Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

DATA SCIENCE AND COMPUTATION

Ciclo XXXIII

Settore Concorsuale: 06/A1 – GENETICA MEDICA

Settore Scientifico Disciplinare: MED/03 – GENETICA MEDICA

A machine learning based method to detect genomic imbalances exploiting X chromosome exome reads

Presentata da: **Paola Dimartino**

Coordinatore Dottorato

Prof. Andrea Cavalli

Supervisore

Prof. Marco Seri

Co-supervisore

Dr. Tommaso Pippucci

Esame finale anno 2022

Index

Glossary

Term	Definition
AUTOSOME/AUTOSOMAL	Any numbered chromosome – sex chromosomes excluded.
BAM FILE	Binary compressed file containing tab-delimited sequence alignment data.
BREAKPOINTS	Junctions of structurally variable genomic segments.
CANONICAL TRANSCRIPT	Most conserved transcript, highly expressed and with the longest coding sequence, chosen to represent a gene.
GC CONTENT	Percentage of G and C nucleotides in a given region or the whole genome. While AT pairs are bound by 2 hydrogen bonds, GC pairs have 3 bonds and are more thermostable. Thus, GC-rich regions are difficult to anneal and amplify during WES enrichment.
COVERAGE/MEAN COVERAGE	Number of unique reads aligned to a reference nucleotide. Mean Coverage is given by the average number of reads aligned to the genome or targeted regions.
GOLD STANDARD/TRUTH SET (genetics)	Set of curated, validated and high-quality SVs called in an individual. It's assumed that all calls reported in this set are true variations.
MAPPABILITY	Measure of a region complexity. High complexity sequences map in unique regions; low complexity sequences can map in other genomic regions.
NEXT GENERATION SEQUENCING	High throughput massively parallel sequencing technology. Used to ascertain the order of nucleotides in whole genomes or specific regions. Apply to both DNA and RNA.
RESOLUTION (SVs)	Minimum SV size that can be reliably detected by the technology used. The higher the resolution, the better the capability of also defining the real breakpoints.
SEGMENTAL DUPLICATIONS	Long DNA sequences (> 1kb) that are repeated in the genome and share >90% sequence identity.
SOFT CLIPPING	Part of the read (start or end) that doesn't perfectly match the reference genome. These

	reads are usually flagged and ignored if not used for SV discovery workflows.
SV DISCOVERY APPROACH: ASSEMBLY (AS)	SV are detected with de novo alignment of the contigs to the reference genome. It also uses unmapped sequencing reads.
SV DISCOVERY APPROACH: READ DEPTH (RD)	The divergence of read depth distribution across the genome is used to detect CNVs.
SV DISCOVERY APPROACH: READ PAIR (RP)	Paired-end reads having inconsistent span and orientation are collected and used to derive information about SVs.
SV DISCOVERY APPROACH: SPLIT READS (SR)	Single- or paired-end soft-clipped reads spanning the breakpoints of an SV are used.

Introduction

Next generation sequencing and the challenge to identify Copy Number Variants in the clinical context

Whole Exome Sequencing as a first-tier genetic test: advantages and limitations

All along the last decade, Next-Generation Sequencing (NGS) technologies has enormously increased our capability to diagnose genetic diseases in clinical laboratories. The key of this success largely resides in Whole-Exome Sequencing (WES), and on its focus on mutations directly altering protein-coding regions (Chong et al., 2015). WES is a type of targeted short-read NGS that enriches only the protein-coding (1-2%) regions of the genome (Balachandran et al., 2020). As approximatively 85% of Mendelian disorders are explained by variants in protein-coding regions, WES is a cost-effective and highly parallelizable technique widely used for detection of small variants (Gordeeva et al., 2021; Gilissen et al., 2012). Moreover, over the last 10 years, a set of alignment, variant-calling and prioritization best practice workflows for Single-Nucleotide Variants (SNVs) and short insertion/deletions (InDels) has been established, making the process reproducible worldwide while also giving the maximum obtainable sensitivity and validation rate (>99%) (Gilissen et al., 2012; Koboldt et al., 2020). The constantly declining costs, accompanied by the automatized and streamlined clinical interpretation of most coding variants enabled by modern bioinformatics pipelines (Lassman et al., 2020) has made WES the strategy of choice in the hands of geneticists facing a suspected genetic disease. WES is thus rapidly becoming a first-tier test for routine care in many countries and for different diseases (Arts et al., 2019; Klau et al., 2021; Mergnac et al., 2021).

However, WES has also limitations that prevent its use to accurately detect the entire range of clinical variant types. While indeed WES is best at identifying small variants, in particular coding SNVs, other clinically relevant classes of variations can be refractory to be captured by WES. One of the most important examples regards the Structural

Variants (SVs), a set of structural and quantitative chromosomal rearrangements widely studied for their increasingly recognized role in human evolution, adaptation, phenotypic variation and disease (Hastings et al., 2009, Le Scouarnec et al., 2012). SVs include classes of variants that differ both in size and type. By size they range from a minimal size of 50 bp to entire chromosomes (Ho et al., 2020). By type, SVs can be sub-grouped in two main classes: balanced and imbalanced rearrangements. Balanced rearrangements include SVs that do not result in a loss or gain of genetical material, such as inversions, reciprocal translocations and copy-number neutral insertions; whereas unbalanced rearrangements, also known as copy number variations (CNVs), includes deletions, duplications and insertions that alter the diploid status of the DNA by changing the copy number of chromosomes or chromosomal regions (Ho et al., 2020, Spielmann et al., 2018, Zarrei et al., 2015).

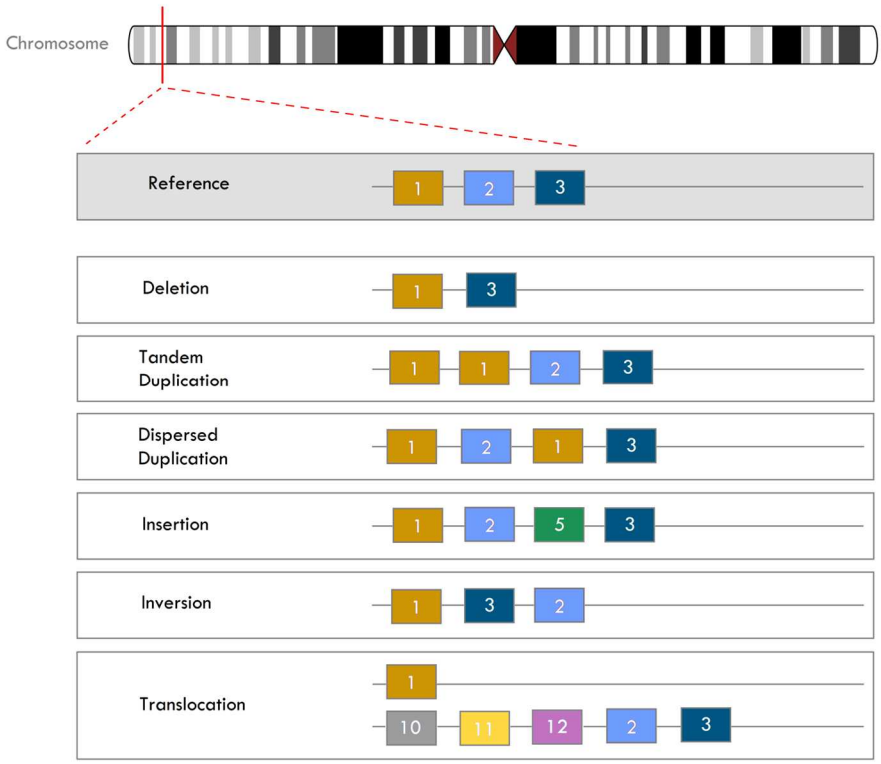


Figure 1 Example of several balanced and unbalanced SV events: the reference panel show 3 genes (boxes 1, 2 and 3) with the correct copy number and position; when this is not true, an SV occurs.

Thus far, there are several possible molecular mechanisms identified that link these rearrangements to clinical phenotypes, of which the most common are 'gene dosage' (Lupski et al., 1992) and 'position effects' (Biederman and Bowen, 1976). While the latter changes gene expression by moving a gene from its normal non-coding *cis*-

regulatory environment (Spielmann et al., 2018), gene dosage is due to the change in the number of copies of a gene or in its regulatory elements (Stankiewicz et al., 2010). More than 17% of protein-coding genes have >90% probability of being intolerant to deletions or duplications changing the wild-type number of copies (Lek et al., 2016; Stankiewicz et al., 2010). In general, genomic rearrangements are at least 1000 to 10.000 folds more frequent than single nucleotide variants – with mutation rates ranging between 10^{-4} and 10^{-5} (Lupski, 2007) and genomic disorders arising with similar frequencies (Stankiewicz et al., 2010).

In spite of their importance as source of clinical variations, most types of SVs are not detectable by WES analysis, as most of these balanced events have breakpoints outside the exonic sequences. Nonetheless, as better explained below, WES can be exploited to detect at least CNVs, the most frequent SV type with renowned clinical consequences. Although possible, however, CNV identification from WES data is not exempt from problems and pitfalls that prevent its implementation as part of the clinical routine (Marchuk et al., 2018). The necessity of accurate CNV detection methods in genetic testing is proven by considering the prevalence of intragenic CNVs in different disease groups, such as neurological disorders (~35%), cancer syndromes (~8.3%) and pediatric and rare disorders (~7.7%) encompassing single exons, several exons or, mostly with duplications, entire genes (Truty et al., 2019; Fig. 2).

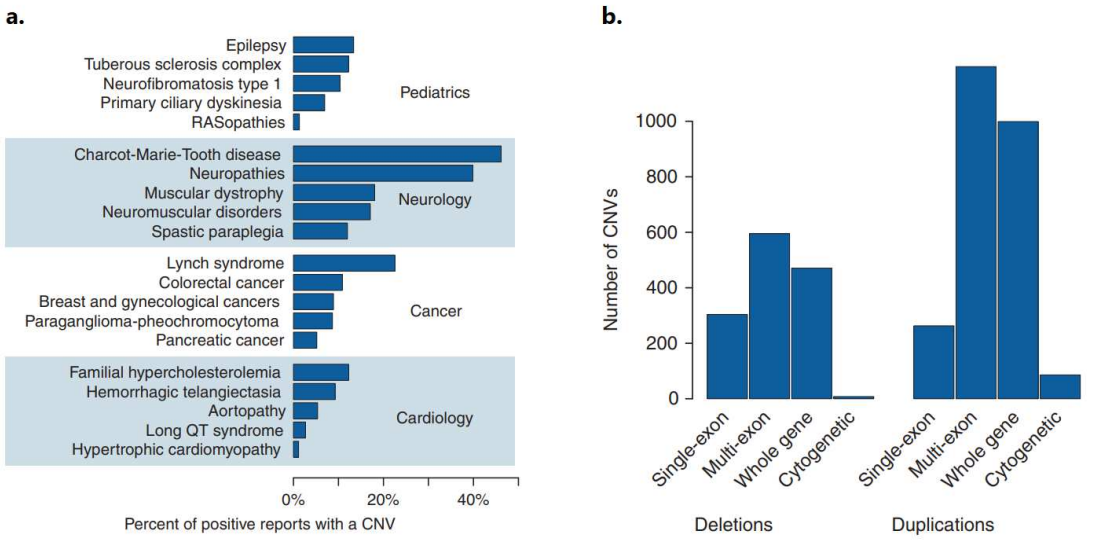


Figure 2 Pathogenic CNVs from a large cohort (from Truty et al., 2019). a. Percentage of positive cases with pathogenic CNVs in different disease groups, showing the highest percentage within the Neurology group. b. Bar plots showing the length distribution of deletions and duplications, from single/several exons (mostly deletions) to entire genes or regions of several Mb (higher for duplications).

