmentation stage, neighbouring windows with similar copy numbers are merged to form a single large detection. The size of windows is either predefined or randomly assigned to make sure that reads mapping to a single window would not cross a specified threshold. Notably, circular binary segmentation (CBS) and Hidden Markov Model (HMM) are very popular methods to perform segmentation. [77]

One of the drawbacks of PEM and SR methods is their inability to handle single-read data. This is not a limitation to RD methods. In addition, unlike PEM and SR methods that rely on the position of reads, the RD-based technique can report absolute copy numbers. [50] As a result, the RD-based approach is a good fit for tools that are specifically designed for CNV detection [77, 68]. The data used in this study was generated almost exclusively with

RD-based methods. The RD method also outperforms PEM and SR methods in identifying large CNVs in the repetitive region [50]. Conversely, breakpoint resolution capability of RD-based method is extremely inferior compared to PEM and SR approaches [77]. Furthermore, GC content bias affects RD-based techniques the most [67]. A sequencing bias that causes variable coverage of reads due to richness and poorness of G and C bases is known as GC content bias [82]. Identification of duplication events with RD method is also less efficient because the difference in read depth is not always very significant [77]. CNV-seq, BIC-seq, CNVnator and ReadDepth are some popular tools that are based on read depth computation [47].

De novo assembly of a genome

In the methods explained above, the first step is to align reads against reference genome and a variant is called by investigating the difference between the two sequences. Unlike this approach, AS-based methods first align short reads against themselves and builds a contig. Such contigs are then used to reconstruct a whole DNA sequence known as the consensus sequence. Then, consensus sequence is compared with the known reference genome to identify SVs/CNVs. Although not required at all, reference genome can also be used during the assembly process to achieve speed and improvement of contigs. The consensus sequence assembled with the AS-based approach is called *de novo* assembly. [50] In theory, the AS-based method can detect a full range of SVs/CNVs along with *de novo* variants. Moreover, it can potentially resolve complex regions of the genome. On the contrary, accurate and long reads beyond the capabilities of today's sequencing platform are required to make this approach computationally viable. As a result, the method is not used for routine variant calling. The AS-based method performs poorly in repetitive regions because repeats collapse into one during the reconstruction process. [77] Magnolya is an assembly based tool that can be used for CNV identification [47].

The mixed method

Yoon et al. [64] predicted the use of the combined approach more than ten years ago. Therefore, the combinatorial approach is not exactly a novel idea. The PEM method is extremely efficient in detecting small deletions with outstanding breakpoint resolution. Nevertheless, it cannot report absolute copy numbers and performs poorly in identifying large CNVs. On the other hand, the RD-based methods fail to handle translocations and

inversions because of their copy neutral nature but detect large CNVs quite efficiently. Similarly, read depth approaches can estimate copy numbers quite accurately but are unable to identify breakpoints. In the matter of AS method, it excels in detecting *de novo* variants but requires intensive computational power and is unable to deal with variants located at repetitive regions. As for SR-based methods, smaller changes and precise detection of breakpoints is an advantage but the low-complexity region of the genome is a problem. [50]

As discussed above, one single strategy is not enough to detect the wide variety of CNVs. Hence, a number of tools have been developed by combining strategies of two or more than two of the four methods. As there are known limitations to each of the four methods, the idea behind combining them is to make them complement each other's weaknesses and maximize the sensitivity and specificity of the calls. [50] Tools based on a combined approach also demonstrated a low false positive rate [47]. The pair, PEM and RD is a very popular combination used by many tools [77]. SVDetect, CNVer and DELLY are some popular tools that are based on a combinatorial approach [47]. Although, this approach ouputs a different set of variants, it is not a different method by itself because it merely combines different methods.

4 Datasets

In this pilot study, we used datasets from several different sources. Since we are benchmarking, the data can be categorized into two main categories, the "gold standard", also known as the truth set or the benchmark set and the data to compare against it, also referred to as the callset or the comparison set. We received the callsets from The European Molecular Genetics Quality Network (EMQN) and Genomic Quality Assessment (GenQA), who collaborated with the participating laboratories. For sequencing, participating laboratories used the germline of widely available DNA of the son (HG002/NA24385) from the Personal Genome Project's Ashkenazi Jewish trio. Then, we downloaded the truth set for the same sample from the publicly available data source [34]. In addition to that, we used chain files (to convert coordinates from GRCh38 to GRCh37 genome assembly), exome coordinates and reference genome assembly that are publicly available in ensembl's ftp server [36].

4.1 The gold standard

The gold standard is a set of high quality variants that we used as a benchmark set in our analysis. Notably, it is by no means a perfect list of variants, but only the best available at the time when this work was done. In our analysis, we used the Tier 1 germline SVs published by the GIAB as the benchmark set that covers 2.51 Gbp of the GRCh37 reference genome assembly. The final Tier 1 set is a combination of results from 19 different variant calling methods containing 12745 variants. The variant detection methods were both *de novo* assembly and alignment based and all of them produced sequence-resolved variants.

Moreover, the variants catalogued in the truth set are deletions and insertions that are ≥50 bp. The variants that were detected by no less than two sequencing platforms or 5 callsets were classified as a true mutation event and recorded in the final truth set. In the context of preparing this truth set, Zook et al. define a "callset" as an output produced by a specific variant calling method, which is then later used to integrate into the truth set. [34]

The callsets were produced with several different sequencing platforms such as Illumina, 10x Genomics, PacBio and Complete Genomics. Illumina and Complete Genomics are short read sequencing technologies, whereas 10x Genomics and PacBio are linked read and long-read technologies respectively. Furthermore, Bionano's optical and Nabsys's electronic mapping was used to approximate the size of SVs. Currently, GIAB is working towards the use of new technologies such as PacBio circular consensus sequencing (CCS) and Oxford Nanopore sequencing to produce a more robust and comprehensive version of the benchmark set. Also, they are working on expanding the truth set to support the GRCh38 genome assembly. [34]

4.2 The callsets

Within the scope of this pilot study, we define "callset" as the list of variants submitted by the participating laboratories. In total, 25 laboratories took part in this study, thus we received 25 submissions for analysis. For data collection, participants were asked to fill up a survey describing their test process along with a VCF file containing variant calls and the browser extensible data (BED) file that contained the information about the target

region. Optionally, participants could also submit FASTQ or BAM file or both. Figure 4.1 provides information about the variants in the callsets.

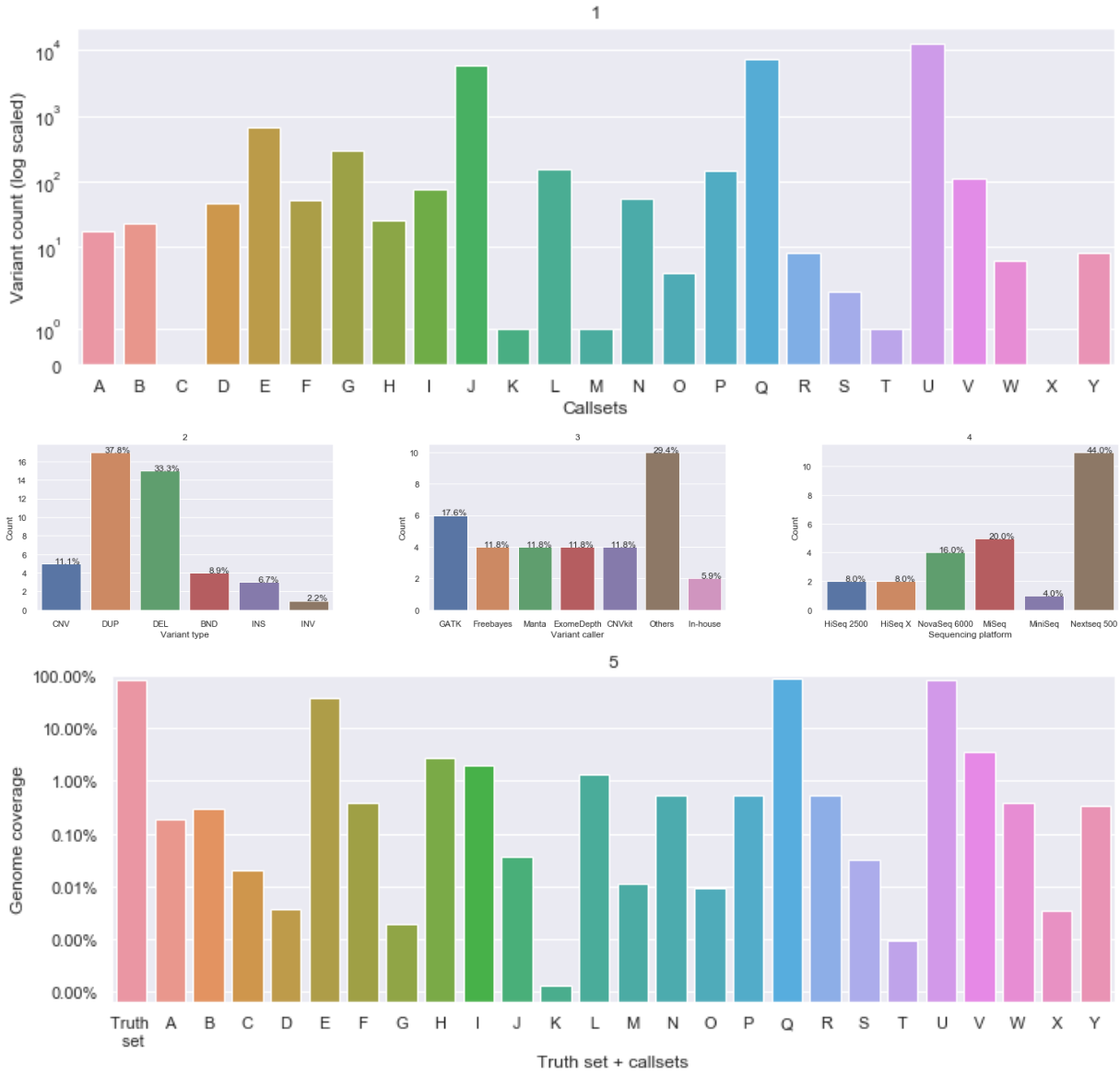

Figure 4.1: Some insights from the submitted callsets. 1. Varying number of variants reported by the participants. 2. Distribution of different type of structural variants such as copy number variation (CNV), duplication (DUP), deletion (DEL), breakend (BND), insertion (INS) and inversion reported by the participants. 3. Different type of variant callers used by the participants. 4. Different type of Illumina sequencing platforms used by the participants. 5. Genome regions covered by the truth set and the callsets.

