Computational prediction of disease-causing CNVs from exome sequence data

Pubudu Saneth Samarakoon

Department of Medical Genetics, Faculty of Medicine

UNIVERSITY OF OSLO

June 2016
Acknowledgements

First and foremost I want to thank my supervisors; Robert Lyle and Torbjørn Rognes, you have been tremendous mentors for me. Robert, thank you very much for listening to all my questions and ideas, and guiding me during this period. Your input on academic writing and our discussions on various scientific topics truly helped me to grow as a research scientist. I am grateful to Torbjørn for introducing me to advance computational methods and for sharing your programming knowledge. I am thankful to Norwegian quota scholarship fund for providing financial support for the PhD.

I would like to express my gratitude to our research group, Hanne Sorte, Asbjørg Stray-Pedersen and Olaug Rødningen. Our fruitful discussions were invaluable and clearly helped me to expand my skills and expertise to new horizons.

A warm thank to all my colleagues and friends at the department of Medical Genetics, University of Oslo. I owe a special thanks to my dear colleagues Arvind Sundaram and Marius Bjørnstad for sharing the office, serious discussions and for always lightening up my day.

I am truly thankful to my family and friends in Norway and Sri Lanka, especially my wife Nisanka for sharing all my ups and downs. Also baby Sayon, who joined my life in the latter part of my PhD.
List of publications

Paper 1

Paper 2

Paper 3
Stray-Pedersen A, Sorte HS, Samarakoon PS, Gambin T, Chinn IK, Akdemir ZHC and et al. Primary immunodeficiency diseases - genomic approaches delineate heterogeneous Mendelian disorders. Journal of Allergy and Clinical Immunology 2016;
# Table of content

1 Introduction ................................................................................................................. 1  
  1.1 The human genome ................................................................................................. 1  
  1.2 Variation in the human genome ............................................................................... 2  
  1.3 Genetic diseases .................................................................................................... 3  
  1.3.1 Disease-causing genetic variants ..................................................................... 4  
  1.4 Diagnosis of genetic disease .................................................................................. 7  
  1.4.1 Genetic tests to detect SNVs and indels ............................................................. 7  
  1.4.2 Genetic tests to detect chromosomal abnormalities and CNVs .................... 8  
  1.5 Exome sequencing ............................................................................................... 9  
   1.5.1 Exome-based disease diagnosis ..................................................................... 9  
2 Disease-causing CNV detection from exome data ....................................................... 12  
  2.1 CNV prediction ..................................................................................................... 12  
  2.2 Coverage-based CNV prediction .......................................................................... 14  
  2.3 CNV annotation ..................................................................................................... 17  
  2.4 Causal CNV detection ........................................................................................... 18  
  2.5 Validation of disease-causing CNVs .................................................................... 18  
3 Aims of the study ....................................................................................................... 20  
4 Methodological considerations .................................................................................... 21  
  4.1 Exome sequence analysis pipeline ....................................................................... 21  
  4.2 CNV prediction from exome sequence data ......................................................... 23  
  4.3 Resolution of microarray platforms ...................................................................... 23  
  4.4 Considerations when identifying true positive, false positive and false negative CNVs ..... 24  
5 Summary of papers .................................................................................................. 28  
  5.1 Paper 1 Identification of copy number variants from exome sequence data ........ 28  
  5.2 Paper 2 cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data ................. 30  
  5.3 Paper 3 Primary immunodeficiency diseases - genomic approaches delineate heterogeneous Mendelian disorders ................................................................. 31  
6 Discussion .................................................................................................................. 33  
  6.1 Performance comparison of CNV prediction programs ....................................... 33  
  6.2 Effect of cnvScan in-house database in CNV analysis ......................................... 35  
  6.3 Characterization of true positives excluded in cnvScan implementation ............ 37  
  6.4 Comparison with other disease-causing CNV detection methods ......................... 40  
  6.4.1 Advantages of cnvScan annotation compared to other annotation programs ..... 40  
  6.4.2 Rate of disease-causing CNV detection ........................................................... 41  
  6.4.3 Short CNV prediction ...................................................................................... 43  
  6.4.4 Utility of exaCGH in detecting short CNVs ..................................................... 44  
7 Future directions ....................................................................................................... 45  
  7.1 Optimizations to reduce the batch effect in cnvScan implementation ................ 45  
  7.2 Whole-genome CNV analysis .............................................................................. 45  
Abbreviations .................................................................................................................. 47  
Commonly used terms ................................................................................................. 48  
References ...................................................................................................................... 50
1 Introduction

1.1 The human genome

The human genome contains ~3.2 billion bases packaged into 23 pairs of chromosomes. Since the completion of the human genome project [1, 2], scientists have been motivated to develop high-resolution maps and annotation datasets representing a broad spectrum of genomic elements in the human genome. These maps and genome annotation datasets include genes and other functional elements, repetitive elements, mobile elements and genomic variants.

In 2004, with the availability of the complete euchromatic sequence of the human genome, scientists estimated that the human genome contains 20000-25000 protein-coding genes [3], which was recently revised downwards to 19,815 (GENCODE version 24; August 2015 freeze, GRCh38) [4]. These genes contain exons (protein-coding regions) that account for ~1.8% of the genome, introns (non-coding regions) and regulatory regions.

RefSeq [5, 6], ENCODE [7], HAVANA [8] and CCDS [9] were the early initiatives, which are still involved in providing comprehensive genomic annotations. For example, the ENCODE project was originally launched to identify and characterize functional elements in 1% of the human genome [7, 10], and later it was scaled up to provide integrated high-quality annotations for the entire genome (GENCODE) [11, 12]. Similarly, HAVANA, initiated by Ensembl, is dedicated to providing high-quality manually curated genes.

In order to effectively utilize this vast amount of information, various visualization solutions were developed for scientists to browse the entire genome with relevant annotations. NCBI [13], Ensembl [14], UCSC genome browser [15] were a few of the key efforts initiated to provide broader access to this information.

Today, various methods (eg. genetic, evolutionary and biochemical approaches) are used to improve our knowledge on structure and content of the human genome, propelling a broad
range of current research and applications [16]. In particular, this has fuelled the development of systematic approaches that identify genes and genetic variants underlying human diseases.

1.2 Variation in the human genome

Variants in the human genome are the genetic differences between individuals, within and among populations. Genetic variants are changes in DNA sequence or structural alterations of the genome and are classified into three groups: sequence variants (eg. single nucleotide variants, insertions and deletions), structural variants (eg. copy number variants, translocations and inversions) and changes in chromosome number (eg. aneuploidy).

Single nucleotide variants (SNVs) are single base pair changes in the genome of an organism. Insertions and deletions (indels) are changes in DNA sequence where short DNA fragments are inserted or deleted [17]. Chemical and physical mutagens, and errors occurring during DNA replication can induce SNV and indel formation.

Copy number variants are a class of structural variants containing deletions and duplications [18]. Errors in DNA repair and DNA replication pathways contribute to CNV formation via several mechanisms (eg. non-allelic homologous recombination - NAHR [19], non-homologous end joining - NHEJ [20], microhomology mediated end joining - MMEJ [20], fork stalling and template switching - FoSTeS [21] and microhomology mediated break induced replication - MMBIR [22]). The majority of these mechanisms are mediated by homologous or microhomologous regions [23-26]. Thus, most of the CNVs in human genome are enriched within and near these regions (eg. 70% of the deletion breakpoints are in microhomologous regions) [27, 28].

Our current understanding of genetic variation comes from analysing the genomes of a large number of individuals. The human genome project [3], the SNP consortium (dbSNP) [29] and the international HapMap project [30, 31] were the early efforts that identified ~10 million variants, primarily SNVs. Early large-scale efforts focused on detecting CNVs include the HapMap project phase III [32] and Wellcome Trust Sanger CNV project [33, 34]. Collectively these projects identified a vast amount of common variants (frequency >5%) found in a limited number of populations.
While these initial efforts were limited by the available technology, recent studies, fuelled by more advanced variant detection methods (section 1.5) characterized a large amount of common and rare variants in a broad range of populations worldwide. For example, a total of 88 million variants, including 84.7 million SNVs, 3.6 million indels and 60000 structural variants, were characterized with the completion of the 1000 genomes project [35-37]. In addition to this, the NHLBI GO Exome Sequencing Project (ESP) [38, 39] and UK10K project [40] were other efforts initiated to identify common and rare genetic variants.

Considering all the variants identified from these projects, scientists have estimated the average genome-wide variant counts (SNVs: ~3.7 million and indels: ~35000 [36]) and the average genome-wide coverage of CNVs (contributing to 217.1Mbp, 7.01% of the genome) [41]. Based on these figures, SNVs are the most frequent type of variant and structural variants affect more bases than other types of genetic variants in the genome of an individual [37].

With the goal of improving personal health, these vast amounts of population-specific genomic variants were made available in public data repositories [42]. Current diagnostic approaches use these datasets to help distinguish disease and non-disease causing variants in patients with genetic diseases.

1.3 Genetic diseases

Today, the vast majority of human diseases are believed to have some genetic contribution. Based on the nature of this genetic contribution, these diseases are grouped into Mendelian diseases, complex diseases and chromosomal diseases.

Mendelian diseases [43] are caused by mutations at a single genetic locus that segregates in families according to Mendel’s theory of inheritance [44]. Therefore Mendelian diseases are also known as single-gene diseases. In contrast, complex diseases are associated with multiple factors (gene and environmental factors) and do not follow Mendel’s laws [45]. While the majority of diseases are caused by alterations in genes, chromosomal aberrations
can change the structure of the genome and cause chromosomal diseases including aneuploidy and a range of cancer types [46].