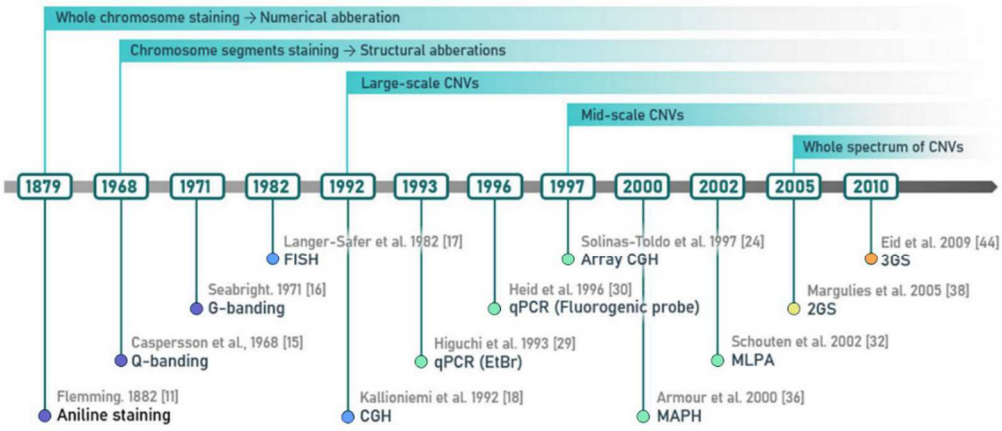
Considering these percentages of prevalence, some of the negative cases in WES could be explained by a deletion or duplication of a dosage-sensitive gene.

Although sequencing whole genomes in place of WES is suggested as a more comprehensive way to detect CNVs – as it is able to supplant WES, microarray and karyotyping - there are still several obstacles that make this a difficult transfer, the most important being the requirement of a powerful computational infrastructure and the overwhelming amount of results produced, that have to be interpreted and reported (Marshall et al., 2020; Stranneheim et al., 2021). Therefore, methods that maximize the use of WES data beyond small variants to CNVs are important in the clinics in order to make WES a comprehensive and efficient first-tier genetic test.

Genome-wide methods for the identification of SVs

The prevalence of SVs in the human genome has been calculated, over the past decades, based on the detection capability and resolution of the available technologies. By improving technologies, the estimated number and size of SVs has been progressively and hugely changed. (Ho et al., 2020). Starting from the early 1970s, the invention of two cytogenetic methods, chromosome banding and karyotyping, made possible the detection of numerical chromosomal aberrations and very large (several Mb) microscopic SVs (Le Scouarnec et al., 2012). Even if these techniques have a very low resolution, they are still widely used as first-tier tests in clinical diagnostics (Speicher et al., 2005). The advent of new molecular cytogenetic approaches, collectively referred as FISH-based techniques, made possible the detection of sub-microscopic SVs, increasing the detection's resolution. New alternative targeted molecular approaches, real-time qPCR and MLPA, simplified the detection of large-scale CNVs, even if the real improvement in SVs detection came with the introduction of array-based techniques. The main advantage of these techniques are the specificity, sensitivity and throughput in comparison with the previous methods, generating data for thousands of genomic regions in a single experiment. However, even if the resolution is high, array-based techniques mostly detect unbalanced rearrangements. With the advent of NGS in 2005, millions of DNA molecules can be sequenced in parallel, generating sequences of base pairs called

reads. Based on read length, NGS is divided in short-read (30-400bp) and long-read NGS (few kilobases). In both approaches, sequence reads are aligned using bioinformatics tools to the reference genome and all differences between the reference and the aligned sequences, ranging from SNVs to SVs, can be detected (Le Scouarnec et al., 2012). Considering all the techniques developed in the last 50 years, NGS technology is superior in terms of sensitivity (it can theoretically detect all types and sizes of rearrangements) and resolution (to the base pair level) reaching with long-read NGS 2.000-8.000 SVs detected per human genome (Ho et al., 2020; Le Scouarnec et al., 2012; Sudmant et al., 2015; Hehir-Kwa et al., 2016). The base-pair resolution implies that the breakpoints of a rearrangement can be easily mapped, making also possible the characterization of complex rearrangements with multiple breakpoints (Le Scouarnec et al., 2012).



Technique	Detectable SV types	Resolution	Key features
Karyotyping	Large CNV, TRA, aneuploidies	> 3 Mbp	Staining
FISH	CNV, TRA, CGR	10 kbp	Fluorescent probes
aCGH	CNV, CGR	5–10 kbp	Hybridization
SNP array	CNV, LOH, CGR	100 bp	SNP probes
Illumina short-read sequencing	Simple SVs, CGR	bp	Paired-end reads
WES	Simple SVs	> 50 bp (generally more than 3 exons)	Protein coding region
PacBio SMRT sequencing	Simple SVs, CGR	bp	CCS and CLR
Oxford nanopore sequencing	Simple SVs, CGR	bp	Ultra-long reads

Figure 3 Methods for SV identification (adapted from Balachandran et al., 2020 and Pös et al., 2021). CNV: copy number variation, TRA: translocation, CGR: complex genomic rearrangement, LOH: loss of heterozygosity, CCS: circular consensus sequencing, CLR: continuous long read.

There are four approaches that can be used, both independently and combined, to detect SVs using short- and long-read NGS: read pairs (RP), read depth (RD), split read (SR) and assembly (AS) (Kosugi et al., 2019).

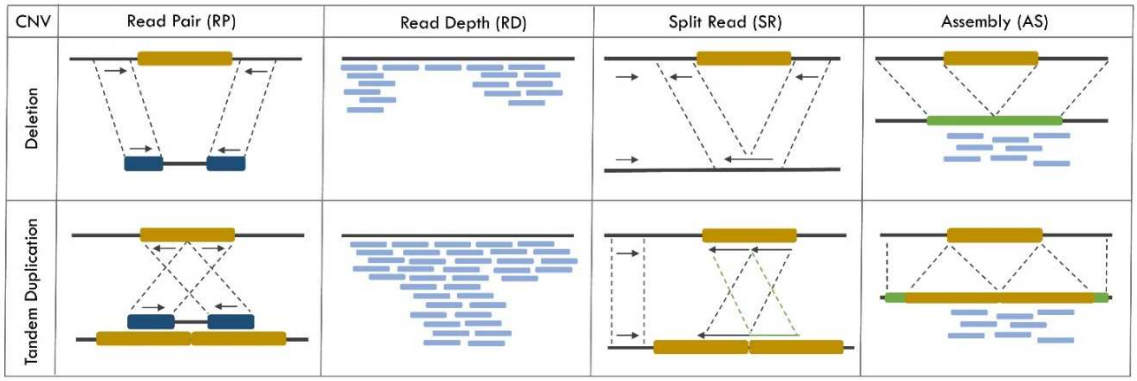


Figure 4 Example of how deletions and tandem duplications are discovered using each NGS approach.

RP and RD collect evidence of an SV based on paired-end reads spanning the entire SV and/or the breakpoints. Since, as already mentioned, exome sequencing targets 1% of non-contiguous genomic regions, the majority of breakpoints will not be part of the experiment; moreover, most target regions cover on average 100-300 bp, making the possibility of spanning an SV more challenging (Fromer et al., 2012, Tan et al., 2014). Even if theoretically these approaches could be used with WES, they would miss most SVs. As for the AS approach, generating reference genomes *de novo* requires reads to be contiguous to obtain a continuous information, automatically discarding WES as potential data input (Zhang et al., 2019). RD, instead, relies on the number of reads mapped to a target region, without making assumptions on read-pairs distance or soft-clipping, making this the only successfully consolidated approach for CNV calling using WES (Tan et al., 2014).

CNV detection using Whole-Exome Sequencing

Read Depth approach and its limitations

As already mentioned, WES can be used to detect CNVs exploiting the Read Depth approach. This approach is usually divided in 4 main steps; firstly, all reads mapping to bins or windows of fixed sizes are counted and then the resulting read counts (RC) are normalized by taking into account the local GC content and mappability. The normalized RC signal is then split into segments having the same number of copies of DNA using algorithms like the circular binary segmentation (CBS) or those based on Hidden Markov models (HMM). Finally, for each segment the actual number of copies

is estimated by considering that the number of reads mapping to a region should be proportional to the number of times the region is present in the DNA (Tattini et al., 2015).

However, the accurate detection of CNVs using WES is hampered by intrinsic biases of the technique that cannot be completely removed within the normalization step of the RD approach (Gordeeva et al., 2021); some of these biases will be discussed in the following paragraph. While WES offers greater coverage than WGS, its distribution cannot be assumed to be gaussian as it varies significantly across the enriched regions. This variability is mainly caused by an unbalanced capture efficiency across the genomic regions, due to several factors that are intrinsic to the nature of the enriched regions (such as an extreme GC-content, high homology with other regions or poor mappability) or due to technical errors in the library preparation, capture and sequencing. Specifically, regions with a very high (>70%) or very low (<30%) GC-content are not easily hybridized to capture probes and poorly amplified using PCR during library preparation, producing a lower number of targeted sequences than expected (Roca et al., 2019). Highly homologous regions are characterized by similar or identical sequences that can be contiguous (repetitive regions) or interspersed in the genome and results in sequenced reads difficult to align correctly to the reference. Regions with poor mappability are prone to be aligned incorrectly to the reference genome, and this could be caused by either the presence of repetitive regions, mutations or sequencing errors (Roca et al., 2019; Gordeeva et al., 2021). Finally, another source of variability can be pinned to technical bias arising during sample processing and sequencing. Since these sources of error are usually batch-specific, most RD strategies can mitigate them by normalizing the signal of the case against a set of controls that are part of the same batch (i.e. sequenced with the same target capture kit, library preparation reagents and platform) (Kadalayil et al., 2015).

Taken all together, these factors cause coverage fluctuations that can be mistakenly interpreted as CNVs or false negatives, since the RD approach heavily relies on the assumption that coverage and number of copies of every targeted region are positively correlated (Roca et al., 2019; Gordeeva et al., 2021; Rajagopalan et al., 2020).

Another significant disadvantage of CNV calling using WES is its resolution in detecting breakpoints. Since most breakpoints lie outside the targeted sequenced regions, CNV boundaries could be missed or only partially resolved, meaning that the resolution of clinically relevant CNVs could be limited (Kadalayil et al., 2015).

Benchmark of CNV calling tools

To date, several CNV calling tools using WES data have been published, all varying at least in one of the main steps required for the analysis when using the RD approach and having different performances in the detection of CNV types and sizes (Gordeeva et al., 2021) (Table 1).

Tool	Algorithm detail	Features (specifics)	Year
CANOES	Negative binomial distribution, regression-based normalization (GC-content), HMM	At least 15 samples, average targets 6, distance between targets 70 kb, average rate of CNV occurrence in the exome 10−8	2014
CLAMMS	GC-content and average depth normalization, custom reference set using kNN, mixture model, HMM	0.3 < GC < 0.7, mappability > 0.75	2015
cn.MOPS	GC-content and sample normalization, mixture Poissons model and Bayes approach	At least 6 samples Minimum segments 5	2012
CNVkit	In-target and off-target regions, bias (GC-content, repeat-masked fraction, target density) correction using rolling median, CBS	Exclude poor mappable regions	2016
CODEX	Log-linear decomposition-based normalization, Poisson likelihood-based segmentation	0.2 < GC < 0.8 Target length > 20 bp, median target coverage > 20 ×, mappability > 0.9	2015
CoNIFER	Singular value decomposition normalization, ± 1.5 SVD-ZRPKM threshold	At least 50 samples Probes with median RPKM across samples > 1, samples with a standard deviation of SVD-ZRPKM < 0.5	2012
CONTRA	Base-level log-ratios, GC-content, library-size correction, calling region significant based on normal distribution, CBS for large variation	Include regions at least 10-bp long with coverage > 10	2012
DeAnnCNV	Web-server, GC-normalization, HMM of log read counts ratio	CNV evidence threshold > 80	2015
EXCAVATOR2	In and off-target regions, 3-step normalization (GC-content, mappability, region length) segmentation with shifting level model, FastCall algorithm	Read mapq > 1 Min number of targets in CNV 4	2016
exomeCopy	Negative binomial distribution, HMM using background read depth and positional covariates (GC-content, length)	mapq > 1, overlap to include read into region—1 bp, median value for background, transition probability to CNV 1e-4 Transition probability to normal state 0.05	2011
ExomeDepth	Beta-binomial distribution, optimized reference set, HMM	Read mapq > 20, max distance between target border and the middle of paired read to include read into region 300 bp Transition probability to CNV 0.0001 Expected CNV length 50 kb	2012
ExonDel	Deletion in exome or genes of interest, GC-content median correction, calling by comparing to median depth within the gene	Read mapq > 20, base quality > 20, min percent of covered bp for each exon 0.1, max number of exons in CNV 9	2014
FishingCNV	PCA of RPKM, CBS test sample, comparing segment coverage against control set distribution	Read mapq > 15 Base quality 10, RPKM > 3 FDR adjusted *p*-value 0.05	2013
HMZDelFinder	Only deletion, exon and sample filtering, call region with RPKM < 0.65 as deletion, AOH filtering based on VCF, prioritization based on Z-score	Mean RPKM > 7 across samples, deletion frequency < 0.5% Exclude 2% samples with the highest number of deletion	2017
PatternCNV	Log2-transformed RPKM standardization, average and variability pattern training from control samples, smooth bin within exon	Bin size 10 mapq > 20	2014
XHMM	Gaussian distribution, PCA normalization, HMM	At least 50 samples, 0.1 < GC < 0.9, 10 bp < target < 10 kbp, mean coverage > 10 × across all samples, average targets 6, distance between targets 70 kb, average rate of CNV occurrence in the exome 10−8	2012

Table 1 Table from Gordeeva et al., 2021 showing some of the tools published with their specific algorithms and features

In one of the last and most extensive benchmarks, Gordeeva et al. compared 16 CNV callers (Table 1) finding good concordance between some tools, but in general only when comparing no more than 3 or 4 (Fig. 5), differences in the range of lengths detected and an overall predisposition of these tools in detecting deletions over duplications.