As shown in figure 4.1, the number of variants varied greatly among submissions. Some submissions contained thousands of variants and some did not even report one variant. Similarly, participating laboratories used disparate methods to generate the callsets. The

majority (80%) of participants used the GRCh37 reference assembly. On the other hand, the remaining 20% used the GRCh38 reference assembly. Among the 25 submissions, 23 were targeted submissions (panel or exome), ranging from 4 kbp to 112 Mbp, whereas 2 submissions almost covered the entire genome. The size of the variants also varied among callsets as shown in Figure 4.2.

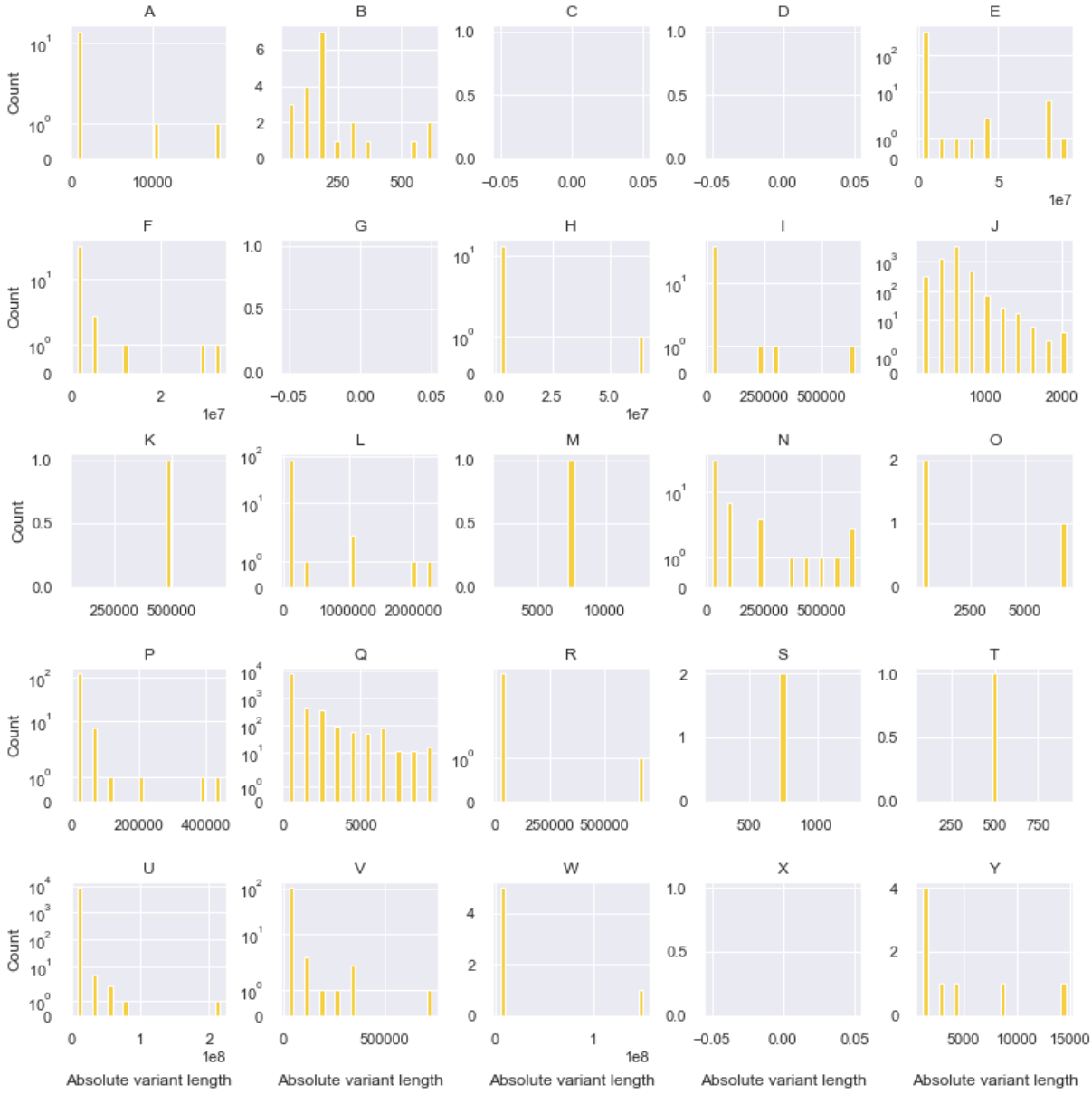

Figure 4.2: Distribution of the length of variants reported in the callsets.

Interestingly, all laboratories used Illumina technology for sequencing, albeit different platforms, hinting towards popularity of short read sequencing in clinical space. As part of the submission survey, participants were also asked to mention the variant caller used.

We found that despite using the same sequencing technology, laboratories used a variety of variant calling methods. Several submissions combined results of multiple variant callers into the same submitted callset. Submission contents varied regarding the types of variants submitted. Some submissions also contained SNPs and Indels (< 50 bp), which were not considered in our analysis.

4.3 Data pre-processing

Standardizing data obtained from the NGS technologies and bioinformatics infrastructure has been a challenge in computational genomics. Although, the field has achieved unprecedented success with the availability of abundance of data, efficiency has been reduced due to lack of standardization. In this study, we received data in the variant call format (VCF) [55] and BED files. The 1000 genome project [72] first introduced the VCF file to represent genetic variation in the human genome. However, its use is not just limited to the human genome. Today, diverse projects use VCF file format and it is a standard file format in a clinical space. The VCF is a generic and extremely flexible file format. Although it is a standard format and there are some conventions to be followed, the information about variants can be represented in many different ways in the VCF file. [55] For instance, one can add as many custom attributes as required in an INFO field. As a consequence, VCF files produced by two different variant callers are always different in some manner. As an example, a chromosome can be represented as either just a number or with a 'chr' prefix, but this subtle difference prevents bioinformatics tools from operating smoothly. As a result, data needs to be tailored to the need of the tool that is being used. In addition, freely available bioinformatics tools are poorly maintained and lack standardization themselves. Typically, lots of tweaking and cleaning is required before a NGS data can be given to the tool for smooth analysis. [70, 11] All the submissions in this study were different from each other, thus each submission was carefully curated and normalized before benchmarking. During our observation, we found that there is a considerable difference in how a variant type is reported by a particular variant caller. For example, some variant callers reported CNV calls as a generic 'CNV' type while others reported as 'DUP'/'DEL' calls. For submissions that used generic 'CNV' type, we normalized them to 'DEL' and 'DUP' types by computing the length of the variant.

In addition to VCFs, we received information about the targeted region in a BED file. The BED file format is considerably simpler than the VCF file format. For that reason, the

ones we received were less disparate and required a lot less cleaning. We only normalized the chromosome representation and removed extra columns other than start and end coordinates. Nevertheless, the two whole genome submissions had millions of rows with small intervals, which made the benchmarking extremely time inefficient. In order to solve this issue, bedtools [3] was used to merge these small intervals into a big one.

At the time of writing this thesis, the truth set was not available for GRCh38 genome assembly. The one we used in this study from GIAB was only available in GRCh37 coordinates. As a result, submissions from five participants that used GRCh38 genome assembly were remapped to GRCh37 coordinates using a tool called CrossMap [20]. Importantly, not all genomic regions in GRCh38 genome assembly are present in its predecessor. So, many reported calls were removed during this process. However, GIAB is in the process of developing the truth set for GRCh38 assembly as well [34]. Once it is public, that will address this issue.

5 Methods

We used two different methods to validate variants in the callset. We primarily used a tool called Truvari [16] developed by GIAB [34] with default parameters to produce evaluation metrics for sequence based analysis. On the other hand, we used our in-house strategy for exome-based benchmarking. Furthermore, we used Integrative Genomics Viewer (IGV) [28] to manually confirm results in many occasions.