1.3.1 Disease-causing genetic variants

About 4000 disease-related genes have been identified to date [47]. These genes carry variants (SNVs, indels or CNVs) that are either associated with or cause human diseases.

SNVs in exonic regions may cause genetic diseases by affecting the structure of proteins (primary and secondary) [48], mRNA stability [49] or gene expression (eg. missense and nonsense mutations). Indels in exonic regions also cause diseases when a variant alters the reading frame of a gene (eg. frame shift mutation) [17]. SNVs and indels in gene regulatory regions may affect gene expression and cause diseases. Today, a broad spectrum of diseases ranging from infectious diseases to cancer have been associated with SNVs and indels [50-52].

While SNVs and indels account for the majority of disease-causing variants identified to date [53], CNVs represent an important subset of variants that cause genetic diseases. CNVs can cause a broad range of common and complex diseases [54, 55] via three mechanisms (gene interruption, dosage effect and position effect) that affect either the structure of a gene or gene expression (Fig. 1, Table 1).
Fig. 1 Gene interruption

a. CNVs affecting single genes (partial or complete, deletions or duplications); b. Unmasking a recessive mutation by a heterozygous deletion; c. Gene interruption and fusion gene formation.
### Disease causing mechanism

<table>
<thead>
<tr>
<th>Gene interruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Partial gene deletion or duplication (Fig. 1A)</td>
</tr>
<tr>
<td>ii. Heterozygous deletion can create hemizygous locus with a recessive mutation (Fig. 1B)</td>
</tr>
<tr>
<td>iii. CNV with multiple genes can create a fusion gene that encodes a mRNA with an open reading frame (Fig. 1C)</td>
</tr>
<tr>
<td>iv. Combination of multiple CNVs that do not individually produce phenotypes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene dosage</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Deletion of dosage-sensitive gene (gene deficiency)</td>
</tr>
<tr>
<td>ii. Duplication of a dosage-sensitive gene</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position effect</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Disease causing mechanism</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene interruption</td>
<td>Change the gene structures and cause diseases. Cause diseases by interrupting a single gene. Partial deletion of <em>PDXDC1</em> is associated with hearing loss [56]. Muscular dystrophy and lipoprotein lipase deficiency are associated with partial duplications [57]. Cause diseases via loss of function mechanism. Opsoclonus-myoclonus ataxia-like syndrome [58], Charcot–Marie–Tooth disease type 1 [59] and Walker–Warburg syndrome [60] are associated with compound heterozygous changes. Cause diseases via gain of function mechanism. In recent years, CNV breakpoints were examined to identify fusion genes that affect intellectual disability, developmental delay, autism spectrum disorders, congenital anomalies and dysmorphic features [61, 62]. Cause complex diseases. Autism, mental retardation [63] and developmental disorders like Smith-Magenis Syndrome [64] are associated with these complex events.</td>
</tr>
<tr>
<td>Gene dosage</td>
<td>CNVs that change the number of copies of a gene present in the genome can alter the gene expression and cause diseases. Gene deficiency in haploinsufficient genes can cause diseases [65, 66]. Increase in gene dosage can create imbalances due to excess gene product and cause diseases [67].</td>
</tr>
<tr>
<td>Position effect</td>
<td>CNVs can alter the gene environment (enhancers, silencers and local chromatin organization), which may change the expression of neighbouring copy number neutral genes and cause diseases.</td>
</tr>
</tbody>
</table>

**Table 1** Disease causing mechanisms of CNVs

In addition to the direct involvement of genomic variants in genetic diseases, variants in modifier genes can modulate the severity of disease-related phenotypes [68-70]. For example, increased copy number of *SMN2* is correlated with a milder phenotype of spinal muscular atrophy caused by a mutation in *SMN1* [71].
Today, linkage studies, genome-wide association studies (GWAS) [72], microarray-based studies and high-throughput sequencing-based approaches have discovered a large number of clinically-relevant variants that are catalogued in a number of different databases. For example, the human genome mutation database (HGMD) [73] lists ~175,000 mutations that are associated with or cause genetic diseases [53]. Similarly, ClinVar [74] archives ~126,000 genomic variants with clinically relevant phenotypes and DECIPHER database archives a large number of clinically relevant CNVs [75]. Additionally, OMIM provides a detailed gene-centric and disease-centric view of clinically relevant variants [76, 77].

While HGMD, ClinVar, DECIPHER and OMIM are the most commonly used databases with variants contributing to a broad range of human diseases, a number of disease-centered databases are available with a wealth of information on certain phenotypes [78]. All these datasets improve the understanding of the mutational spectrum of a particular gene or disease and they are widely used in diagnosing genetic diseases.

### 1.4 Diagnosis of genetic disease

Genetic diagnosis is the identification of genomic variants, genes or gene products that are associated with human diseases [79]. The classical genetic diagnostic decision-making process is guided by a detailed family history study that reveals the mode of inheritance of the disease. Then, genetic tests are performed to confirm the diagnosis of a symptomatic individual or individual with a family history of a specific disease. These genetic tests are designed to detect a broad range of genomic variants [80].

#### 1.4.1 Genetic tests to detect SNVs and indels

Since the advent of genetic testing in clinical laboratories, various techniques have been introduced to detect disease-causing SNVs and indels [79, 81]. For example, restriction fragment analysis [82], DNA sequencing (Sanger sequencing) [83] and allele-specific oligonucleotide ligation assays [84, 85] are some of the initial tests developed. While these techniques are not widely used in current protocols, a few remain useful in causal variant identification. Particularly, Sanger sequencing is still used in diagnosing well-characterized
Mendelian disease genes (eg. cystic fibrosis: CFTR, Rett syndrome: MECP2) and to confirm disease-causing SNVs identified by other methods [86].

The genetic diagnostic tests discussed above are designed to target specific loci, generally identified by accurate phenotyping of patients and prior knowledge of the genetics of the disease. Clinical and genetic heterogeneity can complicate the diagnostic process by making it difficult to identify a likely causal gene from many candidates. Hence, targeted genetic testing for these genes are time consuming, cost ineffective and have low success rates. Furthermore, these tests have limited applicability in detecting rare variants and novel variants. However, as discussed in section 1.5, current diagnostic protocols use high-throughput sequencing based methods to detect disease-causing variants.

1.4.2 Genetic tests to detect chromosomal abnormalities and CNVs

Karyotyping was introduced as the first diagnostic test to detect chromosomal abnormalities [87]. Then, fluorescence in situ hybridisation (FISH) was introduced as a targeted method, which detects submicroscopic anomalies in chromosomes [88, 89]. Next, comparative genomic hybridisation (CGH) techniques [90] marked a turning point in genetic diagnosis by enabling genome-wide CNV detection. CGH was further improved using microarrays to perform high-resolution genome-wide CNV detection [91]. These microarrays contained 1-4 million 45-60bp long probes that target specific sequences spread over the entire human genome sequence [92]. Then, SNV genotypes were used as genetic markers [93] in developing genotype-arrays as an alternative platform to perform genome-wide CNV detection [94, 95].

When microarray platforms were introduced, array results were validated using alternative targeted methods. Initially, quantitative fluorescence PCR (QF-PCR) was used to validate aCGH results and later multiplex ligation-dependent probe amplification (MLPA) was also used to validate CNVs [96]. However, with recent developments in aCGH probe design and advance bioinformatics algorithms, aCGH has become the principal CNV detection platform in current genetic diagnostic protocols [92, 97].
Microarrays offer reliable and efficient methods for genome-wide population-scale CNV detection [98]. However, the sensitivity and breakpoint precision of these platforms vary greatly depending on their probe density distributions. Therefore they have variable performance in detecting CNVs [99, 100]. For example, a microarray benchmark study has shown that large CNVs (>1Mb) were detected with high accuracy even in arrays that have different probe distributions, but these arrays have shown variable efficiencies in detecting short CNVs [101]. With the availability of high-throughput sequencing data, alternative sequencing-based approaches have been developed which can provide higher resolution CNV detection.

1.5 Exome sequencing

With the introduction of next generation high-throughput sequencing technologies [102], whole genomes could rapidly be characterized at single base pair resolution. The continuous development of these technologies lead to the rapid decrease in sequencing cost and therefore massive amounts of sequence data were generated with ever-decreasing effort [103-105]. About six years ago, when whole-genome sequencing was not available at an affordable price, exome sequencing was introduced as a cost-effective approach to sequencing coding regions of the human genome [106, 107].

When exome sequencing was first introduced, it was estimated that ~85% of clinically relevant variants were in exonic regions [108]. While this figure is likely an overestimate due to the scarcity of data on disease causality, scientists still expect the majority of causal variants to be in the exome. Today, exome sequencing is used as an effective genetic diagnostic test that has revolutionized the detection of novel, de novo and rare variants across a wide range of diseases [108-112].

1.5.1 Exome-based disease diagnosis

The genome of a healthy individual differs from the human reference genome at 4.1-5.0 million sites [37]. These differences are due to genetic variants and exome sequencing has the potential to reveal these variants that are in coding regions. For example, variant calling programs identify 30000-50000 SNVs and 2000-6000 indels per exome (passing the quality
thresholds used in bioinformatics pipelines). Therefore identifying disease-causing mutations from thousands of normal variants is the most challenging task in current genetic diagnostic approaches. In order to address this issue, several different diagnostic strategies and guidelines were introduced in recent years, laying the foundation to identify disease-causing variants [113-115].

As detailed in these diagnostic guidelines, current pipelines use a number of different prioritization approaches to select potentially disease-causing variant from patient samples. For example, disease-causing variants can be found in candidate genes common to a group of affected unrelated individuals or the same causal variant could be found in affected family members. Similarly, de novo causal variants can be found by using a family trio (mother-father-child). Therefore exome-based diagnostic methods can be designed to increase the probability of identifying the causal variant. When studying diseases that are known for genetic heterogeneity, candidate genes shared among subsets of individuals can be selected during variant prioritization.