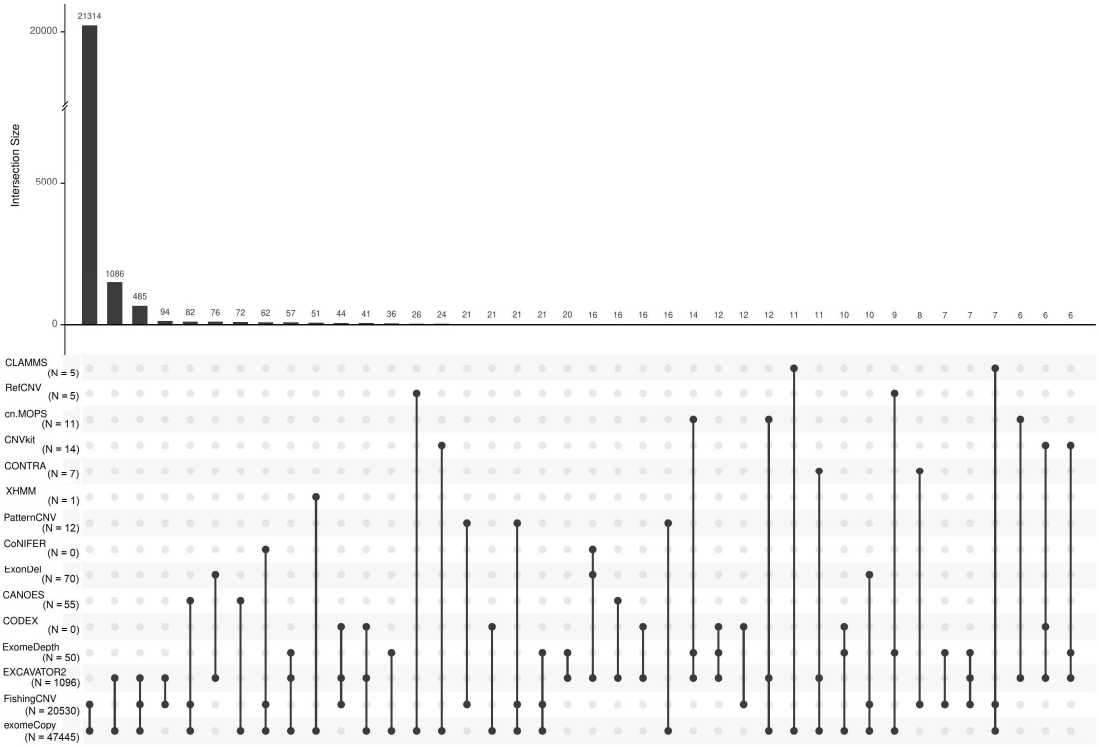


Figure 5 Figure created from Supplementary Table 4 of Gordeeva et al., 2021. Concordance between tested callers; we restricted the plot to intersections between up to 3 tools. Unique CNVs called by each caller are on brackets.

In terms of performance, none of these tools showed a high F1-score (Fig. 6) and the same pattern have been seen in other benchmarks (D'Aurizio et al., 2016; Roca et al., 2019; Zhao et al., 2020).

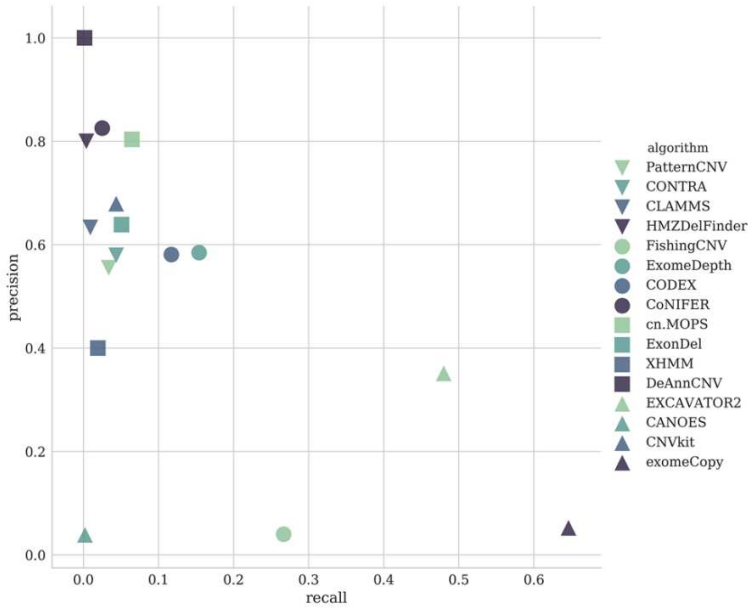


Figure 6 Plot from Gordeeva et al., 2021: overall performance of several CNV callers using WES data

The low performance is not only caused by a poor capture efficiency in specific regions of the genome, increasing both false positives and false negatives, but also by the lack of a well-defined true set covering the entire landscape of CNVs that could be used for the evaluation of these tools, potentially causing the mislabelling of calls as false positives (Gordeeva et al., 2021; Kosugi et al., 2019). Moreover, there are still not extensive benchmarks that stratify the performance on CNV type; thus, the capability of these tools in detecting both deletions and duplications with similar recalls is still unknown.

Also, since most benchmarks use as positive sets variants detected with microarrays, where the minimum size detected is around 1kb (Le Scouarnec et al., 2012), very little has been published regarding the capability of these tools of detecting CNVs including 1-2 exons in WES. One of the few examples is given by the work published by Gordeeva et al., where – without distinguishing between deletions and duplications, they show that a very small number of tools out of the 16 evaluated can detect CNVs encompassing 1 to 3 exons with good sensitivity (Gordeeva et al., 2021). In another recent paper, Rajagopalan et al. evaluated the performance of a single tool (ExomeDepth) in the detection of CNVs < 4 exons, showing a sensitivity of 86% for deletions and 87% for duplications (Rajagopalan et al., 2020). A final example is given by Moreno-Cabrera et al., where the authors indeed evaluate the performance of 5 tools in calling correctly both single target regions and genes. However, performances are related to CNV calling using NGS panels, where a small set of genes of interest is targeted and sequenced with very high coverage. Thus, results from this work cannot be easily translated to a WES setting.

Several new studies have shown that adding CNV discovery to WES analysis could increase the overall diagnostic yield (around 22-28% considering only SNVs and InDels) of about 1.6-4.7% (Marchuk et al., 2018; Truty et al., 2019; Retterer et al., 2016). Moreover, WES could potentially reach a resolution of 200 bp (i.e., the average size of a single exon) while the current gold standard for CNV detection in clinics, microarrays, reach a minimum size of ~40kb, missing almost completely the smallest CNVs. Taking into account the capability of WES of also detecting chromosome aneuploidies, having a validated approach for CNV calling would make WES the compelling first-tier, stand-

alone diagnostic test for a broad spectrum of conditions, especially neurological diseases and cancers (Marchuk et al., 2018).

Aims

Whole-Exome Sequencing is nowadays one of the first-tier tests used in clinics to find mutations in Mendelian diseases; this is due both to its decreasing cost and easiness in processing thanks to a set of well-defined best-practice workflows for SNV and small InDels detections (Chong et al., 2015; Gilissen et al., 2012; Koboldt et al., 2020). However, we still know very little on how CNV detection using WES data could increase WES diagnostic yield. The lack of best practices and reliable gold standard references make this type of analysis still difficult to perform, leaving a great amount of WES data in a "grey area". A plethora of exome CNV callers have been published over the years, however there are very few benchmarks evaluating the minimum resolution obtainable and their capability in detecting 1- and 3-copy imbalances, and no publications at all taking into consideration the minimum target regions required. Nonetheless, it is apparent that the performance of most popular tools is biased towards a certain CNV class (deletions versus duplications) and size range (small versus large), suggesting that the combination of multiple tools is needed to obtain an overall good detection performance. However, such an operation is laborious as it means to choose, tune and maintain a composite set of bioinformatic tools over the years. Considering that clinical CNVs extend from large genomic rearrangements to single gene alterations (Truty et al., 2019), it is quite important to evaluate how accurate WES CNV callers are throughout the entire range of the events to the purpose of being robustly used in the diagnostic setting.

Starting from these considerations, we imagined that Machine Learning could be used to construct a CNV caller able to learn from real data and that it could be at the same time a more flexible and robust solution towards the detection of exonic CNVs differing in class and size. We then reasoned that such a caller could base its functioning on a measure, Normalized Read Counts (NRC), that we could easily retrieve using our previously developed tools EXCAVATOR/EXCAVATOR2 (Magi et al., 2013; D'Aurizio et al., 2016). Furthermore, we exploited the naturally occurring presence in one or two copies of the non pseudo-autosomal X chromosome WES sequences in males and females, respectively. This difference, which is unique among human

chromosomes, can be therefore used to mimic a contrast between a normal-copy status versus a deletion status when taking sequences from females versus males. Similarly, to mimic a duplication status, we merged sequences derived from males and females. We decided to simply merge sequences instead of simulating duplications to preserve the natural data noise occurring in WES sequencing, that is partially lost with simulations. Another advantage of using an almost entire chromosome, selected from either affected or healthy individuals, is the amount of samples available for training a Machine Learning model; this make the approach entirely new, as most ML approaches that works with NGS sequences use for training medium/small positive datasets derived from specific set of mutations and/or cohorts (Smedley et al., 2016; Pellegrino et al., 2021; Zhang et al., 2021).

Here, we developed and tested our caller against a set of the most recent and maintained CNV callers, varying for the method of detection used. We benchmarked all methods using 1000 Genomes Project samples and evaluated their performance in detecting deletions and duplications with selected target regions sizes. Since this type of benchmark has still not been done, we were also interested in understanding how well these tools performed in each different setting. Finally, we exploited a cohort of 251 Italian-derived WES and SNP-array data, part of the Epi25 consortium - a collaborative effort in collecting epilepsy cases worldwide. This valuable set was used to test the robustness of our method.

Overview

This paragraph will briefly introduce all steps described in 'Materials and Methods' section to give a comprehensive idea of the workflow we built.