5.1 Evaluation metrics

We used precision, recall and F1-score as metrics to assess the VCFs submitted by the laboratories. Truvari outputs these numbers by comparing variants (calls) in a submitted VCF (callsets) with variants in the truth set (base calls). Calls can be classified into one of the three categories: true positive (TP), false positive (FP) or false negative (FN). Since we are primarily interested in finding the positives, we discard true negatives (TN) in this study. Ideally, laboratories would want all reported variants to be present in the callset. A call variant is marked as TP when the same variant is present in the truth set as well. These are the calls that are correctly reported by the laboratories. Similarly, a variant is marked as FP when it is present in the callset but not in the truth set. These are the calls

that had been incorrectly reported by the laboratories. Finally, a variant is classified as a false negative when it is absent in the callset but is present in the truth set. These are the calls that had been missed by the laboratories. [21] Figure 5.1 below depicts the above-mentioned idea.

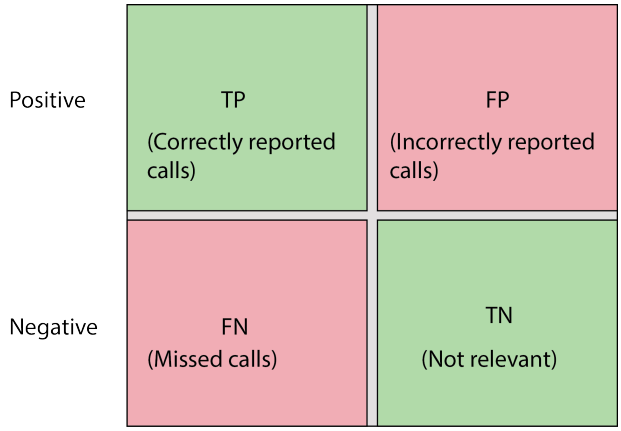

Figure 5.1: The confusion matrix showing the binary classification of variants in the callset.

The precision, recall and F1-score are derived from TP, FP and FN scores. The score is defined as a sum of the total number of variants falling into each category.

- *Precision* (also known as a positive predictive value) is a ratio of correctly reported calls over the total number of reported comparison calls. It is computed as

$$Precision = \frac{TP}{TP + FP}$$

[27]

- *Recall* (also called sensitivity or the true positive rate) is a ratio of correctly reported calls over the total number of base calls. It is computed as

$$Recall = \frac{TP}{TP + FN}$$

[27]

- *F1-score:* Neither precision nor recall alone is enough to evaluate the performance. F1-score is an aggregate metric that combines both recall and precision. Simply, the harmonic mean of the precision and recall is defined as the F1-score. It is computed with the formula

$$F1 = \frac{2 \quad precision \quad recall}{precision + recall}$$

[27]

5.2 Assessment workflow

The workflow that we used in our assessment of callsets is quite simple. First of all, we preprocess the callset and the truth set as described in Section 4.3. In addition, we only retain variants from the truth set that has a "PASS" filter. For stratified analysis, we filter variants from both the callset and the truth set depending on the variant type to be analyzed. Since VCF files can be extremely large, we then compress and index both the callset and the truth set for efficient use of space and fast access to the data. Subsequently, we benchmark the call set against the truth set. Once benchmarking is complete, evaluation metrics are calculated and assessment is performed based on those metrics. Figure 5.2 illustrates our assessment workflow, which is a multi-step process.

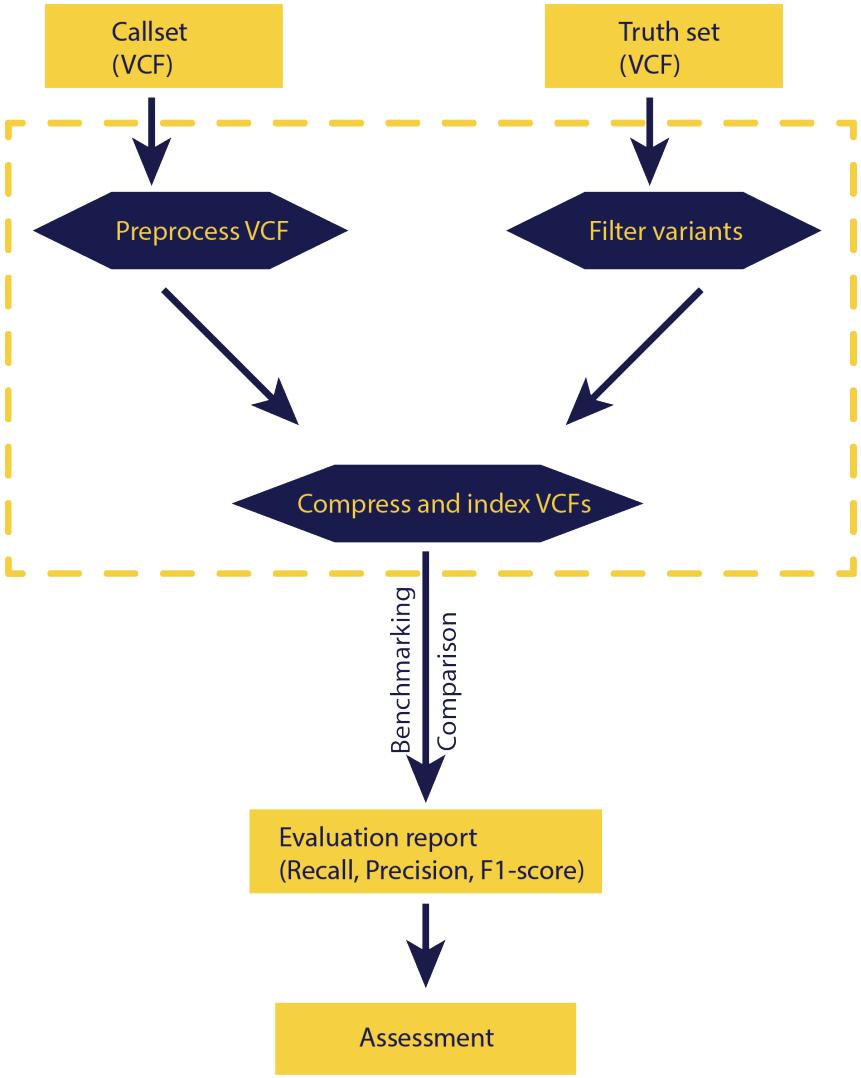

Figure 5.2: Assessment workflow used in the study.

5.3 Evaluated regions

During our observation, we found that many of the targeted submissions called variants outside their sequencing target indicated by the submitted BED file. This is likely due to a capture-based targeting strategy, where regions adjacent to the targeted region (within the captured fragment) are also sequenced. In order to accommodate this, the submissions were evaluated in the area, which was common with the Tier 1 high-confidence region of the benchmark set and the target of the submission after expanding each region in the target by ±250 bases in each direction. The extension of the target region was performed while doing the separate analysis of the deletion calls but not during the first round of analysis i.e. analysis including duplication calls.

5.4 Variant comparison strategy

5.4.1 Sequence based comparison

The sequence-based comparison is a straightforward way of validating variants in the callset. In this method, we simply compare variant sequences present in the callset and the truth set against each other.

Sequence comparison method

Truvari uses an elaborate algorithm to compare and classify variants. Once the comparison and classification is complete, it outputs evaluation metrics necessary to assess the quality of the callset. Algorithm 1 shows the pseudo-code of the algorithm used by Truvari [16]. Parameters used in the Algorithm 1 are as follow:

- *Refdist* is a maximum distance a comparison call can be within from base call's start and end co-ordinates to be marked as a neighbor. 500 is a default value of refdist. [16]
- *Pctsim* is a Levenshtein distance ratio between base and comparison call. Default value of pctsim is 0.7. However, it can be set to 0 if calls are not sequence resolved (variants with actual bases not just starting base and start/end co-ordinate) in either comparison or base set. [16]

- *Pctsize* is the size ratio of call and base and is computed as:

$$pctsize = \frac{min(comp\_length, base\_length)}{max(comp\_length, base\_length)}$$

  Hence, if the value of pctsize is 1 then that would mean the length of both comparison call and base call is the same. We do not want to compare variants if their size difference is extremely large because other criteria for matching would easily be satisfied in cases when either comparison or base variant is extremely large. The default value of pctsize is 0.7, this means the size of both comparison or base call needs to be at least 70 percent of the other. [16]

- *Pctovl* is a reciprocal overlap between comparison and base call. It is computed as:

$$pctovl = \frac{sizeofoverlappingbases}{max(comp\_length, base\_length)}$$

  For a variant to be classified as TP, We would like the major portion of the comparison call to overlap with the base call. The default value for pctovl is 0.7. As a result, if the overlap between comparison and base call is less than 70% then that variant is not classified as TP. [16]

- *TruScore* (also called Truvari score) is an aggregate value derived from pctsim, pctsize and pctovl. Neighbor getting the highest truscore is classified as TP. Truscore is computed as:

$$truscore = \frac{2 \quad pctsim + pctsize + pctovl}{3}$$

  Pctsim is given two times the weight than the others. [16]

Algorithm 1: Classify variants as TP, FN and FP [16]

Input: CompCalls, TruthSet, CompBed, TruthBed

Output: list of TP, FP and FN variants

```
init
    Build interval tree of CompCalls;
for each variant t in TruthSet do
    fetch CompCall variants overlapping within refdist from interval tree;
    for each fetched CompCall variant c do
        if (c.variant_type == t.variant_type
        AND Levenshtein distance ratio of c and t >= pctsim
        AND sizeRatio of c and t >= pctsize
        AND ratio of reciprocal overlap of c and t >= pctovl) then
            Add c to the list of neighbors;
        end
    end
    sort list of neighbors by TruScore;
    Mark highest scoring c and t as TP;
    Use c only once unless –multimatch parameter is given;
    if neighbors==0 then
        Mark t as FN;
    end
end
for each variant c in CompCall do
    if c is not marked as TP then
        Mark as FP;
    end
end
```

As shown in Algorithm 1, sequences are matched based on the Levenshtein distance [76] and reciprocal overlap. The absolute sequence matching would not produce lots of matches because having nucleotide level accuracy for large changes like CNVs is not practical. Moreover, the clinical significance of the mutation probably would not change much even if variants are matched with a stringent scheme.