Variant prioritization can be further improved by using population-specific datasets and clinically relevant datasets. Here, population-specific datasets are used to filter-out common variants and clinically relevant datasets are used to identify already known disease-relevant variants (Table 2). Following prioritization, the next step is to identify a pathogenic variant from the prioritized list. Here, a broad range of datasets including scoring systems that improve variant interpretation [115], published literature, functional genomic annotations and model organism phenotypes can be used to assess the pathogenicity of a genetic variant.

Current variant calling programs [116] detect SNVs [117] and indels [118, 119] with high accuracy, and annotation programs provide a wealth of annotations to assist the disease-causing variant detection [120, 121]. Today, exome-based diagnostic approaches have shown a success rate of 25%-40% in detecting disease-causing SNVs and indels [122, 123]. However, only a 2%-3% CNV detection rate has been published for heterogeneous Mendelian cohorts [124, 125].
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Application</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population-specific datasets</strong></td>
<td><strong>Used to identify known:</strong></td>
<td></td>
</tr>
<tr>
<td>1000 genomes data</td>
<td>SNVs, indels and CNVs</td>
<td>[35-37, 42]</td>
</tr>
<tr>
<td>NHLBI GO Exome Sequencing Project</td>
<td>SNVs and indels</td>
<td>[38]</td>
</tr>
<tr>
<td>Database of Genomic Variants</td>
<td>CNVs</td>
<td>[126]</td>
</tr>
<tr>
<td>Sanger CNV project</td>
<td>CNVs</td>
<td>[34]</td>
</tr>
<tr>
<td><strong>Clinically-relevant datasets</strong></td>
<td><strong>Used to identify disease-relevant:</strong></td>
<td></td>
</tr>
<tr>
<td>OMIM</td>
<td>Genes containing causal variants</td>
<td>[76, 77]</td>
</tr>
<tr>
<td>HGMD</td>
<td>SNVs, indels and CNVs</td>
<td>[73]</td>
</tr>
<tr>
<td>ClinVar</td>
<td>SNVs, indels and CNVs</td>
<td>[74]</td>
</tr>
<tr>
<td>DECIPHER database (Database of</td>
<td>CNVs</td>
<td>[127]</td>
</tr>
<tr>
<td>Chromosomal Imbalance and Phenotype</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in Humans Using Ensembl Resources)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deciphering Developmental Disorders (DDD)</td>
<td>SNVs, indels and CNVs</td>
<td>[128]</td>
</tr>
</tbody>
</table>

**Table 2** Public datasets used in exome-based disease diagnosis
2 Disease-causing CNV detection from exome data

Since the introduction of exome sequencing as an effective diagnostic tool, various CNV prediction and annotation programs have been developed to detect disease-causing CNVs from patient exomes. These disease-causing CNV detection methods can be discussed in four main steps: CNV prediction, annotation, causal CNV detection and validation.

2.1 CNV prediction

Computational programs use paired-end or coverage data to predict CNVs from sequence data. Based on the most commonly used CNV prediction strategies, these programs are grouped into three categories: paired-end-, split-read- and coverage-based CNV prediction.

In short-read sequencing (eg. Illumina sequencing [129]), paired-end reads are generated with approximately known distances. Read alignment and mapping programs are expected to maintain these distances [130]. If CNV breakpoints are internal to paired-end intervals, these reads are mapped in wrong orientation or mapped at distances that are substantially different from the expected distances [131, 132]. Therefore paired-end data is used in computational programs to predict CNVs from sequence data (Fig. 2) [133].

Split-read- based CNV prediction programs detect CNV breakpoints using unmapped or partially mapped (soft-clipped) reads (Fig. 3). If unmapped or partially mapped reads contain CNV breakpoints, these programs use split-mapping algorithm to identify CNV breakpoints at single base-pair resolution [131, 132].

Paired-end- and split-read-based methods are widely implemented to call CNVs in whole-genome sequence (WGS) data. For example, BreakDancer [134], PEMer [135], VariationHunter [136], commonLAW [137], GASV [138], AGE [139], Pindel [118], SLOPE [140], SRiC [141] and STRiP [142] predict CNVs in whole-genome sequences. However, when predicting CNVs in exome sequences, paired-end information is only used in rare instances (eg. SPLITREAD [143]).
**Fig. 2** Paired-end-based CNV identification

**a.** Deletion (reads are mapped too far apart); **b.** Insertions (reads are mapped too close together); **c.** Inversions (the orientation of one read in the pair is changed); **d.** Duplications (reads change their relative order while retaining their original orientation)

**Fig. 3** Split-read-based CNV identification

**a.** Deletion (continuous stretch of gaps in the split-read); **b.** Insertion (continuous stretch of gaps in the reference); **c.** Inversion (section of the split-read is mapped in wrong orientation); **d.** Duplication (both sections of split-read alter their relative order during the alignment)
Both paired-end- and split-read-based CNV prediction methods depend on the availability of paired-end reads encompassing or containing CNV breakpoints (Fig. 2, 3). Therefore, when CNV breakpoints are in regions that are not captured in the exome (eg. non-coding regions), these two methods are not capable of detecting CNVs. Therefore coverage-based methods have been implemented to overcome this challenge, when predicting CNVs in exomes.

### 2.2 Coverage-based CNV prediction

In short-read sequencing, copy number variable regions generate higher or lower read counts compared to diploid regions in the genome [144]. Therefore when calling CNVs, prediction programs analyse the coverage distribution of exonic regions, expecting that the coverage of an exon is proportional to the copy number of the exon (Fig. 4).

![Coverage-based CNV prediction](image)

**Fig. 4 Coverage-based CNV prediction**

- **a.** Copy number neutral control dataset; **b.** Increase in coverage of exon 2 (duplication) compared to control dataset; **c.** Decrease in coverage in exon 3 compared to control dataset. Coverage-based programs follow three main stages when predicting CNVs from exome sequence data: coverage extraction, normalization and CNV calling. In the first stage, alignment files (BAM file) are used to extract coverage data for target regions defined in an exome definition file. The majority of programs extract coverage data from multiple exomes,
which are used in the second stage, coverage normalization. Here, extracted coverage data is normalized using different statistical approaches to create a reference (background) dataset. Finally, this normalized reference (background) dataset is used in variant calling algorithms to predict CNVs from target exomes.

The efficiency of coverage-based programs depends on various factors that are associated with the exome sequencing process. In exome sequencing, short exonic regions (about 200bp) and off-target regions (eg. segmental duplicates) are captured with variable efficiencies. This can affect CNV prediction by introducing bias into the statistical models used to call CNVs. In addition to these sequencing artifacts, factors including GC percentage, mappability, duplicated reads and ambiguously mapped reads influence the coverage distribution of target exomes and thereby affect the efficiency of CNV prediction. As discussed in Table 3, a number of different techniques have been implemented in coverage-based programs to minimize the effects from these issues.

<table>
<thead>
<tr>
<th>Issues affecting CNV prediction</th>
<th>Technique used to minimize the effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Duplicated reads</strong></td>
<td>SAMtools rmdup [148] and Picard MarkDuplicates [149] can identify duplicated reads in alignment files. Once identified, duplicated reads can be excluded during coverage data extraction.</td>
</tr>
<tr>
<td><strong>Ambiguously mapped reads</strong></td>
<td>Ambiguously or inaccurately mapped reads have low mapping quality scores [150]. Therefore, quality score threshold can be used to exclude these reads.</td>
</tr>
<tr>
<td><strong>GC content</strong></td>
<td>Local coverage correction methods can reduce the effect of GC content [153].</td>
</tr>
<tr>
<td><strong>Mappability</strong></td>
<td>Local mappability score correction methods can introduce to reduce these effects [147, 155].</td>
</tr>
</tbody>
</table>
Coverage data from regions with low mappability scores can introduce bias to the CNV prediction.

**Technical artefacts in sequencing**
Technical artefacts could occur during the sequencing process and these effects may vary depending on capture technology and sequencing batch.

Advance statistical methods can be implemented to normalize coverage data across the sample collection used for CNV prediction. For example, principal component analysis based methods can be used to reduce these biases [156, 157].

**Table 3 Coverage-based CNV prediction: issues and strategies**

Coverage-based methods were initially introduced to predict CNVs in paired case-control samples. For example, to predict CNVs in a tumour (case) sample compared to a healthy control [158, 159]. Soon, it was implemented in population scale CNV prediction [153]. Table 4 provides a summary of coverage-based programs that can predict CNVs in exome sequence data.