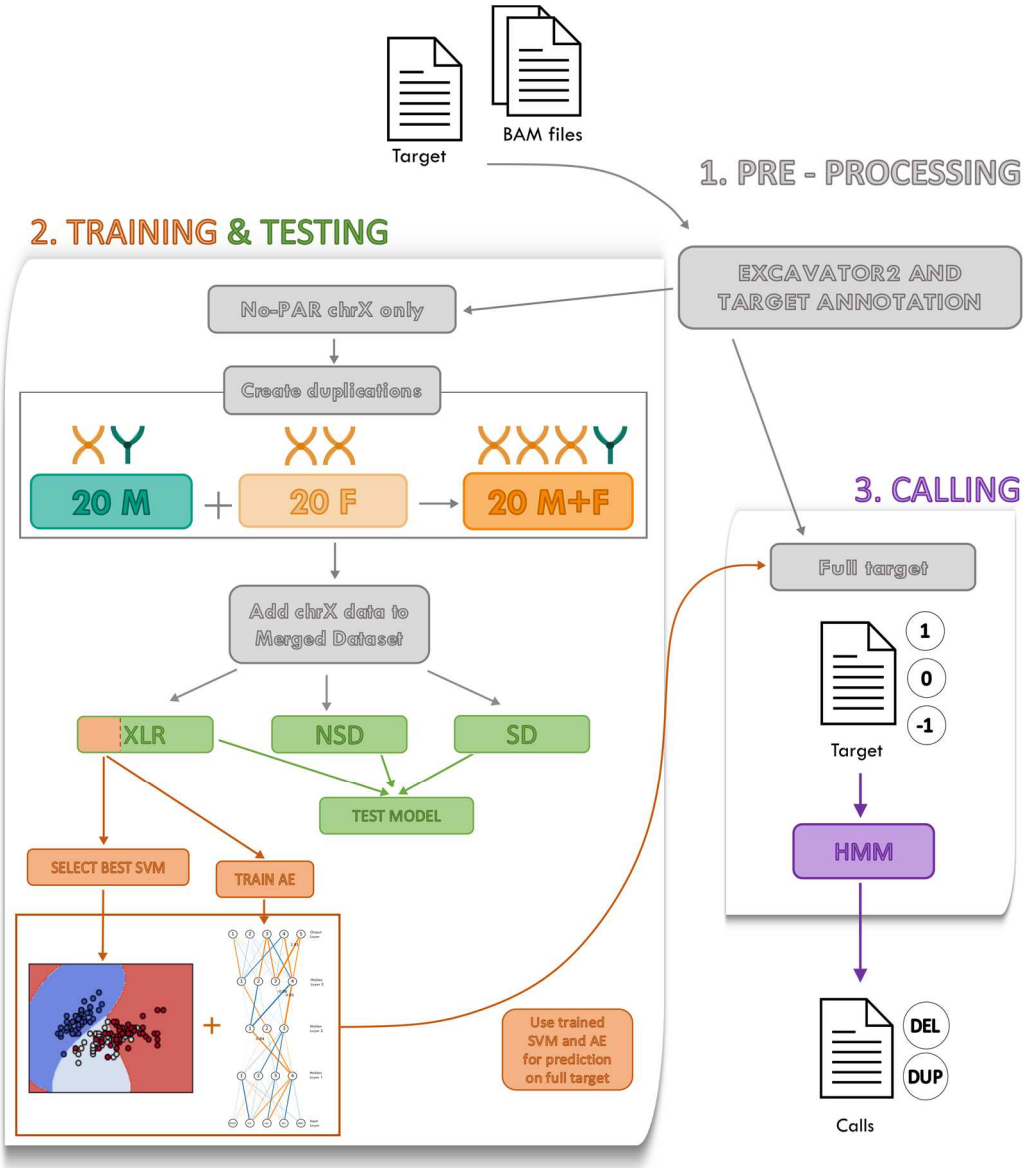


Figure 7 TrainX workflow: in a first pre-processing step, ML datasets with Read Counts and additional features are created starting from samples' BAM files and their targets. No-PseudoAutosomal (PAR) chrX regions are extracted from the datasets and used to create the training set (XLR: X-Linked Recessive, SD: Segmental Duplications, NSD: No Segmental Duplications). Trained SVC and Autoencoder (AE) are used to make predictions on the whole ML dataset. Finally, predictions are used to estimate HMM parameters and call deletions and duplications.

1. Pre-processing

The pre-processing step is required to create the read count datasets needed for Machine Learning algorithms. The workflow starts from a list of BAM files (WES) and their bed file, containing the coordinates of the enriched regions. At the end of this step, a new dataset will be generated for each BAM file containing the following features:

Dataset column	Explanation
Chr	Target region chromosome (not used for ML).
Start	Start position of target region (not used for ML).
End	End position of target region (not used for ML).
GC_content	Target region average GC content.
Length	Target region length.
Dist3	Distance (in bp) between current Start position and End position of previous target region.
MeanCvg	Processed sample autosomal mean coverage.
NRC_poolNorm	Processed sample NRC (computed with EXCAVATOR2), normalized against the average NRC derived from a pool of 10 control female samples and log2.
ID	Processed sample ID.

Table 2 Description of all features inside the datasets created for all samples to call. Number of lanes will be equal to the number of target regions inside the bed file. See 'Dataset construction' in the Material and Methods sections for more details

2. Training and testing

If, for a specific target, there are enough BAM files (we choose to pick 20 males and 20 females but could be lower, as long as the merged training dataset has around 200

samples as seen in the Results, Fig. 16), these samples can be used for training and therefore added to the training dataset as detailed in the following workflow.

For each of the 40 samples chosen for training, the dataset is sliced to keep only data for non-pseudoautosomal regions of chromosome X. A new feature with the class to predict will be added to the dataset: 0 for females to represent a neutral number of copies and -1 for males, corresponding to 1-copy deletions.

Since having 3 copies of chromosome X is not a natural occurring, we recreated this setting by merging male-derived and female-derived chromosome X alignments. As we had data from 20 males and 20 females, by merging data from each couple we obtained 20 new alignment files. Datasets containing read counts and all other features were created in the same way as for males and females, with the exception of MeanCvg, that contains the autosomal mean coverage of the female used during the merging step since it corresponds to the normal number of copies. Finally, these samples are labelled with class “1” to represent 1-copy duplications.

Then, for each of the 60 samples, data on chromosome X is grouped into 3 regions of interest: target regions intersecting X-Linked Recessive genes, segmental duplications or all the remaining regions. These 3 types of target regions are added respectively to the XLR, SD and NSD datasets. At this point, these 3 datasets contain only regions from the batch of 60 samples processed. Before starting to train models, target regions are randomly sampled for all IDs from each batch of target-specific datasets that have been created up to that point. These regions are added to a general training (Merged XLR) and test sets (Merged SD and NSD).

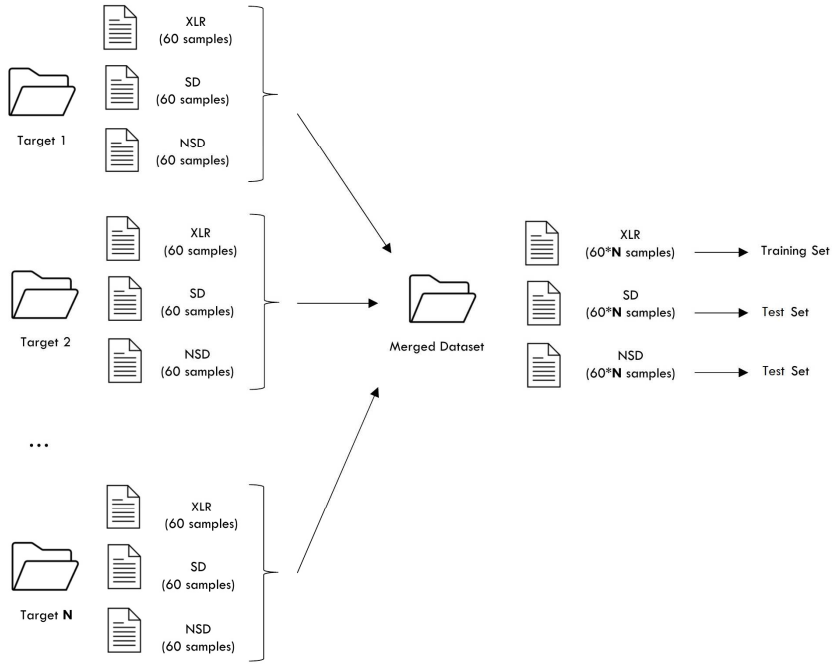


Figure 8 Training datasets data structure. For each target, 3 datasets containing all individuals will be created. The final Merged datasets contain sampled regions from each target and for all individuals.

Thus, when there are enough new samples for training, training and test sets are additively updated with them. Otherwise, the idea behind trainX is that predictions can still be made on the new processed samples using the pre-trained model (i.e., using the algorithm trained on the merged training set without those samples). As for now, the Merged dataset always include RC data from 5 enrichment kits: BGI, MedExome, Nextera, SureSelect V6 and Twist (detailed in Methods).

A fraction of the training set is then used to select the best SVM model while performance is tested on the hold out and the two test sets. As a way of filtering outlier target regions, we used the same training fraction to train an autoencoder (AE). The mean reconstruction error and its standard deviation (over all the reconstructed target regions) for the XLR training fraction is then recorded to be used in the following step.

3. Calling

For all samples selected for CNV calling, their datasets with read counts and annotation for the whole target are used for prediction using the trained SVM. Meanwhile, the trained AE is used to compute the reconstruction error for all target regions; if a target region has a reconstruction error that is higher than the mean XLR reconstruction error, that region is reclassified as neutral copy (class 0).

SVM predictions are used to estimate, for each sample, HMM transition and emission matrices. Viterbi algorithm is used to decode the HMM and add to each target region the most probable state (deletion or duplication). Regions having the same state are merged in single windows, originating the final file with CNV calls.

4. Benchmark

Rationale. We evaluated our method performance by comparing it against a set of well-known CNV callers, and its robustness by running the method on a large set of WES; for all samples used we had a set of gold standard CNVs. All metrics in this benchmark were computed focusing on called target regions instead of considering the entire called window. The reason behind this choice is that our method is focused on using EXCAVATOR2 Read Counts generated for all target regions. Thus, we wanted to evaluate how good our method was, compared to other CNV callers, in classifying correctly small events (encompassing from 1 to 5 target regions).

Real data. We used two sets of real WES data: NA12878 and the epi25 dataset. NA12878 is a well-studied individual derived from 1000 genome project. For this sample there are several gold standard sets of CNVs discovered using different technologies. We selected the gold standard sets built using arrays (array CGH and SNP array) or short read NGS and evaluated how many of these variations were called in NA12878 using TrainX or other CNV callers. Whereas epi25 is a consortium collecting and studying data derived from epilepsy patients. Having access to the Italian WES cohort and the matching SNP-arrays, we could use these data to evaluate the robustness of TrainX performance.

Synthetic data. Due to the lack of universally accepted gold standards for benchmarking CNVs and knowing how unreliable results can be when using existing truth sets to evaluate a caller performance on NA12878, we also evaluated all callers capability of detecting synthetic CNVs. We selected a set of known clinically pathogenic CNVs, ranging from small variations encompassing single target regions to very large ones, and artificially introduced them in NA12878 alignment file.

TrainX model selection. Finally, these datasets were also a precious resource to better understand the best model to use when translating from the X chromosome to the autosomes. As already mentioned in the introduction, read depth across target regions in never uniform. Thus, when we initially tested several models and had to choose the best to use, apart from computing performance in the test sets we also found extremely important to evaluate performance in the autosomes by using metrics obtained for NA12878 real CNVs. When passing from the good set of regions chosen for training to the entire target, we saw that models that showed best performance in the X chromosome during model selection had a drop in performance (with an high increase in number of false positives) in the autosomes. However, out of all models tested, we obtained good results with the SVM Classifier, observing a good generalization property. Thus, we choose SVM as the final model to use with TrainX.

In Material and Methods, all these steps will be described sequentially in 2 main paragraphs, based on what process in needed first to proceed to the next one. Thus, “TrainX” paragraph will describe pre-processing and training steps using the 5 kits we included in the pre-trained model, while the “Benchmark” paragraph will contain information about model re-training, calling and method validation using the benchmark datasets.

Materials and Methods

TrainX

Dataset construction

Sample selection.

We selected 5 groups of WES data differing for the experimental design of the enrichment kit, reference build and sequencing provider:

Capture Technology	Bait Size (Mb)	Reference Build	Library Preparation	Platform	Sequencing Provider
MGIEasy Exome Capture V4 (BGI Genomics, CHN)	58.97	hg19	Hybridization capture	Illumina	BGI
SeqCap EZ MedExome (Roche NimbleGen Inc, USA)	46.58	hg19	Hybridization capture	Illumina	In-House
Nextera Rapid Capture Exome V1.2 (Illumina Inc, USA)	45.33	GRCh37	Hybridization capture	Illumina	Broad Institute
SureSelect Human All Exon V6 (Agilent Technologies, USA)	60.46	hg19	Hybridization capture	Illumina	In-Service
Twist Human Core Exome (Twist Bioscience, USA)	36.71	GRCh38	Hybridization capture	Illumina	In-Service

Table 3 List of target enrichment groups chosen for dataset construction. Each group contains 20 Males and 30 Females.