Limiting truth set within the target region of the callset

During our observation, we noticed that the average length of variants in the truth set is much longer than that of in the callsets. This is reasonable because the truth set covered the whole genome and used long-read sequencing platforms along with short-read technologies. Conversely, callsets were prepared with short-read Illumina platforms with an average read sizes ranging from 150 bp to 300 bp. It leads to a problem that smaller variants in the callset would never match with much longer variants in the truth set because their size ratio would never be within the threshold. Figure 5.4 shows the real case from one of the submissions.

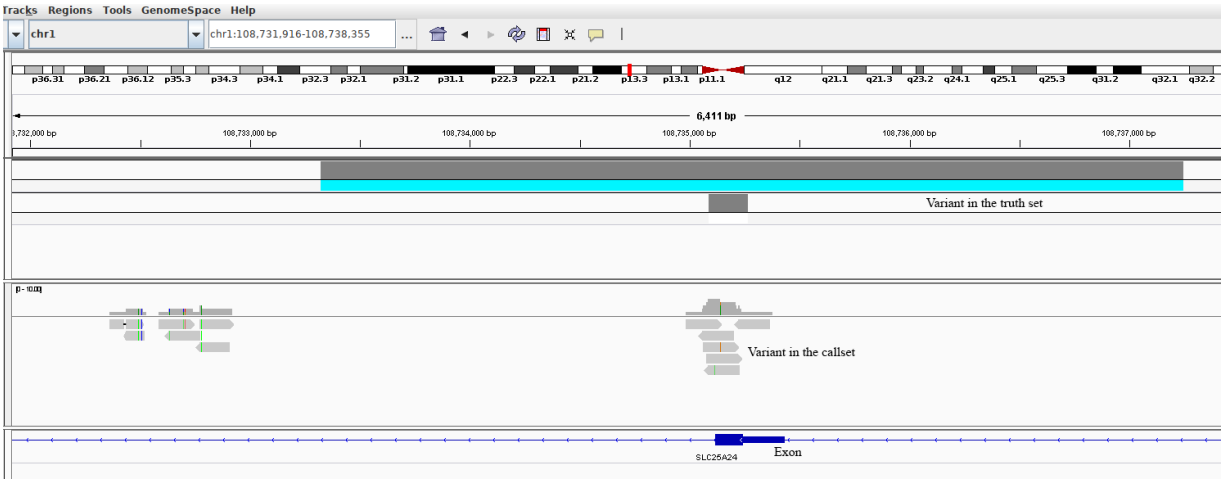

Figure 5.3: A variant in the callset that is completely inside the larger variant in the truth set.

The variant shown in Figure 5.4 was marked as false positive due to the enormous disparity in the size ratio. Nevertheless, these two variants are probably the same, it is just that targeted sequencing did not target the whole region. To address this issue, we limited variants in the truth set only to the high confidence region of both the truth set and the callset. The steps taken to do this is as follow:

1. Get overlapping regions between the truth set and the call set.
2. Limit (cut) variants in the truth set to the overlapping intervals obtained in the previous step.

We used bedtools [3] to get the overlapping intervals.

Stratified analysis

Delage et al. [79] demonstrated that only 17-28% of insertion calls (≥50 bp) could be detected by the short-read platforms. While building the truth set used in this study, Zook et al. [34] experienced the same. Clearly, short read based technologies struggle to detect insertions more than deletions. With all these insights from prior work and our observation of the comparison of the full set of data, we hypothesized that perhaps automated concordance analysis for deletion calls is more feasible than that of duplication calls. Hence, in our second round of analysis, we only benchmarked deletion calls.

Limiting analysis to the WGS submissions

As previously mentioned, we received two WGS submissions covering more than 80% of the genome in this pilot. Among all the submissions, these two were the most similar to the truth set and yielded the most promising result. Therefore, we further analyzed WGS submissions with two extra tools, namely, Readdi [59] and SVanalyzer [52] to validate the results obtained from Truvari. Notably, the tool Readdi can only benchmark DEL calls.

5.4.2 Exome based comparison

Exons are protein-coding segments of any gene [5, 62]. They cover about 2% of the human genome. Nevertheless, exons comprise 85% of the known pathogenic variants. [12] Due to this correlation, many researchers only focus on exonic variants [15]. Furthermore, exome sequencing is a common first step while dealing with many genetic disorders [26]. On the other hand, many genetic diseases have well-known phenotype, albeit not all. Usually, these distinct signals direct clinicians to test for mutations in a handful of genes. For instance, a mutation in the Cystic fibrosis transmembrane conductance regulator (CFTR) gene causes cystic fibrosis and patient develops a peculiar phenotype. In this case, clinicians just need to know if mutation has occurred in the CFTR gene to verify the phenotypic signal [56], resolution down to nucleotide level change is not required.

Since exonic variants are clinically more relevant and nucleotide level resolution of variants are not always required to establish medical significance, we used these two concepts to develop our exome based strategy. Compared to sequence based method explained in Section 5.4.1, our alternative strategy is very lenient. In our new approach, we only check if the exon has been affected by a mutation or not. The underlying idea behind this

approach is that if the same exon in both the callset and the truth set has been affected by the same type of mutation event, for example, a deletion, then it should have the same clinical consequence as far as the patient is concerned. Algorithm 2 illustrates our exome based strategy.

Algorithm 2: Classify variants based on presence of mutation event affecting an exon

Input: CompCalls, TruthSet, CompBed, TruthBed, exonBed

Output: list of TP, FP and FN variants

```
init
    Only retain variants that overlap with exon in both CompCalls and TruthSet;
    t-dict = {};
    c-dict = {};
for each variant t in TruthSet do
    t-dict[exon affected by t] = t.variant_type;
end
for each variant c in CompCalls do
    c-dict[exon affected by c] = c.variant_type;
end
for each item e in min_size(t-dict, c-dict) do
    for each item i in max_size(t-dict, c-dict) do
        if e.exon_number == i.exon_number then
            Mark as TP;
        end
    end
end
for each item i in t-dict do
    if i is not marked as TP then
        Mark as FN;
    end
end
for each item i in c-dict do
    if i is not marked as TP then
        Mark as FP;
    end
end
```

As shown above, algorithm 2 takes five inputs.

- *CompCalls* is a list of variants in the callset. It is a VCF file.
- *TruthSet* is a list of variants in the truth set. It is also a VCF file.
- *CompBed* is a target regions of the callset. It is a BED file.
- *TruthBed* is high confidence region of the truth set. It is also a BED file.
- *exonBed* is also a BED file containing start and end coordinates of all the known exons in the human genome. It was downloaded from ensembl database [36].

In our exome based strategy, we start by removing variants that do not affect any exons from both the truth set and the callset. In other words, if a variant does not touch an exon, we remove it from our analysis. We use normalized VCF in this analysis. Therefore, VCF only contains variants of type DEL and DUP. Once the intronic variants are removed, we make a list of DEL and DUP calls in each exon for both the truth set and the callset. We only keep one mutation per exon, whenever there are two deletions reported in an exon, we only list one in our data structure. The data structure we use is a key-value pair dictionary where exon id is a key and mutation type is a value. Figure 5.4 depicts the data structure used in this process.

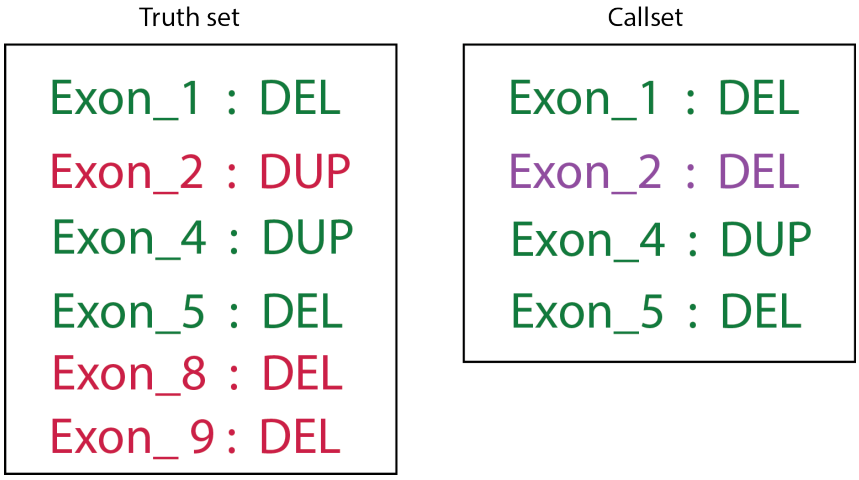

Figure 5.4: Illustration of key-value pair of exon id and variant type stored in a data structure.