<table>
<thead>
<tr>
<th>Program (Google citations)</th>
<th>Features</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExomeCNV (190)</td>
<td>Use paired case-control exome samples to call CNVs.</td>
<td>[160]</td>
</tr>
<tr>
<td>CONTRA (94)</td>
<td>Create a reference dataset from multiple exomes and call CNVs using a method similar to ExomeCNV.</td>
<td>[161]</td>
</tr>
<tr>
<td>ExomeCopy (32)</td>
<td>Use a negative binomial model to normalize coverage data according to tile length and GC content. CNV calling is performed based on a hidden markov model (HMM).</td>
<td>[162]</td>
</tr>
<tr>
<td>ExomeDepth (99)</td>
<td>Uses a robust strategy to model coverage data and to build an optimized reference set that maximize the CNV detection power.</td>
<td>[163]</td>
</tr>
<tr>
<td>CoNIFER (180)</td>
<td>Performs a singular value decomposition (SVD) based normalization of reads per thousand bases per million reads sequenced (RPKM).</td>
<td>[157]</td>
</tr>
<tr>
<td>XHMM (138)</td>
<td>Uses principal-component analysis (PCA) to normalize coverage data and CNV calling is performed by a HMM.</td>
<td>[156]</td>
</tr>
<tr>
<td>ExoCNVTest (17)</td>
<td>Identify disease-associated CNVs and generate absolute copy number genotypes at putatively associated loci.</td>
<td>[164]</td>
</tr>
<tr>
<td>FishingCNV (26)</td>
<td>Graphical software package to call CNVs using principal component analysis based method.</td>
<td>[165]</td>
</tr>
<tr>
<td>CANOES (9)</td>
<td>Models coverage data using a negative binomial distribution and creates a reference dataset when calling CNVs.</td>
<td>[166]</td>
</tr>
<tr>
<td>PatternCNV (7)</td>
<td>Estimate CNVs based on coverage variability patterns summarized from a set of reference samples.</td>
<td>[167]</td>
</tr>
<tr>
<td>cnvOffSeq (2)</td>
<td>Estimate CNVs using a singular value decomposition based normalization framework for off-target coverage data.</td>
<td>[168]</td>
</tr>
<tr>
<td>EXCAVATOR</td>
<td>Implements a three-step normalization procedure to</td>
<td>[155]</td>
</tr>
</tbody>
</table>
mitigate the effects of sequencing biases and use HMM-based segmentation algorithm to call CNVs.

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>eXome</td>
<td>Use HMM based approach to call CNVs.</td>
<td>[169]</td>
</tr>
<tr>
<td>CODEX</td>
<td>Normalize coverage data to reduce the effect of mappability and GC content.</td>
<td>[170]</td>
</tr>
<tr>
<td>CLAMMS</td>
<td>Highly scalable copy number estimation tool based on lattice-aligned mixture models.</td>
<td>[171]</td>
</tr>
<tr>
<td>CoNVaDING</td>
<td>Includes a stringent quality control (QC) metric that flags low-quality exons when calling CNVs.</td>
<td>[172]</td>
</tr>
</tbody>
</table>

Table 4 Exome sequence-based CNV prediction programs

* Exome-based programs that are specially optimized to detect CNVs in cancer patients include, VarScan 2 [173], exome2cnv [174], CoNVEX [175], CEQer [176] and Control-FREEC [177].

While a number of exome-based CNV prediction programs are currently available, only a few are commonly used and cited in the published literature. For example, a Google citation count of each CNV prediction program (Table 4) indicates that only a few programs are commonly cited by the genomics community. Two benchmark studies have tested these commonly used programs and demonstrated the potential utility in detecting disease-causing CNVs [178, 179].

Despite the potential clinical utility of exonic CNV prediction programs, there are several pitfalls that affect widespread implementation. For example, these programs predict drastically different CNV counts when tested with the same sample collection. Similarly, CNV length distributions of these programs also vary greatly from one program to another. Moreover, some of these programs are not optimized to detect short CNVs (CNVs containing 1-4 exons) and show high false positive CNV counts [178-181]. Additionally, high stringent CNV calling programs may have high false negative CNV counts. While these issues are not yet resolved, the clinical importance of exome-based CNV prediction was repeatedly highlighted in recent years [124, 178, 179, 182].

2.3 CNV annotation

Annotation programs provide information needed to understand the functional effects and clinical relevance of predicted variants. ANNOVAR [120] and VEP [121] are the most commonly used programs that can annotate CNVs and other genomic variants. Additionally,
CNVAnnotator [183], SG-ADVISER CNV [184] and DeAnnCNV [185] provide web-portals for CNV annotation. These programs use different source datasets and provide a broad range of information including gene and functional effect information, information on already known and common CNVs, and clinically relevant information. These annotations are effectively used to identify disease-causing CNVs in the next stage. However, since the choice of source dataset has a major effect on the variant interpretation [186], efficiency in disease-causing CNV identification may depend on the selected annotation program.

2.4 Causal CNV detection

Following prediction and annotation, CNVs are filtered and prioritized (using approaches discussed in section 1.5.1) to detect potentially disease-causing variants. Then, the pathogenicity of each of the prioritized variant is assessed to identify disease-causing CNVs that could explain patient phenotypes.

As discussed in section 2.2, there are a number of limitations to the use of exome-based CNV prediction programs. These limitations have a direct effect on disease-causing CNV identification. For example, a high false positive count increases the risk of selecting a false positive CNV as the disease-causing variant. Similarly, high false negative counts (predicting few CNVs) increase the risk of not detecting the causal variant. Additionally, if thousands of CNVs (eg. CONTRA and ExomeCopy) are predicted, this greatly increases the time and effort involved in disease-causing CNV identification.

Following the identification of disease-causing CNVs, current diagnostic approaches implement a secondary experimental step to validate computationally predicted causal variants.

2.5 Validation of disease-causing CNVs

As discussed in section 2.2, CNV prediction programs that use exome sequence data have shown high false positive CNV counts. Thus, in order to confirm the molecular diagnosis, computationally detected disease-causing variants can be validated using various experimental techniques. For example, aCGH can be used to confirm CNV predictions and
when these platforms do not contain probes covering the causal variant (eg. CNVs in pseudogenes), targeted techniques (MLPA, QF-PCR) can be used.

The majority of commercially available aCGH platforms have low probe counts over exonic regions. Therefore these aCGH platforms have limited applicability in detecting short exonic CNV (with at least 3 consecutive probes). For example, the median probe spacing of Agilent high-resolution 1x1M array is 2.1kb and only 5.3% of the array is composed of probes covering exonic regions. Hence, scientists were motivated in developing custom solutions to detect short exonic CNV. Recent studies have demonstrated the utility of these platforms in confirming computationally predicted disease-causing CNVs [178, 179].

In summary, disease-causing CNVs can be detected from exome sequence data in four main steps: CNV prediction, annotation, causal variant detection and validation. While the clinical utility of these approaches has been demonstrated in recent years, a number of challenges that are associated with these steps limits their widespread applicability. Therefore the main objective of this thesis is to overcome these challenges and develop an improved method to detect disease-causing CNVs from exome sequence data.
3 Aims of the study

The main objective of the work presented in this thesis was to develop a complete pipeline to improve the detection of disease-causing CNVs in exome sequence data. Specific aims were to:

1. Evaluate the performance of commonly-used CNV prediction programs for exome sequence data.

2. Study the ability of CNV prediction programs to predict short CNVs (containing 1-4 exons) and to develop new methods that improve short CNV detection.

3. Develop a custom aCGH to experimentally validate computational predictions and to improve short CNV detection compared to existing arrays.

4. Study CNV prediction quality and to develop methods that can assess CNV prediction quality scores.

5. Improve CNV annotation using multiple source databases.
4 Methodological considerations

4.1 Exome sequence analysis pipeline

Computational programs tested in our study use exome sequence data as the source for CNV prediction. DNA samples for exome sequencing were obtained from primary immunodeficiency diseases (PIDD) patients referred to Oslo University Hospital, Oslo Norway. Exome capture was performed using the Agilent SureSelect Human All Exome capture kit v.5 and Illumina rapidcapture exome. Then captured libraries were sequenced on an Illumina HiSeq 2500 at the Norwegian Sequencing Centre (www.sequencing.uio.no).

Sequence data (fastq files) were analysed using an in-house developed exome pipeline (Fig. 5) combining NovoAlign (v2.07.17) [187], Picard (v1.74) [149], GATK (v2.4 and v2.8) [117, 188, 189] and ANNOVAR (2013Aug23)[120]. Alignment (BAM) files generated from the pipeline showed an average coverage ranging from 124x-210x (Illumina rapidcapture exome: 124x-146x and Agilent SureSelect v5: 160x-210x) and they were used as the input for CNV prediction programs tested in our study (section 4.2).

As shown in Fig. 5, our pipeline uses an in-house developed database and scripts to improve variant interpretation. For example, ANNOVAR annotated SNV and indel calls were further annotated to provide population-specific information (using an in-house variant database) and disease-relevant information (using in-house software). During disease-relevant annotation, SNVs and indels affecting primary immunodeficiency diseases (PIDD) candidate genes were identified and then each variant was annotated with the phenotype and inheritance pattern. Finally, these variants were filtered and prioritized using Filtus [190] when identifying PIDD-causing SNVs and indels.
Fig. 5 Exome sequence analysis pipeline
4.2 CNV prediction from exome sequence data

Six CNV prediction programs (ExomeCNV [160], CONTRA [161], ExomeCopy [162], ExomeDepth [163], CoNIFER [157] and XHMM [156]) were tested on alignment files, which were generated from the pipeline discussed in section 4.1. Additionally, a custom algorithm - ExCopyDepth was developed using modules from ExomeCopy (coverage data extraction module) and ExomeDepth (CNV calling module). Then, in order to validate CNVs predicted from these programs, a custom aCGH (exaCGH) focusing exonic regions was developed using Agilent eArray system [191]. Paper 1 and various sections of this thesis provide detailed descriptions of steps involved in CNV prediction, exaCGH development and CNV validation.

4.3 Resolution of microarray platforms

Since their introduction, microarray platforms have been continuously improved to increase the coverage and resolution of genome-wide CNV detection [192]. Today, high-resolution microarray platforms offer genome-wide CNV profiling with high accuracy [98, 193, 194]. However, due to the low probe coverage over exonic regions, these platforms show restrictions in short exonic CNV detection. Thus, we developed a custom CGH array (exaCGH) to perform high-resolution CNV detection in exonic regions [179]. exaCGH contain 1 million probes and cover ~80% of exons (Havana or Ensemble exons in Gencode v.15) with at least 4 probes. Exonic coverage of exaCGH increases up to ~90%, when considering Gencode v.15 exons that are longer than 100bp. The proportion of Gencode v.15 exons covered by 4 or more probes is highest in exaCGH compared to Agilent 1x1M and CytoScan HD array platforms [179].