For each group, we selected WES data of 20 females and 20 males to use for the creation of the training set and 10 additional females to use as controls (i.e., controls for EXCAVATOR2, see "EXCAVATOR2 Normalized Read Count" section in "Pre-processing and feature construction").

Pre-processing and feature construction.

Autosomes mean coverage. In-target mean coverage (MeanCvg) was computed using mosdepth (Pedersen and Quinlan, 2018) giving to option -b the target without chrX and chrY. MeanCvg was computed for the 20 females and 20 males of each group and used as a sample-specific feature.

Duplicated samples. To mimic 3 copies duplications, for each enrichment group we used the SAMtools (Li et al., 2019) command "merge" to merge females and males BAM files having the most similar autosome MeanCvg, creating 20 new alignment files.

For each duplicated sample, we used the matching female autosomes MeanCvg as feature as it is considered the wild-type reference in our model.

Class. The categorical feature we want to predict depends on the sample's gender and was added following these criteria:

Gender	Class	Copy Number	Status
Male	-1	1	DEL
Female	0	2	WT
Merged sample	1	3	DUP

Table 4 Target variable was created considering Males as 1-copy deleted, Females as wild-type and Merged samples as 1-copy duplicated.

EXCAVATOR2 Normalized Read Count. Our sets of 20 females, 20 males, 20 merged samples and the remaining 10 female control samples were processed using EXCAVATOR2 (D'Aurizio et al., 2016). We first run the TargetPerla.pl module to process each enrichment target file, setting a 50kb window. We then run the second module, EXCAVATORDataPrepare.pl, for all bam files to obtain the Normalized Read Counts (NRC) and filtered the RData files keeping only in-target regions (corresponding to our enrichment target file positions).

To normalize the NRC values against the set of female controls, we used the EXCAVATORDataAnalysis.pl module with option "-e pooling" to create data for the pool of females, extracted the in-target NRC values from the Pooling Control RData file, and calculated the ratio:

$$NRC\_PoolNorm(i) = log_2\left(\frac{sWMRC_i}{cWMRC_i}\right)$$

where $WMRC_i$ corresponds to EXCAVATOR2 Window Mean Read Count of sample *s* and the control pool *c* for region *i* of the target file. We then applied the $\log_2$ function to the ratio.

Target-specific features and filtering. As target-specific features, we added the region' GC-content, mean mappability, length and distance to the next 3' region. GC-content percentage was added using pybedtools (Dale et al., 2011) command "nuc" with the genome fasta file used for alignment. To compute the mean region mappability we intersected each target file with the mappability file, specific for hg19/GRCh37 or

GRCh38 depending on the WES group and available within EXCAVATOR2, using pybedtools command “intersect” and computed the interval mean value using pybedtools “merge” with option -o “mean”.

Finally, we filtered each enrichment target removing centromeres, telomeres, scaffolds or other problematic or heterocromatin-rich regions included in the UCSC Gap file provided by EXCAVATOR2 for the two reference builds using pybedtools intersect with option “-v”.

Training datasets.

For each sample of the enrichment groups, we took for training the final target with its annotations limited to chrX and without pseudo-autosomal regions (PAR). PAR are regions of homology between chromosome X and Y located at the end of each arm needed for recombination during meiosis. These are the only regions of the sex chromosomes having an autosomal inheritance, thus we removed them since in males appear as normal copy. Finally, we divided chrX regions into a training dataset and two test sets.

XLR dataset. The training set contains target regions encompassing X-Linked Recessive (XLR) genes. Genes having an XLR transmission are rarely found imbalanced in apparently healthy males and females, making them a conserved group of genes that could be taken as reference for control copy number 1 and 2, respectively. An initial list of XLR genes was drawn by selecting all genes having only a XLR inheritance in our OMIM genemap2 database (2017-10-24 download version). For each of these genes, we selected the chromosomal coordinates corresponding in the GENCODE database (release 19 for hg19/GRCh37 and release 37 for GRCh38) to the canonical transcripts. To be as stringent as possible and avoid selecting regions that could potentially have common CNVs, this list was re-evaluated by checking for the presence of variations encompassing our selected genes in gnomAD, a database containing a collection of variations in the general population. From gnomAD SV v2.1 dataset, we manually selected all CNVs overlapping the CDS regions of our list of genes and their Allele Frequencies (AF) in the population and computed the overall cumulative AF using the R function *cumsum*. Cumulative frequencies were generated separately for deletions and duplications. For both CNV types, we set a cumulative AF threshold < 0.01 and if the

condition was not already met, we removed one by one genes with the most frequent events and re-computed each time the cumulative AF for the new subset.

SD dataset. The "SegDup" test set contains regions of the chrX not overlapping the XLR dataset and intersecting regions with segmental duplications. A bed file with all the chromosomic coordinates of regions with segmental duplications was downloaded from UCSC for all the genomic builds tested and intersected with each target using pybedtools "intersect". For this set, we wanted to evaluate trained models' ability in labelling difficult regions of the genome, expecting a drop in the performance.

NSD dataset. The "NotSegDup" test set include the remaining chrX target regions not overlapping the training set and the SD test set. This set contains the largest number of regions, and was created to evaluate the models' ability to generalize in regions similar to the XLR dataset, but more likely error-prone in the labelling assignment.

Merged datasets.

XLR, SD and NSD datasets from each target enrichment were merged to create 1 single final training set and 2 test sets, respectively. We choose randomly a number of target regions derived from all samples equal to the dimensionality of the smallest target.

Feature importance and selection.

Although the feature vector describing our training dataset is small and does not require any dimensionality reduction, we used feature selection primarily to evaluate which of our features were more informative and to what extent. We used scikit-learn Permutation Feature Importance (PFI) and the python implementation of the Boruta algorithm (Kursa and Rudnicki, 2010) to evaluate feature importance when using Random Forest models trained on the merged XLR dataset.

PFI is described as the decrease in a model performance when a single feature is randomly shuffled. By doing this, the relationship between feature and outcome variable is broken, and changes in model scores suggests how much the model relies on it. Being a model-agnostic method, it can be computed multiple times using different feature permutations. Scikit-learn Random Forest Classifier was set with 20 estimators, a maximum depth of 10, 10 minimum samples for splits and 10 minimum samples in leaves. PFI was then computed on 50% training dataset held-out, evaluating

changes in accuracy score when randomly shuffling each feature 100 times and obtaining an average value.

Boruta is another algorithm designed as a wrapper method, so after training a Random Forest Classifier on the dataset it returns each feature classified as confirmed, rejected or tentative instead of giving a value representing their importance. this algorithm first add randomness to the training set by including shuffled copies of all features, the so called “shadow” features. If the original feature performs better than the best randomized feature, that feature is classified as confirmed. Furthermore, to obtain statistically robust results, this process is iterated n times and the significance of a feature is estimated using a two-tailed binomial test. We used the BorutaPy package together with scikit-learn Random Forest Classifier having the same hyperparameters used for the PFI analysis and performing 100 trials. While training the Random Forest, we also computed the Gini feature importance, or Mean Decrease in Impurity (MDI), defined as the total decrease in node impurity, weighted by the probability of reaching a particular node and approximated by the proportion of samples reaching it, averaged across all trees of the forest.

Minimum number of samples and target regions.

To estimate the minimum number of target regions and samples needed to train our models, we performed a grid-search with a 10-fold cross-validation over a set of parameters using scikit-learn GridSearchCV function, to select the best Random Forest Classifier on increasingly large subsets of our original merged dataset (using fractions: 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.8, 0.9, 1). We held-out 20% of the merged XLR dataset and subsampled the remaining 80% both in terms of number of individuals and number of target regions. Cross-validated Random Forest trained on each subset was then tested for accuracy on the hold-out. The procedure is iterated 10 times to take into account also the randomness of the subsampling.

Algorithm implementation

Training set preparation.

The XLR dataset was split into 20:80 fractions for training and test respectively; we selected a smaller fraction for training to avoid overfitting. We then added Gaussian

noise to our training fraction. Adding noise means creating new samples that are close to the training samples, smoothing the distribution of the input space and increasing its randomness but also expanding the size of the dataset. Thus, a model is less prone to learn training samples and start to learn more general features, increasing generalization. We added white noise to all XLR samples using numpy *random.normal* function, with μ = 0, σ = 0.05 and size equal to the dataset's shape, then selected randomly 20% of the noisy data and added it to the training fraction. Finally, we applied feature scaling to our datasets. Since outliers are a frequent occurrence when using WES data, mostly arising from the MeanCvg feature, we used scikit-learn RobustScaler(), as it use statistics that are robust to outliers. This scaler works independently on each feature of the training set, centring samples by subtracting the median and scaling according to the Interquartile Range.

Model selection.

We selected several models to train using our dataset. For each model, the best combination of hyperparameters, configuration variables that cannot be estimated during training, were selected through a grid search. Scikit-learn GridSearchCV() implementation was used for model selection, selecting F1 score as metric on which the models must be evaluated to balance between Precision, Recall and uneven class distribution, and a stratified 10-fold cross-validation to maintain the distribution of the samples' labels in each split equal to the one of the full dataset. Due to computational and time costs of the entire workflow, only a small set of hyperparameters has been tested.

Model	**Hyperparameters tested**
DecisionTreeClassifier()	criterion : [“gini”, “entropy”] max_depth : [2,5,8,10] min_samples_split : [50,100,150,200] min_samples_leaf : [50,100,150,200]
RandomForestClassifier()	criterion : ["entropy", "gini"] n_estimators : [10,20,50,100] max_depth : [5,10] max_features : ["auto", 3] min_samples_leaf : [100,150,200] min_samples_split : [100,150,200]

SVC()	C : [10,20] kernel : ["rbf"] cache_size : [1000]
KNeighborsClassifier()	n_neighbors : [2,5,10] weights : ["uniform", "distance"]
MLPClassifier()	hidden_layer_units : [4,6,8] hidden_layers : [1,2,3,4] activation : ["tanh", "logistic", "relu"] solver : ["sgd"] learning_rate : ["constant","adaptive"] learning_rate_init : [1e-4, 1e-5] early_stopping : ["false"] n_iter_no_change : [10] tool : [35e-4] max_iter : [200]

Table 5 List of ML models tested and the set of hyperparameters evaluated.

Testing.

The resulting best models were evaluated against the 80% test fraction of the XLR dataset and in the NSD and SD test sets. For each sample of the datasets, class was assigned to the label with the higher prediction's probability.

Autoencoder.

An autoencoder is an unsupervised feed-forward neural network trained to predict the input itself. This model is composed by an encoder that compresses the input data deleting unnecessary information, and a decoder that attempts to reconstruct the input starting from the encoder compressed version. The training is repeated until the reconstruction error (RE), the distance between the original input and the reconstructed data, is minimized. We used an autoencoder to detect and filter outliers, defined in this case as target regions having a high RE. This workflow follows the pre-processing steps used for the supervised training, using the same Gaussian noise and splitting fractions for the XLR dataset. We than scaled the samples using scikit-learn MinMaxScaler() and trained scikit-learn's MLPRegressor(), setting a first compressing layer of size $N_F - 1$ (N_F = number of features), one encoding layer of size $N_F - 2$ and a decoding layer of size $N_F - 1$; we also changed two hyperparameters from their default values: learning_rate_init = 0.0001 and max_iter = 10. REs were computed as the

Euclidean distance between the original sample and the decoded one, and from the resulting vectors we computed the Mean AE. The trained model was then tested on the test fraction of the XLR dataset, the NSD and SD dataset. Predicted target regions having a RE > Mean AE_{XLR} + 2 σ were reclassified as WT.

Hidden Markov Model calling using models' predictions.