Following the preparation of the two separate lists for the truth set and the callset, we make a comparison of these two lists. The match criteria is very simple, identical rows are marked as true positives. Provided there is a deletion in Exon_1 in both the truth set and the callset as shown in Figure 5.4, we mark it as a true positive. On the contrary, rows that fail to match are marked as false negatives and false positives. In Figure 5.4, records

in green, red and purple are true positives, false negatives and false positives respectively. The time complexity of Algorithm 2 is proportional to the number of exons affected by the variants in both the truth set and the callset. Assuming m and n are number of exons affected by mutation in the truth set and the callset respectively, the time complexity of Algorithm 2 is $O(m \times n)$. The quadratic running time of Algorithm 2 can be improved to $O(n \log m)$ by using the search routine of the dictionary, but here the quadratic solution is sufficiently efficient.

Our exome based strategy is experimental and suffers from some known limitations. For instance, if an exon has been affected by both deletion and duplication, then this approach would only retain one of the two, whichever comes last. However, this is a perfectly valid biological scenario. Moreover, we are discarding intronic variants in this method, hence the results obtained will never be complete. This approach most likely would not be comprehensive enough to validate the correctness of the callset. However, it can be used as a rapid screening method. At the least, it can help narrow down the scope of genetic testing and help guide the therapeutic process in the right direction. In addition, this approach can be used as an orthogonal method to validate exonic CNVs reported by other tools. For future work on validating CNVs affecting exons, the high quality validation set publish by Mahamdallie et al. [66] could be a great resource.

6 Results

We applied two strategies to perform CNV concordance analysis on a submitted sequence data. We used Truvari as a primary tool for sequence based benchmarking and our in-house strategy for exome based comparison. Firstly, the number of TPs, FPs and FNs variants were counted from which the precision, recall, and F1-score were computed. For two better performing WGS submissions, we conducted further analysis with Readdi [59] and SVanalyzer [52]. Moreover, we conducted pairwise concordance analysis of the two WGS submissions with Truvari, Readdi and SVanalyzer to have more confidence on the data because in principle they should be identical. For the pairwise benchmarking, we did not have to remap coordinates because both WGS callsets were based on the GRCh38 genome assembly. This preserved the variants that otherwise would have been removed during the remapping process.

6.1 Method 1: Sequence based comparison

6.1.1 Analysis of deletion and duplication calls with Truvari

Among the 25 submissions, 23 were used for the benchmarking of both DEL and DUP calls because 2 submissions did not report any CNVs. Based on the benchmarking with Truvari (Table 6.1), out of the 23 submissions, only 4 contained at least one true positive variant. Unsurprisingly, these 4 callsets were WGS and WES submissions. The precision obtained from WES submissions E and I were 18% and 33% with recall of 1% and 6% and F1-score 0.02 and 0.068 respectively. On the contrary, both WGS submissions Q and U performed well with high precision of 89%, recall of 29%, and F1-score 0.43 and 0.44 respectively. All other submissions were found to have 0 precision and recall.

6.1.2 Stratified analysis of deletion calls with Truvari

For the benchmarking of only deletion calls with Truvari, we extended the target region by 250 bases on each direction. This extension and stratification technique improved the scores for the 2 WGS and 2 WES callsets from before. However, precision and recall for

all the other submissions remained 0. Based on these benchmarking results (Table 6.2), the precision obtained from WES submissions E and I were 24% and 80% with the recall of 2.5% and 5.6% and F1-score 0.046 and 0.105 respectively. The score for WGS callsets Q and U also improved to a precision of 92% and 89% and to a recall of 68% and 69% and F1-score 0.78 and 0.77 respectively.

Table 6.1: Method 1: Benchmark results of both deletion and duplication calls with Truvari.

Submissions	Target region	Precision	Recall	F1-score	TPs	FPs	FNs
A	Whole Exome	0	0	NaN	0	0	6
B(GRCh38)	Gene panel	0	0	NaN	0	0	16
C(GRCh38)	Gene panel	0	0	NaN	0	0	4
D	Gene panel	No CNVs, therefore not assessed.					
E	Whole Exome	0.188	0.01	0.02	45	194	4135
F	Gene panel	0	0	NaN	0	0	18
G	Gene panel	No CNVs, therefore not assessed.					
H(GRCh38)	Whole Exome	0	0	NaN	0	0	163
I	Whole Exome	0.33	0.068	0.113	5	10	68
J	Gene panel	0	0	NaN	0	0	0
K	Clinical Exome	0	0	NaN	0	0	0
L	Gene panel	0	0	NaN	0	9	70
M	Gene panel	0	0	NaN	0	0	0
N	Gene panel	0	0	NaN	0	0	23
O	Gene panel	0	0	NaN	0	0	0
P	Gene panel	0	0	NaN	0	0	23
Q(GRCh38)	Whole Genome	0.89	0.29	0.43	2832	332	6877
R	Clinical Exome	0	0	NaN	0	0	24
S	Gene panel	0	0	NaN	0	0	1
T	Gene panel	0	0	NaN	0	0	1
U(GRCh38)	Whole Genome	0.89	0.29	0.44	2857	352	6881
V	Whole Exome	0	0	NaN	0	4	66
W	Gene panel	0	0	NaN	0	0	18
X	Gene panel	0	0	NaN	0	0	0
Y	Clinical Exome	0	0	NaN	0	0	12

Table 6.2: Stratified analysis: Benchmark results of deletion calls with Truvari after extension of the target region by 250 bases on each side

Submissions	Target region	Precision	Recall	F1-score	TPs	FPs	FNs
A	Whole Exome	0	0	NaN	0	3	7
B(GRCh38)	Gene panel	0	0	NaN	0	4	14
C(GRCh38)	Gene panel	NaN	NaN	NaN	0	0	0
D	Gene panel	No CNVs, therefore not assessed.					
E	Whole Exome	0.244	0.025	0.046	44	136	1688
F	Gene panel	0	0	NaN	0	3	17
G	Gene panel	No CNVs, therefore not assessed.					
H(GRCh38)	Whole Exome	NaN	0	NaN	0	0	124
I	Whole Exome	0.8	0.056	0.105	4	1	67
J	Gene panel	NaN	NaN	NaN	0	0	0
K	Clinical Exome	NaN	NaN	NaN	0	0	0
L	Gene panel	0	0	NaN	0	36	65
M	Gene panel	NaN	NaN	NaN	0	0	0
N	Gene panel	NaN	0	NaN	0	0	23
O	Gene panel	NaN	NaN	NaN	0	0	0
P	Gene panel	0	0	NaN	0	9	22
Q(GRCh38)	Whole Genome	0.921	0.688	0.787	2832	242	1284
R	Clinical Exome	NaN	0	NaN	0	0	21
S	Gene panel	0	0	NaN	0	1	1
T	Gene panel	NaN	NaN	NaN	0	0	0
U(GRCh38)	Whole Genome	0.89	0.692	0.778	2857	352	1270
V	Whole Exome	0	0	NaN	0	1	137
W	Gene panel	NaN	0	NaN	0	0	17
X	Gene panel	NaN	NaN	NaN	0	0	0
Y	Clinical Exome	NaN	0	NaN	0	0	3

6.1.3 Further analysis of WGS submissions

The WGS submissions Q and U were further analysed for deletion calls with Readdi and SVanalyzer (Table 6.3). Compared to Truvari, both Readdi and SVanalyzer achieved lower precision, recall and F1-score for submission Q. The precision, recall and F1-score with Readdi were 56%, 66% and 0.60 and with SVanalyzer were 57%, 67% and 0.61 respectively. For submission U, Readdi performed slightly better than Truvari with 94% precision, 67% recall and F1-score of 0.78. On the other hand, performance of SVanalyzer was almost similar to Truvari with 89% precision, 67% recall and F1-score of 0.77.

In addition, the pairwise comparison of WGS submissions Q and U were performed with Truvari, Readdi and SVanalyzer as shown in Table 6.4. Truvari performed best with 89% precision, 88% recall and F1-score 0.89. However, SVanalyzer and Readdi provided low precision of 53% and 52% with recall of 85% and 87% respectively and F1-score of 0.65. Truvari's slightly better performance might be because of its ability to consider the targeted region provided in BED files. Also, since Truvari was our first choice, variants were preprocessed before analyzing, which may have helped for slightly better results, too.

Table 6.3: Benchmark results of deletion calls for two WGS submissions with Readdi and SVanalyzer compared to Truvari

Submissions	Tool	Precision	Recall	F1-score	TPs	FPs	FNs
Q	Truvari	0.92	0.68	0.78	2832	242	1284
	Readdi	0.56	0.66	0.60	2856	2243	1459
	SVanalyzer	0.57	0.67	0.61	2920	2180	1402
U	Truvari	0.89	0.69	0.77	2857	352	1270
	Readdi	0.94	0.67	0.78	2876	172	1439
	SVanalyzer	0.89	0.67	0.77	2935	366	1387

Table 6.4: Benchmark results of pairwise comparison of WGS submissions with Truvari, Readdi and SVanalyzer

Submissions	Tool	Precision	Recall	F1-score	TPs	FPs	FNs
Q vs U	Truvari	0.89	0.88	0.89	2831	333	371
Q vs U	Readdi (only DELs)	0.52	0.87	0.65	2669	2430	379
Q vs U	SVanalyzer	0.53	0.85	0.65	2802	2470	497

6.2 Method 2: Exome based comparison

With our alternative exome-based method, we hoped to improve the precision and recall scores. However, only 10 callsets contained at least one key-value pair of exon id and mutation type that matched with the key-value pair in the truth set. Again, two WGS callets along with one WES callset scored the same highest F1-score of 13%. The F1-score of all other submissions were less than 7%. Table 6.5 shows the benchmarking results for all the callsets with our exome-based approach.