The resolution of a microarray platform is mainly defined by the mean spacing of array elements [100]. However, the mean probe spacing of an array does not represent variation in local resolution, especially when array elements are not uniformly distributed throughout the genome [100]. The characterization of local variation in probe spacing can provide more insight into the resolution of an array and subsequently better understanding of its utility. Hence, we compared the local probe spacing in exaCGH, Agilent 1x1M and CytoScan HD to study the resolution of these three array platforms.
During this comparison, we first calculated the probe count of each Gencode exon and CDS region (including 800bp flanking regions in 5’ and 3’ ends [179]). In order to compare the variation of probe spacing in regions (Gencode exonic and CDS) that differ in length, these regions were grouped into three categories: 1-2kb, 2-3kb and over 3kb. Then, we calculated the mean probe spacing in each group (Fig. 6).

As expected, exaCGH had the lowest probe spacing in each group ranging ~300bp to ~350bp. Probe spacing in CytoScan HD was ranging from ~400bp to 1.2kb and Agilent 1x1M was ranging from ~1kb to ~2.4kb. This analysis confirmed that exaCGH has the highest resolution compared to two of the tested commercially available high-resolution arrays and optimized to detect short exonic CNVs. As a secondary verification of our array results, several exaCGH CNV calls were randomly selected and manually visualized using Agilent genome workbench.

Fig. 6 Resolution of microarray platforms

4.4 Considerations when identifying true positive, false positive and false negative CNVs

Following the evaluation of exaCGH design, array experiments were performed and exaCGH results were used to validate computationally predicted CNVs. Here, computational
predictions were compared to exaCGH results and CNVs detected by both methods were identified as true positives. False positives CNVs were the computational predictions, which were not detected in exaCGH. False negatives were the CNVs identified by the array but not by the prediction program.

Prediction programs and exaCGH use different CNV breakpoint detection strategies. For example, CNV breakpoint estimation in computational programs depends on the segmentation algorithm used [147], while CNV breakpoints in microarrays were mainly determined the probe density distribution [195]. Therefore computationally predicted CNV breakpoints may differ from breakpoints identified from exaCGH results.

Due to the CNV breakpoint inconsistencies in computational predictions and exaCGH results, a certain degree of overlap between these two datasets needs to be considered when identifying true positives. Additionally, CNVs predicted from different computational programs, have shown very different length distributions [179]. Therefore, the degree of overlap between CNV predictions and exaCGH results may differ from program to program (Fig. 7).

As shown in Fig. 7, computational predictions and exaCGH results overlap to a variable degree depending on the length of CNVs. Therefore when identifying true positive and false positive CNVs, it is challenging to determine the required threshold overlap. Hence, we used 1 base-pair overlap as the threshold to be less stringent in true positive identification.

Using 1 base-pair overlap between computational predictions and exaCGH results can improve the true positive yield while minimizing the false positive CNV count. However, selecting CNVs with very low overlap can potentially classify false positives as exaCGH confirmed variants and thereby introduce bias into our true positive set. Therefore, true positive CNVs were further analysed to study how true positive count (in each program) change with the degree of overlap (between exaCGH results and computational predictions).

As shown in the cumulative frequency distributions (Fig. 8), only a small fraction of CNVs (~0.01%-0.02%) had less than 100bp overlap between the two datasets used. Further decrease in overlap (<100bp) clearly reduces the cumulative frequency, highlighting the presence of few true positive CNVs. This analysis confirmed that the threshold overlap used in CNV
validation stage (1bp overlap), did not inflate the true positive count and thus, did not introduce a significant bias into our study.

Fig. 7 Manual examination of CNVs

a. Computationally predicted CNVs and the exaCGH detection overlap to a variable degree;

b. Computational CNV predictions are internal to the exaCGH detection;

c. The exaCGH detection is internal to computationally predicted CNVs. * exaCGH parameters: minimum number of probes – 3, minimum average absolute log ratio for deletion and duplication - 0.25,
Fig. 8 Cumulative frequency distribution for true positive CNVs
5 Summary of papers

5.1 Paper 1 Identification of copy number variants from exome sequence data

In this study, we evaluated six commonly used CNV prediction programs (ExomeCNV, CONTRA, ExomeCopy, ExomeDepth, CoNIFER, XHMM) and also developed an in-house modification to ExomeCopy and ExomeDepth (ExCopyDepth) to optimize short CNV (containing 1-4 exons) prediction. Additionally, we studied how multiple prediction program combinations can be used effectively to detect CNVs. Crucially, we developed a custom aCGH (exaCGH) focussing on exonic regions to validate computational CNV predictions. In this analysis, CNV prediction was performed on 30 exomes from the 1000 genomes project and 9 exomes from patients with primary immunodeficiency diseases (PIDD). Nine samples from each group were run in exaCGH and both computational predictions and array results were used to identify true positive (TP) and false positive (FP) CNVs.

During the analysis, we compared CNV programs in terms of CNV prediction length, count and the number of exons internal to predicted variants. Then we calculated the sensitivity and false positive rate of these programs.

The comparison of computational programs showed a striking variation in the length and number of predicted CNVs (Paper 1: Fig. 1a, b). These programs also showed a clear variation in the number of exons internal to predicted CNVs (Paper 1: Fig. 2a, b). Compared to individual programs, a narrow range was shown for the number exons internal to CNVs that were predicted from multiple programs (Paper 1: Fig. 2c, d).

The number of TP CNVs identified from different programs also showed clear variation (Paper 1: Table 2). All the programs except CoNIFER and ExomeDepth had low TP/FP count ratios indicating high FP CNV counts (Paper 1: Table 2). Comparison of short CNVs (CNVs containing 1-4 exons) indicated improved performance for ExomeCopy, ExCopyDepth and intersection of ExomeCopy and ExCopyDepth compared to CoNIFER and XHMM (Paper 1: Fig. 3).
1000 genomes exomes used in the study had higher coverage variance compared to the PIDD patient exomes. Additionally 1000 genomes exomes were analysed using a relatively small reference sample collection compared to PIDD patient exomes [179]. When analysing our results, it was possible to observe how these factors (differences in two exome collections) affect the CNV prediction. For example, CONIFER, XHMM, ExomeDepth were not able to predict short TP CNVs from 1000 genomes exomes. However these programs predicted 4-50 short TP CNVs with relatively high TP/FP ratios (0.17-0.67) in PIDD patient exomes. With 1000 genomes exomes, intersection of ExomeCopy and ExCopyDepth showed an increased performance (highest TP/FP ratio) compared to other programs (Paper 1: Table 3).

We then studied the consistency in assigning copy number state (CNS) when a CNV is predicted from different programs. Here we showed that the same CNSs were assigned to CNVs predicted from all the programs or programs that use similar variant calling methods (Paper 1: Table 4).

In order to benchmark the accuracy of CNV prediction, we compared the sensitivity and false positive rate (FPR) of programs tested in our study. ExomeCopy had the highest FPR (38.9% of the samples showed FPR > 0.95) and CoNIFER had the lowest FPR. ExomeDepth, CoNIFER and XHMM were optimized to detect rare CNVs. Therefore these programs had high false negative counts affecting the sensitivity. For example, sensitivity of ExomeDepth, CoNIFER and XHMM was lower than the other tested programs.

Next, the clinical utility of these programs was tested by searching for PIDD-causing variants. All the programs except XHMM identified a PIDD-causing deletion in FANCA (exon 26–37), demonstrating the potential clinical importance of CNV prediction programs that use exome sequence data.

Finally, we proposed a protocol with two stages, especially applicable in candidate gene studies. Stage one: selecting the program or program combination to predict CNVs depending on the target exome collection and user requirements (Paper 1 provides a detailed discussion on conditions need to be considered at this stage). Stage two: validation of computational predictions using exaCGH experiments.
In conclusion, we introduced a customized algorithm to predict short CNVs (ExCopyDepth) and a custom aCGH (exaCGH) focussing exonic regions. Using computational programs and exaCGH, we highlighted that programs are capable of predicting disease-causing CNVs despite of the relatively high false positive counts. Finally, we presented a protocol that guides the CNV prediction and validation especially in candidate gene studies.

5.2 Paper 2 cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data

This paper describes a new software tool (cnvScan) that considerably improves the clinical utility of CNV prediction from exome sequencing data. cnvScan can read input from any prediction program, which may be affected by issues discussed in section 2.2. Amongst these issues, high false positive CNV count is one of the main challenges in clinically relevant CNV identification [178-180]. Therefore in Paper II, cnvScan was introduced to address this challenge.

Firstly, it was demonstrated that the default CNV prediction quality (assigned by prediction programs) was useful as a stable measure to detect likely false positive predictions (Paper 2: Fig. 1, Additional file 1: Fig. S2-S5). Then, a variant screening method was introduced to calculate an additional CNV quality score. The new quality score (CNVQ) was designed to help assess the effects of sequencing artefacts that influence FP CNV prediction.