After training and testing our models, we made predictions on the autosome regions of our use cases. Since the models we trained do not take into account the sequentiality of the target regions across the genome and we wanted to filter out classified events occurring spuriously, we thought of using an HMM algorithm to add a spatial dimension to our set of predictions. Taking into consideration that it is more probable to see normal copy regions than imbalances, we tried to model this behaviour and extracted the most probable sequence of states to obtain the real signal the model tried to measure. We started by defining these initial settings:

Settings	Parameters
States	{ 'WT', 'DEL' , 'DUP' }
Observations	TrainX predictions

Start probability	{
	'WT' : 0.95 ,
	'DEL' : 0.025,
	'DUP' : 0.025
	}

Start Transition probability	{
	'WT' : { 'WT' : 0.9 , 'DEL' : 0.05 , 'DUP' : 0.05 },
	'DEL' : { 'WT' : 0.05 , 'DEL' : 0.9 , 'DUP' : 0.05 },
	'DUP' : { 'WT' : 0.05 , 'DEL' : 0.05 , 'DUP' : 0.9 }
	}

Start Emission probability	{
	'WT' : { '0' : 0.9 , '-1' : 0.05 , '1' : 0.05 },
	'DEL' : { '0' : 0.05 , '-1' : 0.9 , '1' : 0.05 },
	'DUP' : { '0' : 0.05 , '-1' : 0.05 , '1' : 0.9 }
	}

Table 6 Initial HMM settings

Baum-Welch algorithm from R HMM package was used to estimate, for each sample in analysis, transition and emission matrices starting by an interrupted sequence of

observations (a single chromosome) and its states, and iterating the process 20 times. As sequence of observation, we decided to use chr1 as it is the longest autosome.

Viterbi algorithm was used to identify, for each chromosome independently, the succession of hidden states with the highest probability of generating the sequence of observations. For each state, posterior probabilities were computed using the posterior() function from the R package. Finally, we defined windows containing target regions with constant HMM state and selected those classified as deleted and duplicated to create our set of calls; To each call, we added its posterior probability, given by the average probability of each single state of its window.

Benchmark

To evaluate TrainX performance we used 2 set of data collections, WES data from 1000 Genomes Project and Epi25 consortium, for which we had specific sets of gold standard CNVs. Since TrainX is a method based on the classification of single target regions, we decided to measure performances for all tools used in the benchmark (listed in the following paragraph) for each target region independently. For both data collections, we were as stringent as possible in defining a positive and negative space in which to look for called CNVs. The positive set (PS) was defined as a group of target regions encompassing our gold standard (>50% overlap), whereas the negative set (NS) contains target regions that with good probability are present in normal copy. Starting by this target separation, we defined the 4 possible outcomes as follow:

True positives (**TP**). Number of PS regions overlapping calls (min 50% overlap).

False negatives (**FN**). Number of PS regions not called or called with a discordant CNV type.

False positives (**FP**). Number of NS regions called (min 50% overlap).

True negatives (**TN**). Number of NS regions not called or with <50% overlap with calls.

Then, we derived the following metrics:

Accuracy	F1-score	TPR	TNR	FPR	FNR	Balanced accuracy
$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{2*TP}{(2*TP)+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{FP}{FP+TN}$	$\frac{FN}{FN+TP}$	$\frac{TPR+TNR}{2}$

1000 Genomes Project

Samples selection.

We downloaded 56 WES individual data from 1000 Genomes Project Data Portal selecting the "1000 Genomes on GRCh38" data collection. We restricted our selection to samples from Broad Institute as the main project center (thus, as described in the portal, all samples from this center were enriched with Agilent SureSelect All Exon V2), sequenced in an Illumina Platform and having more than 80% of bases covered more than 20X. We then selected from different populations a balanced set of males and females (25 F, 25 M), 5 additional F to use as controls for TrainX and NA12878 as the use case.

Index	Sample ID	Population	Gender	Main Project Exome Center	Main Project Exome Platform	% Target Covered >20x
1	HG00731	PUR	M	Broad Institute	ILLUMINA	0.96
2	HG00734	PUR	F	Broad Institute	ILLUMINA	0.93
3	NA11920	CEU	F	Broad Institute	ILLUMINA	0.94
4	NA12842	CEU	M	Broad Institute	ILLUMINA	0.93
5	NA18853	YRI	M	Broad Institute	ILLUMINA	0.93
6	NA18867	YRI	F	Broad Institute	ILLUMINA	0.95
7	NA18964	JPT	F	Broad Institute	ILLUMINA	0.94
8	NA19000	JPT	M	Broad Institute	ILLUMINA	0.93
9	NA19347	LWK	M	Broad Institute	ILLUMINA	0.94
10	NA19474	LWK	F	Broad Institute	ILLUMINA	0.93
11	HG00551	PUR	F	Broad Institute	ILLUMINA	0.92
12	HG00739	PUR	M	Broad Institute	ILLUMINA	0.92
13	HG01054	PUR	M	Broad Institute	ILLUMINA	0.94
14	HG01067	PUR	F	Broad Institute	ILLUMINA	0.91
15	HG00737	PUR	F	Broad Institute	ILLUMINA	0.95
16	NA06986	CEU	M	Broad Institute	ILLUMINA	0.93
17	NA07048	CEU	M	Broad Institute	ILLUMINA	0.92
18	NA11918	CEU	F	Broad Institute	ILLUMINA	0.92
19	NA12890	CEU	F	Broad Institute	ILLUMINA	0.93

20	NA18499	YRI	F	Broad Institute	ILLUMINA	0.93
21	NA18516	YRI	M	Broad Institute	ILLUMINA	0.93
22	NA18870	YRI	F	Broad Institute	ILLUMINA	0.93
23	NA18910	YRI	M	Broad Institute	ILLUMINA	0.92
24	NA18986	JPT	M	Broad Institute	ILLUMINA	0.93
25	NA18999	JPT	F	Broad Institute	ILLUMINA	0.93
26	NA19055	JPT	M	Broad Institute	ILLUMINA	0.91
27	NA19065	JPT	F	Broad Institute	ILLUMINA	0.93
28	NA19334	LWK	M	Broad Institute	ILLUMINA	0.94
29	NA19346	LWK	M	Broad Institute	ILLUMINA	0.93
30	NA19471	LWK	F	Broad Institute	ILLUMINA	0.92
31	HG00641	PUR	F	Broad Institute	ILLUMINA	0.92
32	HG00736	PUR	M	Broad Institute	ILLUMINA	0.91
33	HG01048	PUR	M	Broad Institute	ILLUMINA	0.93
34	HG01080	PUR	F	Broad Institute	ILLUMINA	0.91
35	HG01110	PUR	M	Broad Institute	ILLUMINA	0.93
36	HG01168	PUR	F	Broad Institute	ILLUMINA	0.92
37	NA12399	CEU	M	Broad Institute	ILLUMINA	0.92
38	NA18501	YRI	M	Broad Institute	ILLUMINA	0.93
39	NA18520	YRI	F	Broad Institute	ILLUMINA	0.91
40	NA18873	YRI	F	Broad Institute	ILLUMINA	0.92
41	NA18912	YRI	F	Broad Institute	ILLUMINA	0.93
42	NA18982	JPT	M	Broad Institute	ILLUMINA	0.93
43	NA19003	JPT	F	Broad Institute	ILLUMINA	0.92
44	NA19058	JPT	M	Broad Institute	ILLUMINA	0.93
45	NA19092	YRI	M	Broad Institute	ILLUMINA	0.92
46	NA19116	YRI	F	Broad Institute	ILLUMINA	0.92
47	NA19130	YRI	M	Broad Institute	ILLUMINA	0.92
48	NA19213	YRI	M	Broad Institute	ILLUMINA	0.90
49	NA19235	YRI	F	Broad Institute	ILLUMINA	0.94
50	NA19338	LWK	F	Broad Institute	ILLUMINA	0.91
51	HG00732	PUR	F	Broad Institute	ILLUMINA	0.96
52	NA19093	YRI	F	Broad Institute	ILLUMINA	0.93
53	NA06989	CEU	F	Broad Institute	ILLUMINA	0.89
54	NA12058	CEU	F	Broad Institute	ILLUMINA	0.88
55	NA19064	JPT	F	Broad Institute	ILLUMINA	0.89
56	NA12878	CEU	F	Broad Institute	ILLUMINA	0.88

Table 7 List of BAM files downloaded from 1000 Genomes Data Portal with their characteristics. Index column is color-coded to show how we separated samples in subsets of 10, 30 and 50. In grey: additional samples used to create EXCAVATOR2 pooling control for TrainX. Purple: sample used for benchmarking all callers.

Target reference dataset construction.

Since originally Agilent SureSelect All Exon V2 target was designed using hg19 coordinates, we had first to lift all positions over the GRCh38 assembly using UCSC LiftOver. New contiguous target regions were merged using pybedtools merge. We

then filtered out all the alternative contigs, chrX and chrY and created our target reference dataset.

NA12878 negative space.

We constructed NA12878 NS restricting to the genomic regions with the lowest probability of being CNVs. Thus, we downloaded from DGV the most recent "DGV variants" database mapped to the GRCh38 assembly (GRCh38_hg38_variants_2020-02-25.txt) and selected all variants found in NA12878. We then removed all regions of the target reference dataset with any overlap with this database using pybedtools intersect and option "-v". Finally, we computed NA12878 MeanCvg for each region of this final target using mosdepth and filtered out all target intervals <50X.

NA12878 positive space.

DGV. To create our sets of real NA12878's CNVs we selected from NA12878 DGV's variants 2 subsets based on the discovery technology used (array-based and short-reads). We then intersected each subset with our target reference dataset using pybedtools intersect with option "-wo" and computed for each overlapping region the ratio between the number of shared bases and the length of the target region, selecting only regions with >50% overlap with the variants. Finally, we carefully evaluated the studies to include in our sets when needed.

1. **arrays.** From the set of array-based studies (SNP-array, Oligo aCGH and BAC aCGH) we decided to remove CNVs from Redon et al., 2006 detected using BAC aCGH (leaving those detected with SNP-array), variants from Kidd et al., 2010 because they were classified as novel sequence insertion without distinction between deletions and duplications and, for studies with more than one reference, we removed all precedent studies. After this filtering, we retained CNVs detected with SNP-array from Altshuler et al., 2010, Cooper et al., 2008, McCarroll et al ., 2008, Redon et al., 2006, Wang et al., 2007 and variants detected with Oligo aCGH from Campbell et al., 2011 and Conrad et al., 2009.

   To evaluate the number of target regions in common between these studies, we used pybedtools multiinter.

2. **short-read sequencing.** For our short-read sequencing set we selected 1000 Genomes Consortium Phase 3 study from DGV and excluded all previous studies since they were done with different combinations of technologies.

Synthetic Variants. From dbVar, we downloaded the most recent GRCh38 datasets for pathogenic non-redundant deletions and duplications. We first removed all variants overlapping NA12878's DGV dataset using pybedtools intersect with option "-v" and intersected the remaining CNVs with the target reference dataset, excluding those having at least one target region with less than 50% overlap. Afterwards, we computed NA12878's MeanCvg in these target regions using mosdepth and removed all CNVs having more than 20% of target regions with <50X depth. Finally, we created 5 lists containing CNVs of different sizes (1, 2 to 5, 6 to 10, 11 to 50 and >50 target regions) and separated by at least 10 target regions. We used these final set of files as templates to add in-silico variants in NA12878 BAM file using XomeBlender (Semeraro et al., 2018), recreating germline single copy deletions and duplications and generating 5 BAM files.

Variant Callers used.

NA12878 original BAM file and those containing synthetic CNVs were analysed with a set out of the most up-to-date germline CNV callers published: GATK4 gCNV, ExomeDepth, DECoN, CNVkit and EXCAVATOR2. As all these tools apply normalization against a set of controls, we tested their performance pooling 10, 30 or 50 samples, always balancing the number of males and females included (Table 6).

CNV Caller	Methods applied	Suggested number of control samples	Suggested CNVs detected	Modifications from default options
GATK gCNV v4.1.2 (Babadi et al., 2019)	Negative-binomial factor analysis to remove sequencing biases and hierarchical HMM to detected sample CNVs and global regions of CNV activity	>30	Detect common and rare CNVs; size with best accuracy not specified	*GermlineCNVCaller* with options --cnv-coherence-length 10000.0 and --class-coherence-length 10000.0

ExomeDepth v1.1.15 (Plagnol et al., 2012)	Beta-binomial model considering region GC content and read-count variance, HMM to combine likelihood across multiple exons.	5-10	Detect CNVs not present in the controls; best for 1-50 kb size CNVs	Each case run separately against the controls; *select.reference.set* options n.bins.reduced = 10000 and bin.length = (Target end - Target start)/1000); *CallCNV* expected.CNV.length = 10000
DECoN v1.0.2 (Fowler et al., 2016)	Modification of ExomeDepth with transition probabilities of HMM depending on exons' distance	1 or more	Optimized for target panels; best for small rare CNVs (1 or few exons)	Each case run separately against the controls
CNVkit v0.9.7 (Talevich et al., 2016)	Use both In- and off-target regions, rolling median to estimate GC-content, sequence repeats and target density correction bias trend; signal segmentation using CBS	1 or more	Doesn't detect small common CNVs; higher accuracy with size >1 Mb	No modifications
EXCAVATOR2 (D'Aurizio et al., 2016)	Use both In- and off-target regions, GC-content, mappability and exon size bias correction; segmentation and CNV call using HSLM and FastCall algorithms	1 or more	Detect common and rare CNVs; CNVs up to few Mb (panels or genes)	*TargetPerla.pl* windows' size tested: Window: 10000 Window: 20000 Window: 50000

Table 8 Set of CNV callers tested in the 1000Genomes benchmark, together with the method applied and options that we modified from default.