Table 6.5: Method 2: Benchmark results with in-house exome-based approach.

Submissions	Target region	Precision	Recall	F1-score	TPs	FPs	FNs
A	Whole Exome	0	0	NaN	0	57	1409
B(GRCh38)	Gene panel	0	0	NaN	0	24	1409
C(GRCh38)	Gene panel	No variants reported, therefore not assessed.					
D	Gene panel	No CNVs, therefore not assessed.					
E	Whole Exome	0.001	0.064	0.002	90	87112	1319
F	Gene panel	0.0005	0.0014	0.00072	2	4077	1407
G	Gene panel	No CNVs, therefore not assessed.					
H(GRCh38)	Whole Exome	0.085	0.0035	0.0068	5	54	1404
I	Whole Exome	0.17	0.037	0.06	53	258	1356
J	Gene panel	NaN	NaN	NaN	0	0	0
K	Clinical Exome	NaN	NaN	NaN	0	0	0
L	Gene panel	0.032	0.0234	0.028	33	997	1376
M	Gene panel	NaN	NaN	NaN	0	0	0
N	Gene panel	0.0022	0.0014	0.0017	2	906	1407
O	Gene panel	NaN	NaN	NaN	0	0	0
P	Gene panel	NaN	NaN	NaN	0	0	0
Q(GRCh38)	Whole Genome	0.395	0.0794	0.132	112	171	1297
R	Clinical Exome	0.031	0.0014	0.0027	2	61	1407
S	Gene panel	NaN	NaN	NaN	0	0	0
T	Gene panel	0	0	NaN	0	2	1409
U(GRCh38)	Whole Genome	0.227	0.095	0.134	134	457	1275
V	Whole Exome	0.226	0.095	0.134	39	1133	1370
W	Gene panel	0	0	NaN	0	6329	1409
X	Gene panel	No variants reported, therefore not assessed.					
Y	Clinical Exome	NaN	NaN	NaN	0	0	0

As shown in Table 6.5, our alternative strategy produced lots of FPs and FNs. In some cases, an overwhelmingly large number of FPs. For example, callset E reported 87112 FPs, nevertheless, further analysis revealed that callset E contained several extremely long variants that were millions of bases long as shown below.

CHR	POS	ID	REF	ALT	QUAL	FILTER	INFO
12	25956511	.	N	<DEL>	1000	PASS	END=104373629; SVLEN=78417118

Consequently, one variant affected thousands of exons, therefore the high number of FPs. During our observation, we also found that some (J, K, M, O, P, S, Y) panel-based targeted sequencing did not affect any exons at all. As a result, these callsets did not produce any results because they were empty.

7 Discussion

As new studies keep linking CNVs to rare diseases [53], CNV calling for disease diagnostic laboratories is becoming more relevant than ever. Due to this need and the availability of the number of tools capable to detect and interpret CNV calls, many medical laboratories are starting to incorporate genetic testing for CNVs. As more and more laboratories adopt various CNV calling pipelines, it is imperative that their approaches are standardized to ensure quality of testing via EQA schemes.

The main purpose of this study was to examine whether automated concordance analysis of CNV calls for the EQA is feasible or not. This study showed that the result of concordance analysis is dependant on how the sequence data is generated. The results implied that compared to the callsets generated with WGS, concordance of the callsets generated with the panel-based targeted sequencing is poorer. Additionally, the results also showed that the concordance of deletion calls is better than that of duplication calls.

The low precision and recall of targeted panel-based sequence data are explained by the fact that short-read sequencing produces discontinuous sequences. As a result, the breakpoint resolution capability of short-read sequencing is limited, which adversely affects the CNV detection process. [77] Furthermore, all the participants in this study used RD-based method for CNV calling, which performs better for the WGS data because coverage of WGS sequencing is more uniform than that of targeted sequencing [77]. Moreover, short-read technology lacks the ability to access GC high and repetitive region of the genome [74] that contributed to the poor result of panel-based sequence data. Similarly, the poor concordance of duplication calls was expected because duplications are generally known to be more difficult to detect than deletions [34, 77].

The instructions provided to the participants of this study were not stringent. Consequently, even though participating laboratories were asked to submit CNV calls, all kinds of SVs were received, which added extra time for data pre-processing. Likewise, the participants used a variety of CNV callers. The overall score of precision and recall suffered from it because multiple studies have found that the concordance of variants called by different tools is poor [77]. Moreover, the majority of the participants targeted different regions of the genome for sequencing, which added to the complexity in a concordance analysis. For the concordance study like ours, the dataset generated by the similar procedures would

produce better results. Also, this would help us validate the correctness of the data by conducting a pairwise comparison of the callsets.

One of the major limitations of this study was the poor quality of the submitted data. To our knowledge, this is the first ever pilot study to validate CNV data quality. The nature of this study was experimental and a first step towards automating CNV concordance analysis. Therefore, the participants might not have applied the same rigour as they would for the regular clinical samples to produce the dataset submitted in this study. In addition, since CNV calling is not a regular clinical procedure for the participating laboratories, it might be that users simply did not have comprehensive knowledge about their CNV calling pipeline. Oftentimes, bioinformatics tools are command line based, which diminishes their usability for normal users. It is very common for these tools to take multiple parameters to improve their output. In the case of CNV calling, parameter tweaking is very important to obtain an optimum output [33, 77]. Executing the CNV calling pipeline with the default parameters would reduce the data quality significantly. Another limitation to this study was the truth set itself. The truth set used in this study is based on the whole genome sequencing. Additionally, multiple sequencing platforms including long-read technologies have been used to construct it. Therefore, there is a significant difference between the truth set and the callsets that were generated by the short-read targeted sequencing.

8 Conclusions

The concordance analysis of sequence data is crucial for precision medicine and better healthcare. In this work, we looked into several strategies to investigate the feasibility of an automated CNV concordance analysis for the routine EQA. The result of CNV concordance is heavily dependent on the data produced by the variant calling pipeline that the laboratories have adopted. Although, we expected that CNV concordance strategy used in this study was relevant, the overall performance of the submitted dataset was low. To conclude, CNV detection capabilities of today's state of the art NGS technologies from targeted panel-based sequence data are limited. However, for whole genome sequencing, automated concordance analysis might be feasible for deletion calls.

8.1 Future directions

As discussed in Chapter 7, we believe the quality of the dataset is the major reason for the low concordance obtained in this study. We are continuing this work, in fact, we are already planning to launch the second round of submissions with more precise and stringent instructions to overcome the poor data quality issue faced in this study. In future studies like this, the data generation process needs to be standardized.

Further, it would be interesting to benchmark callsets against different truth sets since there are at least two available for the sample (HG002/NA24385) used in this study [2, 66] to validate the truth set from GIAB. Moreover, pairwise comparison of these three truth sets to observe the differences between them is also intriguing. For the majority of our analysis, we relied on just one tool. A rigorous analysis of the dataset with multiple benchmarking tools could be done to compare the performance of the tools.

Bibliography

[1] A M Maxam, W Gilbert. "A new method for sequencing DNA". In: *Proc Natl Acad Sci USA* 74 (1977), pp. 560–564. DOI: https://doi.org/10.1073/pnas.74.2.560.

[2] Aaron M. Wenger, Paul Peluso, William J. Rowell. "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome". In: *Nature Biotechnology* 37 (2019), pp. 1155–1162. DOI: https://doi.org/10.1038/s41587-019-0217-9.

[3] Aaron R. Quinlan, Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26 (2010), pp. 841–842. DOI: https://doi.org/10.1093/bioinformatics/btq033.

[4] Aquillah M. Kanzi, James Emmanuel San, Benjamin Chimukangara. "Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance". In: *Frontiers in Genetics* 11 (2020). DOI: https://doi.org/10.3389/fgene.2020.544162.

[5] N. M. Bahareh Rabbani Mustafa Tekin. "The promise of whole-exome sequencing in medical genetics". In: *Journal of Human Genetics* 59 (2014), pp. 5–15. DOI: https://doi.org/10.1038/jhg.2013.114.

[6] Bertil Schmidt, Andreas Hildebrandt. "Next-generation sequencing: big data meets high performance computing". In: (2017). Available at: https://doi.org/10.1016/j.drudis.2017.01.014 [Accessed 31 January 2021].

[7] Bo Segerman. "The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases". In: *Frontiers in Cellular and Infection Microbiology* 10 (2020). DOI: https://doi.org/10.3389/fcimb.2020.527102.

[8] Boluwatife A. Adewale. "Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years?" In: *African Journal of Laboratory Medicine* 9 (2020). DOI: https://doi.org/10.4102/ajlm.v9i1.1340.