Sequencing artefacts (eg. high coverage variance across exomes) can affect the coverage normalization step in prediction programs and contribute to false positive CNVs. Although it is not possible to detect and measure sequencing artefacts in normalized data, effects could be observed using an in-house CNV database. For example, a considerable proportion of overlapping false positives, which were predicted from multiple exomes were observed when considering CNVs in an in-house database (Paper 2: Fig. 1, Additional file 1: Fig. S6). As reasoned in Paper 2, these overlapping false positive predictions are likely the effects of sequencing artefacts. Further, we showed that the common TP CNVs also have low quality scores when prediction programs are optimized to detect rare variants. (Paper 2: Fig. 2).
Following the development of a variant screening method, cnvScan was extended with two additional modules: CNV annotation and filtration. CNV annotation module provides different types of functionally and clinically relevant information, and annotation on known CNVs using a broad range of source datasets (Paper 2: Table 1). The filtration module is capable of excluding low quality known non-disease causing CNVs using cnvScan annotations. Clinical utility of these methods were demonstrated in Paper 2 by implementing cnvScan in a PIDD patient cohort (n=64) and detecting three patients with PIDD-causing CNVs (Paper 2: Fig. 4).

Overall, cnvScan can read input from any prediction program and perform a quality assessment on predicted CNVs. Then cnvScan provides a broad range of annotations using multiple source datasets. Finally, cnvScan filtration prioritizes potentially clinically relevant CNVs by excluding false positive, known and non-disease causing variants. Our method considerably improves the efficiency of disease-causing CNV detection by reducing the time and effort in variant analysis.

5.3 **Paper 3** Primary immunodeficiency diseases - genomic approaches delineate heterogeneous Mendelian disorders

In Paper 3, the utility of exome sequencing as a diagnostic test was evaluated using a group of patients with primary immunodeficiency diseases (PIDD). In this study, exome sequencing was performed on a large cohort (n=278) comprising samples collected from two centres: Oslo university hospital, Oslo, Norway (78 probands) and Baylor College of Medicine, Texas, US (200 probands). All the patients in this cohort had been evaluated through conventional methods (section 1.4.1, 1.4.2) and lacked an established molecular diagnosis when included in the study. Therefore, Paper 3 describes the implementation of a broad range of exome-sequence-based diagnostic methods to assess all known PIDD relevant genes in these patients.

All the Oslo samples were run on the exome sequence analysis pipeline discussed in this thesis (section 4.1). The pipeline run on Baylor samples [196] used Atlas2 suite for SNV and indel identification [197, 198]. Variants identified in both centres were filtered using same criterion during the prioritization, and potentially PIDD-causing variants were further studied
to detect causal variants. Identified PIDD-causing SNVs and indels were confirmed by Sanger sequencing.

The pipeline discussed in section 6.4.2 was used to analyse CNVs from all the Oslo samples and 80 samples from Baylor. While in Baylor, ExCopyDepth and an in-house developed program (HMZDelFinder) were used to predict and annotate CNVs. When disease-causing CNVs were identified, they were confirmed by a custom array (exaCGH or Baylor custom array) or MLPA.

Seven PIDD-causing CNVs were identified with the computational methods used in both centres. The pipeline discussed in section 6.4.2 was able to detect five PIDD-causing CNVs (FANCA, NCF1, TERC, MAGT1, IKZF1) and three were detected in the Baylor pipeline (PGM3, SMARCAL1, MAGT1). Exome samples containing PGM3 and SMARCAL1 CNVs were not tested in Oslo. Thus, the ability to detect these two CNVs was not tested in the pipeline discussed in section 6.4.2.

Our cohort study, revealed the potential molecular basis for the disease in 40% (n=110) of the probands. Clinical diagnosis was revised in about half (60/110) and management was directly altered in nearly a quarter (26/110) of probands based on the molecular findings. Further, we found evidence for reduced penetrance (in autosomal dominant traits), founder mutations in outbred populations (in autosomal recessive traits), mutation burden effect and, somatic and revertant mosaicism contributing to disease variability.

This study resulted in the expansion of the phenotypic spectrum associated with known disease genes, and the identification of several new PIDD-causing genes. The majority of computationally predicted disease-causing CNVs identified in this study, were found by the CNV pipeline discussed in this thesis.
6 Discussion

This thesis presents a complete pipeline providing a systematic approach to identify clinically relevant CNVs. Paper 1 provides a detailed study on commonly used exome-based CNV prediction programs while introducing a customized algorithm (ExCopyDepth) and a custom aCGH (exaCGH). Both ExCopyDepth and exaCGH were developed to predict and experimentally validate short exonic CNVs. Paper 2 describes a new software package (cnvScan) that improves the clinical utility of CNV prediction programs. Finally, in Paper 3, the clinical importance of ExCopyDepth, exaCGH and cnvScan were confirmed by detecting disease-causing CNVs in a large patient cohort.

6.1 Performance comparison of CNV prediction programs

In Paper 1, we compared the performance of six commonly used CNV prediction programs. We showed that these programs are capable of predicting CNVs with high sensitivity, which was further highlighted by predicting disease-causing CNVs in PIDD patients. Despite these advantages, we demonstrated that exome-based CNV prediction programs have high false positive counts. These findings were confirmed by increasing number of recent benchmark studies [178, 180, 181].

Although the conclusions of our performance comparison were consistent with recent publications, it is possible to identify several key features that are not addressed in Paper 1. For example, 1000 genomes exomes used in our study showed a high coverage variance (mean coverage: 49.3x-199.7x), which could affect the CNV prediction efficiency. Moreover, during the benchmark study, we had not considered the effect of CNV prediction quality scores. Additionally, we had not provided precision and recall of CNV prediction programs, although they are considered as the most relevant descriptors for performance, especially when true negative counts are not available [199].

In order to address these issues, an additional analysis was designed (ExCopyDepth prediction and exaCGH experiments) using 15 PIDD patient exomes with mean coverage
ranging from 160x-210x. Then a set of true positive and false positive CNVs was identified to calculate precision and recall of CNV prediction programs. Since CNV counts vary depending on the thresholds used in exaCGH experiments, array results were analysed in two stringency levels when calculating precision and recall scores (Fig. 9).

**Fig. 9** Performance comparison of CNV prediction programs

a, b, c. Precision and recall were calculated with low stringency exaCGH calls; d, e, f. Precision and recall were calculated with high stringency exaCGH calls; a, d. Precision vs default CNV quality score; b, e. Recall vs default CNV quality score; c, f Precision vs recall of prediction programs; *CoNIFER does not provide CNV prediction quality by default, thus it was not considered in this analysis; * Low stringency parameters: AMD-2 score threshold - 5, minimum number of probes - 3, minimum average absolute log ratio - 0.20; * High stringency parameters: AMD-2 score threshold - 6, minimum number of probes - 4, minimum average absolute log ratio - 0.25

Due to the high false positive counts in CNV prediction programs [179] low precision (<0.5) was shown for low CNV scores (<20; Fig. 9a, d). However, precision increased with increasing CNV quality scores for all 4 CNV prediction programs. The majority of CNV prediction programs are optimized to detect rare variants. Thus, these programs have high false negative counts, as they are not capable of detecting all the exome-wide CNVs identified in exaCGH experiments. Consequently, CNV programs had low recall scores (Fig.
9b, e). However, with increasing exaCGH stringency (Fig. 9e), recall of CNV prediction programs increased drastically (0-0.06 to 0-0.13). This indicated that CNV prediction programs have high recall (sensitivity) in detecting high-stringency CNV calls.

When considering both precision and recall (Fig. 9c, f), XHMM showed a lower performance compared to ExomeCopy, ExCopyDepth and ExomeDepth. With high-stringency CNV calls, ExomeCopy, ExCopyDepth and ExomeDepth showed a similar performance (Fig. 9f). CoNIFER does not provide quality scores in its default setting, therefore it was not included in the comparison presented in Fig. 9. However, based on recently published literature, CoNIFER can be considered as a program with very low false positive counts [178-180].

Precision and recall plots presented in this thesis and other similar benchmark studies suggest that CNV prediction programs are subject to high false positive counts even though they are capable of predicting disease-causing CNVs [178-181]. This highlights the importance of careful evaluation and additional validation of computational CNV predictions.

6.2 Effect of cnvScan in-house database in CNV analysis

Current exome sequence analysis pipelines use in-house SNV databases to detect sequencing artifacts in SNV calls [200, 201]. Similarly, cnvScan uses an in-house database and provides a CNV quality assessment strategy to improve the detection of sequencing and coverage artefacts. Here, we recommend using the same exome collection for CNV prediction and cnvScan in-house database creation [202]. Alternatively, when prediction programs are run in several batches of exome samples, CNVs from all these exomes can be pooled together to create a relatively large in-house database. Changes in the content and size of the in-house database affect quality assessment, filtration and thereby CNV interpretation. However in Paper 2, we have not discussed how the content and size of cnvScan in-house database affect the CNV analysis.

In order to study this effect, two in-house databases were created from two different datasets: CNV predictions from 18 PIDD patient exomes (db-18) and CNV predictions from two batches containing 18 and 47 PIDD exomes (db-65). Then, cnvScan was implemented using these two in-house databases to detect PIDD-causing CNVs in exomes represented in db-18.
Two PIDD-causing CNVs (two patients from the same family with a deletion in \textit{MAGTI} and one patient with a deletion in \textit{NCFI}) were identified when db-18 was used for the cnvScan implementation. However, only \textit{MAGTI} CNV was detected with the db-65 (\textit{NCFI} variant was not detected). This analysis showed that the in-house database creation approach affects the disease-causing CNV identification.

IGV visualization of \textit{NCFI} variant showed the presence of low quality overlapping database CNVs in db-65 (Fig. 10). Therefore in the db-65 implementation, a low cnvScan quality score (CNVQ) was calculated and \textit{NCFI} variant was filtered out. However, low quality overlapping database CNVs were not observed in db-18 (Fig. 10). Thus, the prediction quality of the \textit{NCFI} CNV was not affected and it was not filtered out during cnvScan db-18 implementation.