For TrainX, we picked 20 males and 30 females (10 of these to use as controls) from the list of BAM files, generated the XLR, NSD and SD datasets as detailed before and added them to the merged datasets. The algorithm was then re-trained and used to make predictions on NA12878 dataset, the latter created separately with features annotated for all target regions.

Epi25

Samples selection.

Epi25 Consortium Year 1 and Year 2 exomes for the Italian Cohort of epilepsy patients were downloaded from Terra (https://anvilproject.org/data/studies/phs001489). This dataset consists of 291 patient's human peripheral blood-derived samples. All samples were sequenced at the Broad Institute, using an Illumina platform and Illumina Nextera Rapid Capture Exomes for target enrichment. BAM files were aligned to the GRCh38 reference build.

For the same set of patients, we also had the relative SNP-array data. All samples were genotyped using a GSA-MD v1.0 array (Illumina, San Diego) with 688032 total markers. SNP-array files were filtered following a genotype-specific QC step and CNVs where then called using PennCNV (Wang et al., 2007); further QC steps were described and applied to produce a high-quality call-set in the study publication (Niestroj et al., 2020). We were given access by the authors to this high-quality dataset, restricted to the Italian cohort.

TrainX calls.

We picked 30 females and 20 males out of 291 exomes to construct the training dataset. Samples were selected excluding those having "high quality" CNVs in the SNP-array and having none or not clinically relevant CNVs in the raw SNP-array dataset. We re-trained the merged XLR dataset after including Epi25 samples and made predictions and calls in 281 samples (pool of 10 females controls excluded).

Target reference dataset construction.

We first used UCSC LiftOver to convert Nextera Rapid Capture Exomes target from hg19 to GRCh38. As for 1000 Genomes, we removed all alternative contigs and sex chromosomes to create the target reference dataset.

Epi25 negative space.

The negative space was created dynamically for each sample using the original raw CNV call-set derived from the SNP-array data. For each sample, we selected the specific set of CNV calls and removed them from the target reference dataset using pybedtools intersect with option "-v".

Epi25 positive space.

To evaluate TrainX performance, we used 3 sets of calls derived from the SNP-array data. We followed some of the steps used by Niestroj et al., 2020, which in turn were mostly derived from Marshall et al., 2017.

As a first step, we removed outlier and not comparable samples from both SNP-array data and TrainX calls. From the raw SNP-array set, we selected patients having >100 CNVs called (as done in Niestroj et al., 2020), and patients for which we did not have the corresponding WES data. From TrainX we selected all patients used for the pooling control and samples having an average autoencoder RE above the 90th quantile of the overall samples' RE distribution. The resulting list of samples was filtered out from both datasets.

This first step was required to precisely evaluate the number of recurrent CNVs in the analysed cohort. We used BEDOPS (Neph et al., 2012) command bedmap with options "--fraction-map 0.5" and "--echo-map-range" to compute the number of CNVs having >50% overlap with each other; a new column describing the number of times a specific CNV was seen has been added to each call using python. The resulting dataset was then annotated using AnnotSV (Geoffroy et al., 2018) using 50% non-reciprocal overlap as option. We proceeded removing CNVs recurrent in >1% of individuals as they were probably array-specific artifacts, and Copy Number Polymorphisms (CNPs) with >1% frequency in DGV, gnomAD or 1000 genomes databases. Following Marshall et al., 2017 analysis workflow, we removed all CNVs encompassing centromeres and telomeres or having ≥50% overlap with repeated sequences, segmental duplications and chromosomal segments containing immunoglobulins, T-cell receptors and MHC genes. All required bed files were downloaded from UCSC TableBrowser, with the exception of the immune response genes segments, that were manually selected from IMGT database for build GRCh38. Finally, we selected all CNVs intersecting with >50% overlap the target reference dataset, excluded calls if they encompassed < 3 target regions and constructed the 3 positive sets (PS).

High quality PS. This is the call-set used by Niestroj et al., 2020 but we further filtered it excluding outliers/not-comparable samples, recurrent variants/CNPs and variants with >50% intersection with problematic regions, as described above. This set contains

CNVs including ≥ 20 SNPs, having ≥ 20kb of length and ≥ 0.0001 SNP density (defined as SNPs/length). If a call was ≥ 1Mb and spanned more than 20 SNP it was included regardless of density.

Medium quality PS. This set includes CNVs containing between ≥10 and <20 SNPs, having ≥10 and <20Kb length and >0.0001 density.

Low quality PS. This final set contains CNVs with <10 SNPs, length <10Kb and >0.0001 density.

TrainX performance evaluation.

TrainX calls filtering. We filtered TrainX call-sets in the same way done for the positive sets except for the density filter; all calls were concatenated in a single file to evaluate recurrent CNVs. To obtain a value similar to the array SNP density, we computed the ratio between the overall number of bases of the target regions included in the call and the length of the call and removed calls having density falling below the 5^{th} quantile of the density distribution.

Metrics and validation. Since we were analysing a consistent number of samples and the positive sets were very small in comparison, we also added as metrics macro-averaged precision and recall using the scikit-learn metrics module. Finally, a variant was considered validated if it was seen by both SNP-array and TrainX.

Results

Feature space evaluation

Mean coverage and NRC evaluation on SureSelect

For MeanCvg and NRC, we did some prior evaluations on samples enriched using SureSelect. Since we expected a slight difference in MeanCvg between males and females due to a 1:2 ratio in chrX sequences, we computed the MeanCvg considering both sex and autosomal chromosomes and exclusively the autosomes.

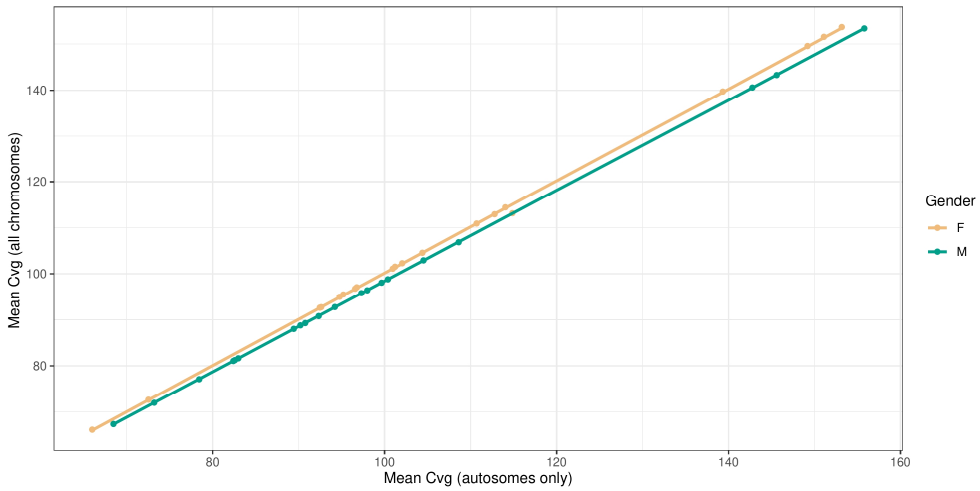


Figure 9 Comparison between Mean Coverage computed using all chromosomes and autosomes only in SureSelect XLR dataset.

Looking at the dot plot, the slight decrease in males' MeanCvg when considering also chrX and chrY was confirmed. As we did not want to influence the training by adding this dependency, we restricted MeanCvg computation to the autosomes.

To find the best set of NRC values to use for training, we tested several conditions: our set of NRC values straight out of EXCAVATOR2 DataPrepare.pl step (i.e., without further normalizations), after normalizing using 1, 5 or 10 females control samples and after normalizing with 10 females control samples and $\log_2$. For all conditions, we then proceeded to create the XLR dataset and evaluated how the NRC was distributed on each class.

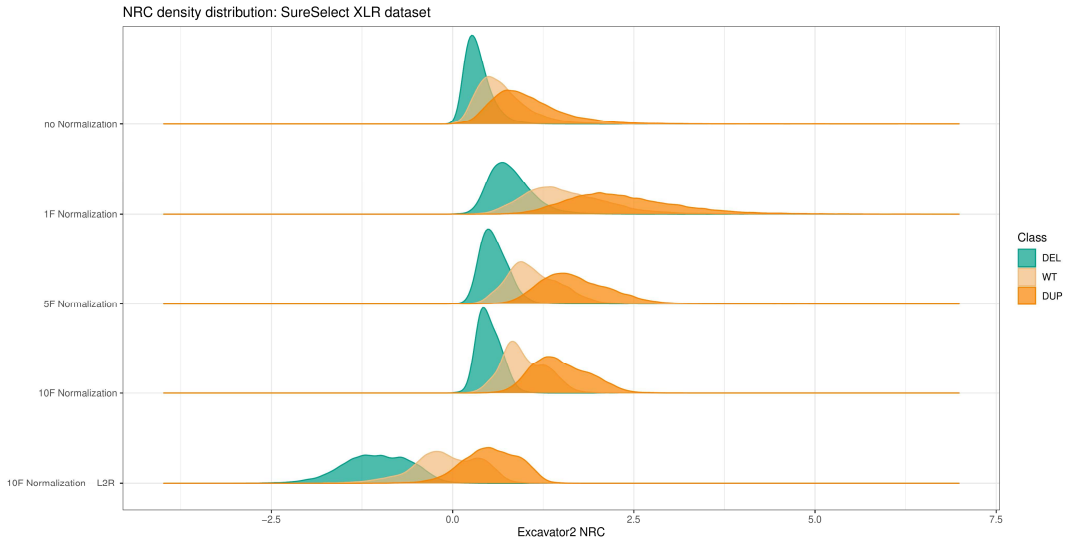


Figure 10 NRC density distribution using different settings in SureSelect XLR dataset.

Type of analysis	Pct overlap (DEL ∩ WT)	Pct overlap (WT ∩ DUP)	Pct overlap (DEL ∩ DUP)
No Normalization	0.5	0.66	0.26
1F Normalization	0.36	0.55	0.11
5F Normalization	0.32	0.52	0.08
10F Normalization	0.29	0.5	0.05
10F Normalization – L2R	0.29	0.5	0.05

Table 9 NRC density distribution using different settings in SureSelect XLR dataset: percentage of overlap between each class

Type of analysis	DEL		WT		DUP	
	Mean	SD	Mean	SD	Mean	SD
No Normalization	0.35	0.18	0.69	0.36	1.03	0.5
1F Normalization	0.8	0.32	1.58	0.62	2.38	0.83
5F Normalization	0.58	0.2	1.13	0.4	1.7	0.5
10F Normalization	0.5	0.2	1	0.3	1.5	0.4
10F Normalization – L2R	-1.06	0.5	-0.1	0.5	0.5	0.4

Table 10 NRC density distribution using different settings in SureSelect XLR dataset: distribution mean value and standard deviation for each class

From this preliminary evaluation, an additional normalization step using a females' control set is required. Furthermore, adding more female samples to the pool

increasingly reduced the distributions' standard deviation and centered the data towards the expected CNV signal (Table 10: with 10 females the average signal for deletions is 0.5, 1 for wt and 1.5 for duplications, as you would expect, for example, when using array CGH). When using a pool of 10 female samples, we also saw a better classes' separation (Table 9). Since a minimum of 10 female samples seems an acceptable requirement to use TrainX, (many WES CNV callers need a higher number of control samples; i.e., see Table 8), we proceeded using this final setting (ratio between the NRC of the sample and the average NRC of 10 female samples) for all samples and enrichment kits. Moreover, we decided to use the $\log_2$ ratios as normalized measures.