[9] E. E. E. Can Alkan Bradley P. Coe. "Genome structural variation discovery and genotyping". In: *Nature Reviews* 12 (2011), pp. 363–376. DOI: https://doi.org/10.1038/nrg2958.

[10] Chiara Di Resta, Maurizio Ferrari. “Next generation sequencing: from research area to clinical practice”. In: *The Journal of the International Federation of Clinical Chemistry and Laboratory Medicine* 29 (2018), pp. 215–220.

[11] Christopher E Mason, Paul Zumbo, Stephan Sanders. “Standardizing the next generation of bioinformatics software development with BioHDF (HDF5)”. In: *Advances in Experimental Medicine and Biology* 680 (2010), pp. 693–700. DOI: https://doi.org/10.1007/978-1-4419-5913-3_77.

[12] Y. J. Erwin L. van Dijk Hélène Auger. “Ten years of next-generation sequencing technology”. In: *Trends in Genetics* 30 (2014), pp. 418–426. DOI: https://doi.org/10.1016/j.tig.2014.07.001.

[13] Evan E. Eichler. “Genetic Variation, Comparative Genomics, and the Diagnosis of Disease”. In: *The New England Journal of Medicine* 381 (2019), pp. 64–74. DOI: https://doi.org/10.1056/NEJMra1809315.

[14] Findlay Bewicke-Copley, Emil Arjun Kumar, Giuseppe Palladino. “Applications and analysis of targeted genomic sequencing in cancer studies”. In: *Computational and Structural Biotechnology Journal* 17 (2019), pp. 1348–1359. DOI: https://doi.org/10.1016/j.csbj.2019.10.004.

[15] F. E. B. Franco Pagani. “Genomic variants in exons and introns: identifying the splicing spoilers”. In: *Nature Reviews Genetics* 5 (2004), pp. 389–396. DOI: https://doi.org/10.1038/nrg1327.

[16] S. Genetics. *Truvari*. [https://github.com/spiralgenetics/truvari].

[17] GenXys. *CNV testing should be universally used in clinical genetic testing*. https://www.genxys.com/content/cnv-testing-should-be-universally-used-in-clinical-genetic-testing/ [Accessed 3 June 2021].

[18] H.P.J. Buermans, J.T. den Dunnen. “Next generation sequencing technology: Advances and applications”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842 (2014), pp. 1932–1941. DOI: https://doi.org/10.1016/j.bbadis.2014.06.015.

[19] E. D. Haley J. Abel. “Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches”. In: *Cancer Genet* 12.206 (2015), pp. 432–440. DOI: https://doi.org/10.1016/j.cancergen.2013.11.002.

[20] Hao Zhao, Zhifu Sun, Jing Wang. "CrossMap: a versatile tool for coordinate conversion between genome assemblies". In: *Bioinformatics* 30 (2014), pp. 1006–1007. DOI: https://doi.org/10.1093/bioinformatics/btt730.

[21] D. Heller. "Structural variant calling using third-generation sequencing data". PhD thesis. Free University of Berlin, 2021.

[22] Illumina. *Advantages of paired-end and single-read sequencing*. [https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html].

[23] Illumina. *Di erences Between NGS and Sanger Sequencing*. Available at: https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html [Accessed 31 January 2021].

[24] S. J. S. Ivan Iossifov Brian J. O'Roak. "The contribution of de novo coding mutations to autism spectrum disorder". In: *Nature* 515 (2014), pp. 216–221. DOI: https://doi.org/10.1038/nature13908.

[25] E. W. M. J. Craig Venter Mark D. Adams. "The Sequence of the Human Genome". In: *Science* 291.5507 (2001), pp. 1304–1351. DOI: http://dx.doi.org/10.1126/science.1058040.

[26] N. A. Jake Lever Martin Krzywinski. "A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data". In: *Genome medicine* 12.14 (2020). DOI: https://doi.org/10.1186/s13073-020-0712-0.

[27] N. A. Jake Lever Martin Krzywinski. "Classification evaluation". In: *Nature methods* 13 (2016), pp. 603–604. DOI: https://doi.org/10.1038/nmeth.3945.

[28] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler. "Integrative Genomics Viewer". In: *Nature Biotechnology* 29 (2011), pp. 24–26. DOI: https://doi.org/10.1038/nbt.1754.

[29] J. P. A. Jan O. Korbel Alexander Eckehart Urban. "Paired-End Mapping Reveals Extensive Structural Variation inthe Human Genome". In: *Science* 318 (2007), pp. 420–426. DOI: https://doi.org/10.1126/science.1149504.

[30] Jasin Hodzic, Lejla Gurbeta, Enisa Omanovic-Miklicanin, Almir Badnjevic. "Overview of Next-generation Sequencing Platforms Used in Published Draft Plant Genomes in Light of Genotypization of Immortelle Plant (Helichrysium Arenarium)". In: *Medical Archives* 71 (2017), pp. 288–292. DOI: https://doi.org/10.5455/medarh.2017.71.288-292.

[31] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick. “An integrated semiconductor device enabling non-optical genome sequencing”. In: *Nature* 475 (2011), pp. 348–352. DOI: https://doi.org/10.1038/nature10242.

[32] J. T. Jonathan Sebat B. Lakshmi. “Large-Scale Copy Number Polymorphism in the Human Genome”. In: *Science* 305 (2004), pp. 525–528. DOI: https://doi.org/10.1126/science.1098918.

[33] José Marcos Moreno-Cabrera, Jesús del Valle, Elisabeth Castellanos. “Evaluation of CNV detection tools for NGS panel data in geneticdiagnostics”. In: *European Journal of Human Genetics* 28 (2020), pp. 1645–1655. DOI: https://doi.org/10.1038/s41431-020-0675-z.

[34] Justin M. Zook, Nancy F. Hansen, Nathan D. Olson, Lesley M. Chapman, James C. Mullikin, Chunlin Xiao, Stephen Sherry, Sergey Koren, Adam M. Phillippy, Paul C. Boutros, Sayed Mohammad E. Sahraeian, Vincent Huang, Alexandre Rouette, Noah Alexander, Christopher E. Mason, Iman Hajirasouliha, Camir Ricketts, Joyce Lee, Rick Tearle, Ian T. Fiddes, Alvaro Martinez Barrio, Jeremiah Wala, Andrew Carroll, Noushin Ghaffari, Oscar L. Rodriguez, Ali Bashir, Shaun Jackman, John J Farrell, Aaron M Wenger, Can Alkan, Arda Soylev, Michael C. Schatz, Shilpa Garg, George Church, Tobias Marschall, Ken Chen, Xian Fan, Adam C. English, Jeffrey A. Rosenfeld, Weichen Zhou, Ryan E. Mills, Jay M. Sage, Jennifer R. Davis, Michael D. Kaiser, John S. Oliver, Anthony P. Catalano, Mark JP Chaisson, Noah Spies, Fritz J. Sedlazeck, Marc Salit, the Genome in a Bottle Consortium. “A robust benchmark for germline structural variant detection”. In: (2019). Available at: https://doi.org/10.1101/664623 [Accessed 7 November 2020].

[35] Z. N. Kai Ye George Hall. “Structural Variation Detection from Next Generation Sequencing”. In: *Journal of Next Generation Sequencing and Applications* (2016). DOI: https://doi.org/10.4172/2469-9853.S1-007.

[36] J. A. Kevin L Howe Premanand Achuthan. “Ensembl 2021”. In: *Nucleic Acids Research* 49 (2021), pp. D884–D891. DOI: https://doi.org/10.1093/nar/gkaa942.

[37] G. M. C. Kimberly Robasky Nathan E. Lewis. “The role of replicates for error mitigation in next-generation sequencing”. In: *Nature reviews, Genetics* 15.1 (2014), pp. 56–62. DOI: http://dx.doi.org/10.1038/nrg3655.

[38] D. C. Koboldt. “Best practices for variant calling in clinical sequencing”. In: *Genome Medicine* 12.91 (2020). DOI: https://doi.org/10.1186/s13073-020-00791-w.

[39] J. K. Kulski. *Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications*. 2016. DOI: https://doi.org/10.5772/61964.

[40] S. W. S. Lars Feuk Andrew R. Carson. "Structural variation in the human genome". In: *Nature reviews. Genetics* 7 (2006), pp. 85–97. DOI: https://doi.org/10.1038/nrg1767.

[41] H. Li. "Toward better understanding of artifacts in variant calling from high-coverage samples". In: *Bioinformatics* 30.20 (2014), pp. 2843–2851. DOI: https://doi.org/10.1093/bioinformatics/btu356.

[42] Margaret Morash, Hannah Mitchell, Himisha Beltran. "The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology". In: *Journal of Personalized Medicine* 8 (2018). DOI: https://doi.org/10.3390/jpm8030030.

[43] Maria Weronika Gutowska-Ding, Zandra C.Deans, Christophe Roos, Jukka Matilainen. "One byte at a time: evidencing the quality of clinical service next-generation sequencing for germline and somatic variants". In: *European Journal of Human Genetics* 28 (2019). DOI: https://doi.org/10.1038/s41431-019-0515-1.

[44] J. K. Martin Kircher. "High-throughput DNA sequencing –concepts and limitations". In: *Bioessays* 32.6 (2010), pp. 524–536. DOI: https://doi.org/10.1002/bies.200900181.