![Fig. 10 Effect of in-house database creation approach](image)

**Fig. 10** Effect of in-house database creation approach

\textbf{a.} db-18 implementation (no low quality overlapping database CNVs in db-18); \textbf{b.} db-65 implementation (low quality overlapping database CNVs were observed in db-65) * Bands with higher and lower colour intensities in \textbf{a} and \textbf{b} represent CNV predictions with higher and lower quality scores.

As mentioned earlier, CNVs in db-65 were predicted in two different batches with 18 and 47 exomes in each collection. CNV prediction quality of these two batches could differ from
each other due to various technical artifacts. Pooling CNVs in these two batches will introduce batch effect to the cnvScan db-65 database and affect the CNV quality assessment and filtration. For example, Table 5 shows that the numbers of CNVs identified in cnvScan db-65 (filtration thresholds:15-65) are slightly lower than that of db-18. This highlights the need for further optimization techniques that can minimize the batch effect. For example, SNV prediction approaches implement quality score recalibration strategies to minimize various types of biases [117, 189]. The current version of cnvScan avoids the batch effect by using the same exome collection for CNV prediction and cnvScan implementation.

<table>
<thead>
<tr>
<th>cnvScan filtration threshold</th>
<th>Number of CNVs identified in cnvScan db-18 in-house database</th>
<th>Number of CNVs identified in cnvScan db-65 in-house database</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>788</td>
<td>791</td>
</tr>
<tr>
<td>15</td>
<td>357</td>
<td>344</td>
</tr>
<tr>
<td>20</td>
<td>220</td>
<td>205</td>
</tr>
<tr>
<td>25</td>
<td>159</td>
<td>151</td>
</tr>
<tr>
<td>30</td>
<td>128</td>
<td>113</td>
</tr>
<tr>
<td>35</td>
<td>86</td>
<td>83</td>
</tr>
<tr>
<td>40</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>45</td>
<td>65</td>
<td>55</td>
</tr>
<tr>
<td>50</td>
<td>56</td>
<td>46</td>
</tr>
<tr>
<td>55</td>
<td>53</td>
<td>37</td>
</tr>
<tr>
<td>60</td>
<td>41</td>
<td>26</td>
</tr>
<tr>
<td>65</td>
<td>33</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 5 CNVs identified in cnvScan implementation

### 6.3 Characterization of true positives excluded in cnvScan implementation

Current SNV analysis pipelines implement different filtration criteria to exclude technical artifacts and population specific variants when searching for clinically relevant variant in exome sequence data. For instance, variant calling programs (samtools [148] and GATK [117, 189]), variant analysis packages (Filtus [190]) and annotation programs (VEP [121]) provide a number of options to apply various filters on SNVs and indels to find important or interesting results.
Similarly, cnvScan and the recently published CoNVaDING [172] provide filtration strategies to exclude low quality CNV predictions. cnvScan provides a filtration strategy for CNVs predicted from any coverage-based prediction program. CoNVaDING deviates from cnvScan by combining both CNV prediction and filtration in a single program.

CoNVaDING performs a coverage analysis and selects a control dataset (exomes with most similar coverage patterns) for each target exome when calling CNVs. During this coverage analysis, a set of quality matrices is created to describe how coverage of exons varies within the target exome and control dataset. Then, these matrices are used in the filtration step to improve the sensitivity and specificity. However, the clinical utility of CoNVaDING has not yet been studied in a large patient cohort. The CNV filtration efficiency of cnvScan is discussed in Paper 2 and the applicability of cnvScan in a large patient cohort is discussed in Paper 3.

While cnvScan was effectively used to identify disease-causing CNVs, our tool can be bias-prone due to the implemented hard-filtering step. Particularly, rare disease-causing CNVs can be excluded when high cnvScan filtration thresholds are applied. A detailed characterization of excluded CNVs would assist cnvScan users to select the optimum threshold, which could be used to minimize the hard-filtration bias. Therefore an additional analysis was designed to study the quality scores of filtered-out or excluded CNVs, especially as this was not discussed in Paper 2.

Here, CNV prediction (ExCopyDepth) and cnvScan implementation was performed using 17 PIDD exomes. During cnvScan implementation, relatively high quality score threshold (default CNV quality and CNVQ: 40) was used to increase the number of filter-out CNVs (n=912). Following the filtration, CNVs excluded in cnvScan implementation were selected for further evaluation. In parallel, exaCGH experiments were performed (on the 17 PIDD exomes) to identify true positive and false positive CNVs within excluded variants.

The quality score analysis of excluded variants (Fig. 11) showed that true positives have higher quality scores compared to false positive CNVs (Fig. 11a). These excluded true positives may contain disease-causing rare variants as well as known variants that are reported in public databases. Therefore these CNVs were further examined to identify known
and novel variant within excluded true positives. Here, CNVs reported in public databases (manually curated DGV, 1000 genomes CNVs and sanger CNVs) were identified as known variants, while CNVs not yet reported in public datasets were identified as novel variants. Then we compared the default quality score and CNVQ distributions of these identified CNVs. Both known and novel true positive CNVs were spread over lower and higher ends of default quality score spectrum (Fig. 11b). However, when considering CNVQ distributions, known variants were spread over a wider range compared to novel CNVs (Fig. 11c).

**Fig. 11** Analysis of CNVs excluded in cnvScan

**a.** Default quality scores of false positive (FP) and true positive (TP) CNVs filtered out (excluded) in cnvScan implementation; **b.** Default quality score distribution of known and novel TP CNVs excluded in cnvScan; **c.** CNVQ distribution of known and novel TP CNVs excluded in cnvScan.

The number of novel true positive CNVs excluded in cnvScan was 32 (~3.5% of the total CNVs excluded) and they were found in at least 4 patient samples. These CNVs could either be population-specific common variants that are not yet reported in public databases or disease-causing CNVs that are present in a subset of patients (n≥4).

Previous studies have demonstrated that CNVs present in <10% of the population can cause or be associated with a broad range of diseases [203-205]. Thus, it is important not to exclude these variants during filtration. Based on Fig. 11, low cnvScan thresholds can be used to retain the majority of these novel CNVs. However, there is a risk of excluding a small fraction of novel CNVs. Therefore, we updated the cnvScan filtration script to produce two
results files: a primary result file containing high quality CNVs identified in cnvScan and a secondary file containing low quality excluded novel variants. Hence, cnvScan users can examine the primary result file when searching for disease-causing CNVs. If a causal variant was not found, users can search the secondary results file.

6.4 Comparison with other disease-causing CNV detection methods

6.4.1 Advantages of cnvScan annotation compared to other annotation programs

Today, a number of programs are available for scientists to annotate CNV predictions; for example, VEP [121], ANNOVAR [120], CNVannotator [183], DeAnnCNV [185]. Despite the availability of these tools, cnvScan was extended to provide additional information on CNVs, especially using sources that are not available in other methods. For example, clinically relevant information from DECIPHER DDD [128] and, known and common CNVs from high quality manually curated database of genomic variants (DGV) [206] were not provided in the majority of other CNV annotation programs. Similarly, these programs provide a limited set of information on novel CNVs when pre-computed scores such as haploinsufficiency scores [66] and gene intolerance scores [207] are useful in functional interpretation of novel CNVs.

A recent study has shown that variant interpretation can be affected by the source transcript dataset used in annotation programs [186]. For example, CNVAnnotator uses a combination of Refseq, UCSC and ENCODE data to annotate genes and other genomic features [183], which could increase the risk of misinterpreting annotated CNVs. cnvScan avoids this issue by using Gencode dataset and provides high quality, consistent and accurate CNV annotations.
6.4.2 Rate of disease-causing CNV detection

Paper 3 discussed in the thesis uses a PIDD patient cohort and describes the utility of our method in identifying disease-causing CNVs. Patient exomes tested in our method were collected from the Department of Medical Genetics, Oslo University Hospital (n=78) and Baylor College of Medicine, Texas (n=80).

ExCopyDepth and cnvScan were able to identify five disease-causing CNVs from these patient populations: four PIDD-causing CNVs (FANCA, MAGT1, NCF1, TREC) in five patients from the Oslo cohort (positive detection rate: 6.4%) and two PIDD-causing CNVs (MAGT1, IKZF1) in the Baylor patient cohort (positive detection rate: 3.75%). All these CNVs were validated by either exaCGH or MLPA experiments.

Our cohort study reported higher positive detection rates (Oslo: 6.4%, Baylor: 3.75%) compared to other heterogeneous Mendelian cohorts and recently published large clinical studies [123-125]. This was mainly due to the carefully designed steps of our complete pipeline: CNV prediction, cnvScan implementation, pathogenicity assessment and exaCGH validation (Fig. 12).
**Fig. 12** Disease-causing CNV detection pipeline

**a.** Computational steps: CNV prediction (using alignment files and exome definition files as main inputs) and cnvScan implementation (using cnvScan quality thresholds, in-house database and various annotation datasets); **b.** Manual step: pathogenicity assessment of each prioritized variant using cnvScan annotations and other relevant resources; **c.** Experimental step: if disease-causing CNVs are identified in the previous step, they are confirmed using exaCGH experiments.

During the implementation of the pipeline (Fig. 12), a low stringency prediction program (ExCopyDepth) was used to minimize the false negative count when calling CNVs. Then during cnvScan implementation, predicted variants were filtered and prioritized to reduce false positive CNV count. Following the prioritization, cnvScan annotations and other relevant information were used for effective CNV interpretation and pathogenicity assessment. If disease-causing CNVs were identified, exaCGH experiments were performed to confirm computational predictions.

When comparing positive detection rates of Oslo and Baylor samples, the Oslo group had higher detection rates than the Baylor group. This could be due to the batch-wise
implementation of the Oslo samples. For example, ExCopyDepth prediction and cnvScan implementation was performed on 20-30 exomes sequenced in 1-3 consecutive batches. However, all the Baylor exomes \((n=80)\) were pooled together, potentially increasing the coverage variance across the collection and thereby affecting the CNV prediction efficiency. Based on these observations, we suggest that ExCopyDepth and ExomeDepth CNV prediction can be further optimized by pooling exomes depending on their sequencing batches. However, when using XHMM and CoNIFER, CNV prediction efficiency can increase with large exome collections [156, 157, 208]

### 6.4.3 Short CNV prediction

As shown in Table 6, our CNV pipeline detected PIDD-causing variants with a broad length spectrum. For example, the \(FANCA\) CNV contained 11 exons and only one exon was internal to the \(TERC\) variant. Moreover, 40% of the PIDD-causing CNVs detected in our study contained 1-2 exons. Therefore, we have demonstrated the ability to detect disease-causing short CNVs (containing < 4 exons), when the majority of CNV prediction programs are not optimized to predict short CNVs [156, 157, 166, 208].

Short CNV detection efficiency in prediction programs depends on various algorithmic features. In ExomeCopy and ExCopyDepth, short CNV prediction was improved due to the implemented tile length normalization step [162, 179].

When considering CoNIFER and XHMM, both programs use principal component analysis based normalization methods to remove experimental and systematic artefacts. CoNIFER CNV calling algorithm search for consecutive runs of three or more targets with normalized scores that are higher than a specified hard threshold [157, 179]. Therefore, CoNIFER restricts its ability to predict CNVs with 1-3 exons. XHMM deviates from CoNIFER by implementing a hidden markov model (HMM) to assess the quality of each CNV signal [156, 179, 208], as opposed to hard thresholds. Thus XHMM shows improved performance in detecting short CNVs compared to CoNIFER.
<table>
<thead>
<tr>
<th>Gene (region)</th>
<th>Number of exons (Length)</th>
<th>CNV</th>
<th>Cohort study</th>
</tr>
</thead>
<tbody>
<tr>
<td>FANCA</td>
<td>10 (22273 bp)</td>
<td>Heterozygous deletion</td>
<td>Oslo</td>
</tr>
<tr>
<td>MAGT1</td>
<td>5 (16253 bp)</td>
<td>Deletion</td>
<td>Oslo and Baylor</td>
</tr>
<tr>
<td>IKZF1</td>
<td>2 (15219 bp)</td>
<td>Heterozygous deletion</td>
<td>Baylor</td>
</tr>
<tr>
<td>TERC</td>
<td>1 (3600 bp)</td>
<td>Heterozygous deletion</td>
<td>Oslo</td>
</tr>
<tr>
<td>NCF1</td>
<td>11 (15350 bp)</td>
<td>Homozygous deletion</td>
<td>Oslo</td>
</tr>
</tbody>
</table>

Table 6 PIDD-causing variants identified by using the pipeline discussed in section 6.4.2

6.4.4 Utility of exaCGH in detecting short CNVs

As discussed in methodological considerations, exaCGH was optimized to detect short exonic CNVs compared to commercial arrays. This directly contributed to the disease-causing short CNV yield of our study. For example, a PIDD-causing single exon deletion in TERC was detected in exaCGH with 7 probes, whereas CytoScan HD and Agilent 1x1M arrays did not contain probes covering the variant. Identification of TERC deletion highlights the advantage of using exaCGH over other commercial arrays. In recent years, custom arrays were used in number of studies to confirm computationally predicted short exonic CNVs with variable degree of success [124, 178, 209].

In conclusion, this thesis discusses a complete pipeline optimized to identify disease-causing CNVs in exome sequence data while solving problems faced by other commonly used methods. Our pipeline is adaptable to user requirements and greatly improves the time and effort in clinically relevant CNV identification [179, 202, 210]. The pipeline was proven to be effective in identifying CNVs with a broad length spectrum and have a positive detection rate of 3.7%-6.4%. Therefore this initiative will provide necessary resources for exome-based CNV detection in current genetic diagnostic protocols and improve the diagnosis of patients with genetic diseases.
7 Future directions

7.1 Optimizations to reduce the batch effect in cnvScan implementation

As discussed in section 6.2, when CNVs were predicted from several batches of exomes, CNV prediction quality could differ from one batch to another (batch effect). Therefore when implementing the current version of cnvScan on multiple batches of exomes, separate in-house databases need to be created and a filtration threshold needs to be selected per batch.

In order to minimize the bias introduced by the batch effect, default CNV quality scores assigned by prediction programs need to be recalibrated using an advance statistical method. Here, special measures needed to be introduced, so that the recalibrated scores do not vary depending on the exome collection used in the prediction program.

With these recalibrated CNV scores, all the in-house CNVs can be pooled to create a relatively large cnvScan database without introducing the batch effect. A larger in-house database would lead to the development of an advance CNVQ (cnvScan quality score) calculation method that could improve the confidence in CNV quality assessment. Both recalibrated quality scores and an advanced cnvScan quality assessment strategy, would help select an optimum cnvScan filtration threshold during variant prioritization.

7.2 Whole-genome CNV analysis

Whole-genome sequencing (WGS) has the potential to capture genome-wide SNVs, indels and CNVs in one experiment. Therefore, WGS will be an attractive alternative to exome sequencing when available at an affordable price [211-214].

Prediction programs tested in our study use exomes as the source to detect CNVs. Therefore these programs will be replaced when WGS become a widespread genetic diagnostic tool. However, cnvScan accept CNV calls from any coverage-based prediction program as the
input [202]. Thus, cnvScan can be extended to analyse coverage-based CNV calls from WGS data. Moreover, paired-end and split-read approaches [215] can be introduced in cnvScan as a secondary CNV evaluation step. For example, when evaluating a CNV, cnvScan can search for reads with insert sizes that significantly deviate from the expected length [216] or partially mapped reads to identify CNV breakpoints [141]. Thus, cnvScan can be extended to assess, further evaluate, annotate and filter CNVs predicted from WGS data.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGH</td>
<td>Comparative genomic hybridization</td>
</tr>
<tr>
<td>CNS</td>
<td>Copy number state</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variant</td>
</tr>
<tr>
<td>CNVQ</td>
<td>cnvScan quality score</td>
</tr>
<tr>
<td>DECIPHER</td>
<td>Database of chromosomal imbalance and phenotype in Humans using ensembl resources</td>
</tr>
<tr>
<td>exaCGH</td>
<td>Exon focussed CGH array</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescence in situ hybridization</td>
</tr>
<tr>
<td>FP</td>
<td>False positives</td>
</tr>
<tr>
<td>FPR</td>
<td>False positive rate</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association studies</td>
</tr>
<tr>
<td>HGMD</td>
<td>Human gene mutation database</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden markov model</td>
</tr>
<tr>
<td>indel</td>
<td>Insertion or deletion</td>
</tr>
<tr>
<td>MLPA</td>
<td>Multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PIDD</td>
<td>Primary immunodeficiency diseases</td>
</tr>
<tr>
<td>QF-PCR</td>
<td>Quantitative fluorescence PCR (polymerase chain reaction)</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads per thousand bases per million reads sequenced</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular value decomposition</td>
</tr>
<tr>
<td>TP</td>
<td>True positive</td>
</tr>
</tbody>
</table>
**Commonly used terms**

**Allele**: Variant forms of the same gene, in the same locus on homologous chromosomes.

**Copy number variant (CNV)**: A DNA sequence that is 1kb or larger and is present at a variable copy number in comparison with a reference genome.

**Gene**: A DNA sequence or locus that encode a functional RNA or protein product and is the molecular unit of heredity.

**Genetic loci**: Specific regions that are mapped within a genome.

**Genome**: Complete set of genetic material present in a cell.

**Genotype**: The genetic constitution of the individual.

**High-throughput sequencing (next-generation sequencing)**: Coverall term used to describe a number of different modern sequencing technologies (eg. Illumina sequencing /Short-read sequencing). These technologies enable DNA and RNA sequencing much more quickly and cheaply than the previously used Sanger sequencing technology.

**Inversion**: A segment of DNA that is reversed in orientation with respect to the rest of the chromosome.

**Karyotype**: An ordered array of metaphase chromosomes from a photomicrograph of a single cell nucleus arranged in pairs in descending order of size and according to the position of the centromere.

**Paired-end reads**: paired reads originating from the same template molecule and separated by a known distance.

**Precision (positive detection rate)**: true positive count / (true positive count + false positive count)
**Recall (sensitivity):** true positive count / (true positive count + false negative count)

**Segmental duplication or low-copy repeat:** A segment of DNA >1kb in size that occurs in two or more copies per haploid genome, with the different copies sharing >90% sequence identity. They are often variable in copy number and can therefore also be CNVs.

**Soft-clipped reads:** Partially mapped reads with bases that are not part of the alignment.

**Structural variant (SV):** Term that encompasses a group of microscopic or submicroscopic genomic alterations involving segments of DNA. Structural variants include CNVs comprising deletions and duplications, translocations and inversions.

**Translocation:** A change in position of a chromosomal segment within a genome. Translocations can be intra- or inter-chromosomal.
References

33. Wellcome Trust Case Control Consortium: Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010, 464(7289):713-720.
38. Exome Sequencing Project (ESP) 6500 exome data [http://evs.gs.washington.edu/EVS/]


149. Picard [http://broadinstitute.github.io/picard/]


189. eArray [https://earray.chem.agilent.com/earray/]

60
192. Gentien D, Reyes C: Microarrays-Based Molecular Profiling to Identify Genomic Alterations. Pan-cancer Integrative Molecular Portrait Towards a New Paradigm in Precision Medicine 2015, Chapter 4: 31-45.


199. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015, 10(3):e0118432.