[45] Matthew Ezewudo, Michael E. Zwick. "Evaluating Rare Variants in Complex Disorders Using Next-Generation Sequencing". In: (2013). Available at: https://doi.org/10.1007/s11920-013-0349-4 [Accessed 31 January 2021].

[46] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz Fritz J. Sedlazeck. "Structural variant calling: the long and the short of it". In: (2019). Available at: https://doi.org/10.1186/s13059-019-1828-7 [Accessed 31 January 2021].

[47] P. P. Z. Mehdi Pirooznia Fernando S. Goes. "Whole-genome CNV analysis: advances in computational approaches". In: *Frontiers in Genetics* 6.138 (2015). DOI: https://doi.org/10.3389/fgene.2015.00138.

[48] Michael A Quail, Miriam Smith, Paul Coupland. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC Genomics* 13 (2012). DOI: https://doi.org/10.1186/1471-2164-13-341.

[49] Y. M. Michal Levy-Sakin Steven Pastor. “Genome maps across 26 human populations reveal population-specific patterns of structural variation”. In: *Nature Communications* 1025 (2019). DOI: https://doi.org/10.1038/s41467-019-08992-7.

[50] Q. W. Min Zhao Qingguo Wang. “Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives”. In: *BMV Bioinformatics* 14.S1 (2013). DOI: https://doi.org/10.1186/1471-2105-14-S11-S1.

[51] Miten Jain, Hugh E. Olsen, Benedict Paten, Mark Akeson. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. In: *Genome Biology* 17 (2016). DOI: https://doi.org/10.1186/s13059-016-1103-0.

[52] Nancy F. Hansen. “SVanalyzer”. In: (). DOI: https://github.com/nhansen/SVanalyzer.

[53] Natasha T. Hill, David Kim, Klaus J. Busam. “Distinct Signatures of Genomic Copy Number Variants Define Subgroups of Merkel Cell Carcinoma Tumors”. In: *Cancers* 13 (2021). DOI: https://doi.org/10.3390/cancers13051134.

[54] S. M. R. P.J. Hastings James R. Lupski. “Mechanisms of change in gene copy number”. In: *Nature Reviews Genetics* 10 (2009), pp. 551–564. DOI: https://doi.org/10.1038/nrg2593.

[55] Petr Danecek, Adam Auton, Goncalo Abecasis. “The variant call format and VCFtools”. In: *Bioinformatics* 27 (2011), pp. 2156–2158. DOI: https://doi.org/10.1093/bioinformatics/btr330.

[56] H. L. Rehm. “Disease-targeted sequencing: a cornerstone in the clinic”. In: *Nature Reviews Genetics* 4 (2013), pp. 295–300. DOI: https://doi.org/10.1038/nrg3463.

[57] K. R. F. Richard Redon Shumpei Ishikawa. “Global variation in copy number in the human genome”. In: *Nature* 444 (2006), pp. 444–454. DOI: https://doi.org/10.1038/nature05329.

[58] rina Abnizova1, Rene te Boekhorst, Yuriy L Orlov. “Computational Errors and Biases in Short Read Next Generation Sequencing”. In: *Journal of Proteomics Bioinformatics* 10 (2017). DOI: https://doi.org/10.4172/jpb.1000420.

[59] Roland Wittler, Tobias Marschall, Alexander Schönhuth, Veli Mäkinen. “Repeat- and error-aware comparison of deletions”. In: *Bioinformatics* 31 (2015), pp. 2947–2954. DOI: https://doi.org/10.1093/bioinformatics/btv304.

[60] C. E. L. Ryan E Mills Christopher T Luttig. “An initial map of insertion and deletion (INDEL) variation in the human genome”. In: *Genome Research* 16 (2006), pp. 1182–1190. DOI: https://doi.org/10.1101/gr.4565806.

[61] W. R. M. Sara Goodwin John D. McPherson. “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17 (2016), pp. 333–351. DOI: https://doi.org/10.1038/nrg.2016.49.

[62] L. Sastre. “Exome sequencing: what clinicians need to know”. In: *Advances in Genomics and Genetics* 4 (2014), pp. 15–27. DOI: https://doi.org/10.2147/AGG.S39108.

[63] A. A. Sen Zhao Oleg Agafonov. “Accuracy and efficiency of germline variant calling pipelines for human genome data”. In: *Scientific Reports* 10.20222 (2020). DOI: https://doi.org/10.1038/s41598-020-77218-4.

[64] V. M. Seungtai Yoon Zhenyu Xuan. “Sensitive and accurate detection of copy numbervariants using read depth of coverage”. In: *Genome Research* 19 (2009), pp. 1586–1592. DOI: https://doi.org/10.1101/gr.092981.109.

[65] Shanika L. Amarasinghe, Shian Su, Xueyi Dong. “Opportunities and challenges in long-read sequencing data analysis”. In: *Genome Biology* 21 (2020). DOI: https://doi.org/10.1186/s13059-020-1935-5.

[66] S. Y. Shazia Mahamdallie Elise Ruark. “The ICR96 exon CNV validation series: a resource for orthogonal assessment of exon CNV calling in NGS data”. In: *Wellcome Open Research* 2 (2017), pp. 35–35. DOI: https://doi.org/10.12688/wellcomeopenres.11689.1.

[67] C. S. K. Shu Mei Teo Yudi Pawitan. “Statistical challenges associated with detecting copy number variations with next-generation sequencing”. In: *Bioinformatics* 28 (2012), pp. 2711–2718. DOI: https://doi.org/10.1093/bioinformatics/bts535.

[68] S. L. M. Simon P Sadedin Justine A Ellis. “Ximmer: a system for improving accuracy and consistency of CNV calling from exome data”. In: *GigaScience* 7 (2018). DOI: https://doi.org/10.1093/gigascience/giy112.

[69] Sohyun Hwang, Eiru Kim, Insuk Lee, Edward M. Marcotte. “Systematic comparison of variant calling pipelines using gold standard personal exome variants”. In: *Scientific Reports* 5 (2015). DOI: https://doi.org/10.1038/srep17875.

[70] “Standardizing data”. In: *Nature Cell Biology* 10 (2008), pp. 1123–1124. DOI: https://doi.org/10.1038/ncb1008-1123.

[71] R. E. M. Steve S. Ho Alexander E. Urban. “Structural variation in the sequencing era”. In: *Nature Reviews Genetics* 21 (2020), pp. 171–189. DOI: https://doi.org/10.1038/s41576-019-0180-9.

[72] The 1000 Genomes Project Consortium. “A global reference for human genetic variation”. In: *Nature* 526 (2015), pp. 68–74. DOI: https://doi.org/10.1038/nature15393.

[73] Tobias P. Loka, Simon H. Tausch, Bernhard Y. Renard. “Reliable variant calling during runtime of Illumina sequencing”. In: *Scientific Reports* 9 (2019). DOI: https://doi.org/10.1038/s41598-019-52991-z.

[74] S. L. S. Todd J. Treangen. “Repetitive DNA and next-generation sequencing: computational challenges and solutions”. In: *Nature Reviews Genetics* 13.1 (2012), pp. 36–46. DOI: https://doi.org/10.1038/nrg3117.

[75] TP Whitehead, FP Woodford. “External quality assessment of clinical laboratories in the United Kingdom”. In: *Journal of Clinical Pathology* 34 (1981), pp. 947–957. DOI: https://doi.org/10.1136/jcp.34.9.947.

[76] V. I. Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Doklady Akademii* 163 (1965), pp. 845–848.

[77] S. Välipakka. “Improving CNV detection from short-read MPS data in neuromuscular disorders”. PhD thesis. University of Helsinki, 2020.

[78] W Greg Miller, Graham RD Jones, Gary L Horowitz, Cas Weykamp. “Proficiency Testing/External Quality Assessment: Current Challenges and Future Directions”. In: *Clinical Chemistry* 57 (2011), pp. 1670–1680. DOI: https://doi.org/10.1373/clinchem.2011.168641.

[79] C. L. Wesley J. Delage Julien Thevenon. “Towards a better understanding of the low recall of insertion variants with short-read based variant callers”. In: *BMC Genomics* 21.762 (2020). DOI: https://doi.org/10.1186/s12864-020-07125-5.

[80] Xiaomin Chen, Yutong Kang, Jing Luo. “Next-Generation Sequencing Reveals the Progression of COVID-19”. In: *Frontiers in Cellular and Infection Microbiology* 11 (2021). DOI: https://doi.org/10.3389/fcimb.2021.632490.

[81] Xiaotu Ma, Ying Shao, Liqing Tian. “Analysis of error profiles in deep next-generation sequencing data”. In: *Genome Biology* 20 (2019). DOI: https://doi.org/10.1186/s13059-019-1659-6.

[82] T. P. S. Yuval Benjamini. “Summarizing and correcting the GC content bias in high-throughput sequencing”. In: *Nucleic Acids Research* 40 (2012), e72. DOI: https://doi.org/10.1093/nar/gks001.

[83] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri. “Big Data: Astronomical or Genomical?” In: *Plos Biology* 13 (2015). DOI: https://doi.org/10.1371/journal.pbio.1002195.

[84] A. M. Zachary S. Bohannan. “Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables”. In: *Computational and structural biotechnology journal* 17 (2019), pp. 561–569. DOI: https://doi.org/10.1016/j.csbj.2019.04.002.