All features distribution

After finalizing the two sample-specific features, we evaluated all features in the XLR dataset. This was done for each kit separately and the resulting Merged dataset.

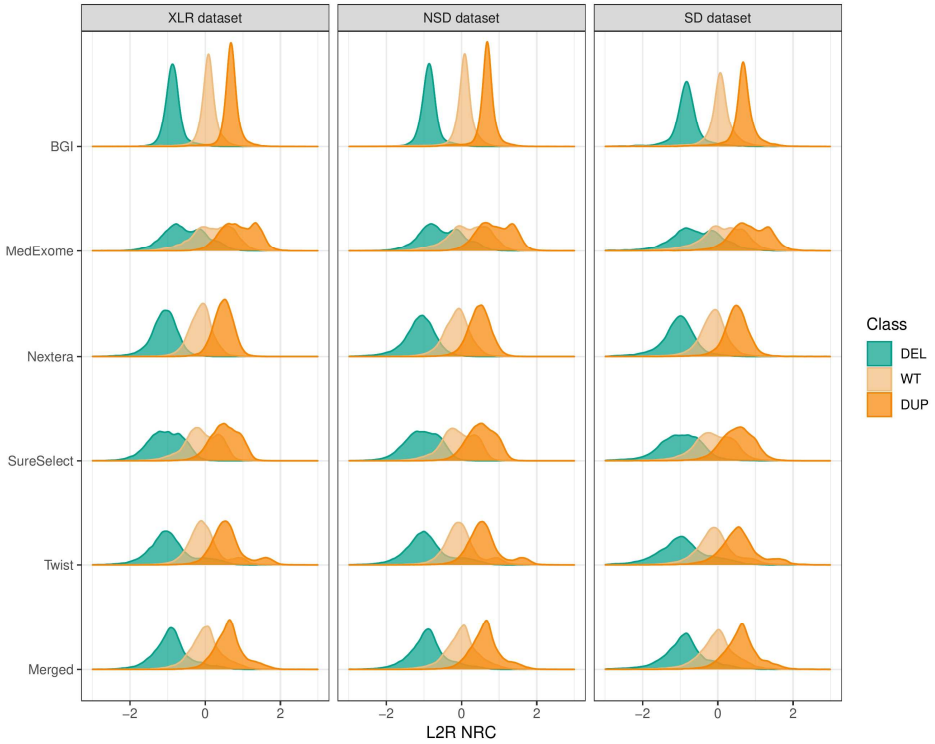


Figure 11 NRC_PoolNorm distribution for all classes in the 3 datasets and for all enrichment kits.

In the XLR dataset, NRC_PoolNorm distribution appears to be different for each target. For BGI especially, the 3 peaks describing each class are narrow and almost perfectly

separated. MedExome, the only set containing a collection of samples sequenced in-house and in different batches (as an example of a situation users could find in their real-world cohorts), shows the worst separation. Altogether, the Merged dataset takes values from all target groups and shows a good NRC distribution. We also evaluated NRC distribution in the 2 test datasets; for all target groups class separation in the NSD dataset is similar to our goldset, while it is slightly worse using the SD dataset, as expected.

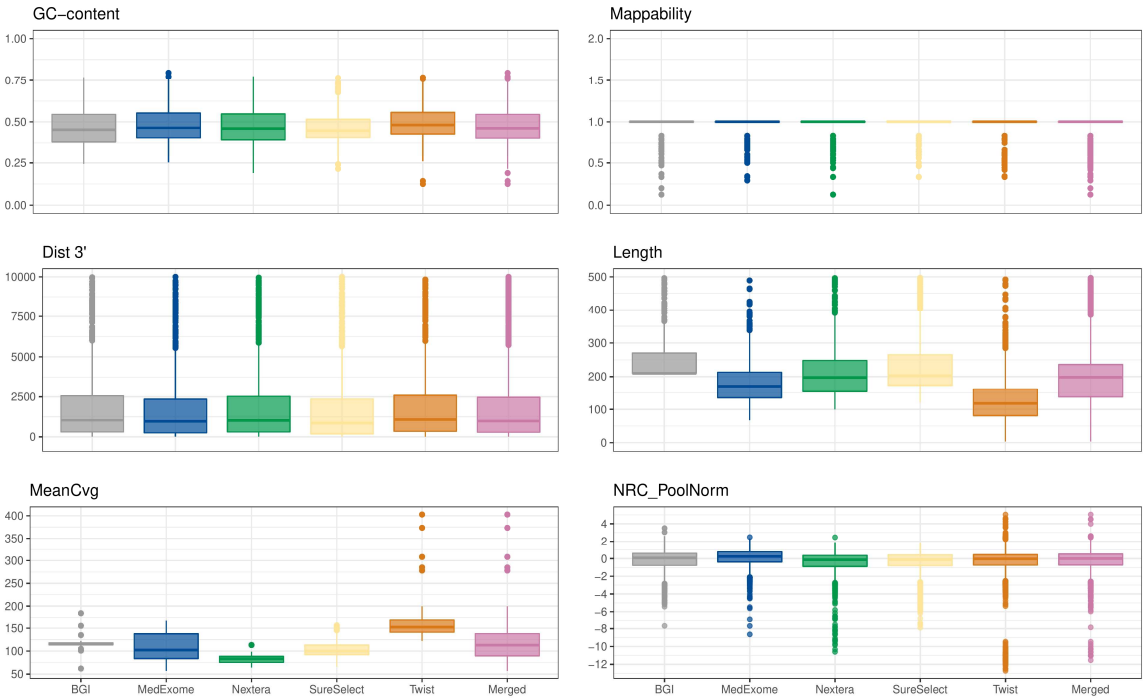


Figure 12 All features distribution in XLR datasets.

We then considered the distribution of all the features in the XLR dataset. For target-specific features, GC-content and Length show the highest variability between enrichment groups. Whereas MeanCvg has specific ranges for each enrichment group based on the theoretical coverage and flowcell chosen for that batch. Since MedExome includes samples from different batches of sequencing, this is the group having the highest variability in terms of distribution. For all features, the Merged dataset embeds most of the differences arising in each enrichment group.

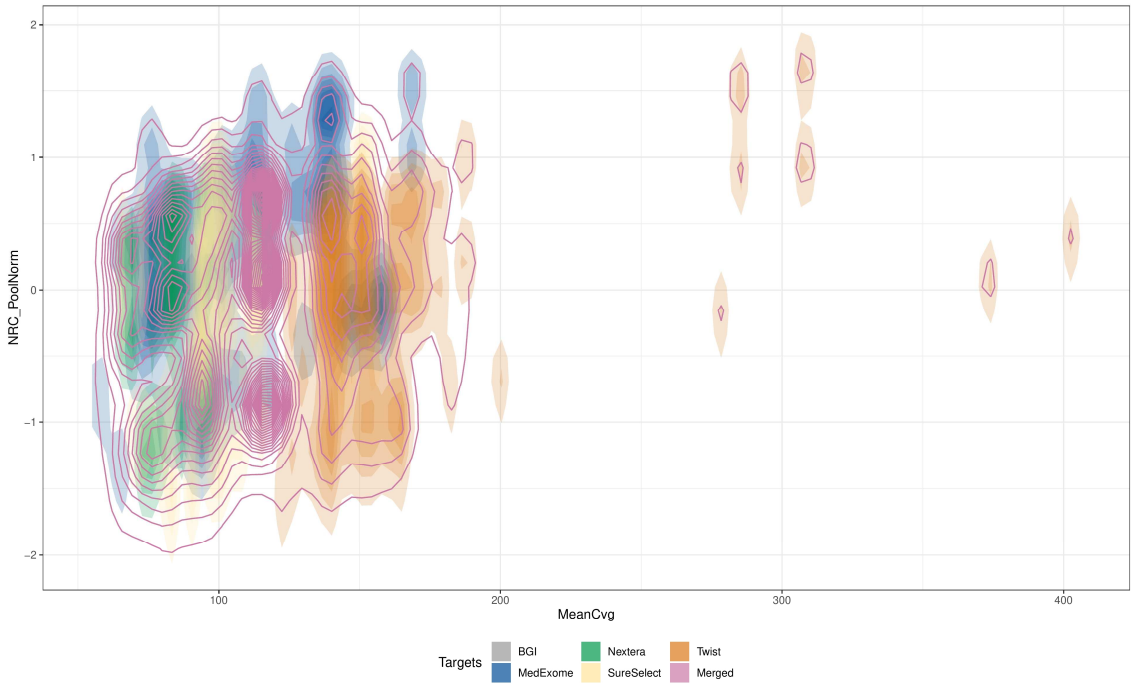


Figure 13 Relationship between NRC_PoolNorm and MeanCvg using all enrichment kits. 2D density plots for all enrichment kits are color-filled, Merged dataset is overlayed on top colouring only level lines.

Finally, we evaluated the relationship between NRC_PoolNorm and MeanCvg with a 2D density plot. For our Merged dataset, the plot shows the highest density in the range of 50-150X for MeanCvg, with NRC_PoolNorm values going on average from -1 to 1. Samples having a higher MeanCvg show NRC_PoolNorm ranges shifted towards the right side of the distribution; these behaviour in our case represents few outliers and shows that predictions could potentially go wrong with samples having depths higher than those mostly represented in our Merged dataset.

XLR genes selection

To construct the training set, we selected a list of 82 XLR-only genes from OMIM database and hand-picked from gnomAD SV v2.1 all CNVs encompassing these genes' CDS.

OMIM XLR gene	gnomAD SVs Variant ID	gnomAD AF
ABCD1, BCAP31	DUP_X_54633	7.58E-03
AFF2	DEL_X_190491	6.20E-05
	DUP_X_54535	1.26E-04
AVPR2	DUP_X_54637	2.59E-03
BRWD3	DEL_X_187516	6.28E-05
	DUP_X_53649	1.26E-03
	DUP_X_53648	3.16E-04
	DUP_X_53650	6.32E-05
	DUP_X_53652	6.32E-05
BTK	DEL_X_188496	3.72E-04
CYBB	DUP_X_52971	6.96E-04
DKC1	DUP_X_54645	1.24E-04
DLG3	DUP_X_53514	6.22E-05
	DUP_X_53512	6.32E-04
F8	DEL_X_190808	1.24E-04
	DUP_X_54658	3.16E-04
	DUP_X_54657	6.20E-05
IDS	DEL_X_190503	2.48E-04
	DUP_X_54543	6.20E-05
	DUP_X_54546	6.21E-05
IGSF1	DUP_X_54261	4.43E-04
IL1RAPL1	DUP_X_52855	1.24E-04
	DUP_X_52854	6.20E-05
LAS1L	DUP_X_53436	3.16E-04
	DUP_X_53442	6.20E-05
MAMLD1	DEL_X_190538	6.43E-05
MID1	DEL_X_184671	8.53E-04
	DUP_X_52617	1.90E-04
	DUP_X_52621	6.21E-05
	DUP_X_52622	6.21E-05
	DUP_X_52615	8.85E-04
OCRL	DUP_X_54242	6.21E-05
OGT	DEL_X_187138	6.20E-05
PAK3	DUP_X_54030	6.32E-05
PGK1	DUP_X_53627	6.32E-05
PGK1, ATP7A	DUP_X_53626	6.20E-05
PHF8	DUP_X_53235	2.53E-04
PHKA1	DUP_X_53547	6.20E-05
PHKA2	DUP_X_52716	1.24E-04
	DUP_X_52723	2.17E-03
	DUP_X_52720	6.20E-05
POF1B	DEL_X_187764	2.56E-04
POLA1	DUP_X_52803	1.26E-04
	DUP_X_52807	6.32E-05
SLC6A8	DEL_X_190712	6.20E-05
	DEL_X_190710	6.21E-05
SSR4	DUP_X_54635	2.09E-02
STS	DEL_X_184530	3.10E-04
	DUP_X_52541	1.24E-04
	DUP_X_52538	2.28E-03
	DUP_X_52545	6.20E-05
TEX11	DEL_X_187098	6.27E-05
	DEL_X_187104	6.33E-05
	DUP_X_53516	1.08E-03
TEX11, DLG3	DUP_X_53513	6.32E-05
TRAPPC2	DUP_X_52644	1.26E-04
TSPAN7	DEL_X_185734	6.20E-05
	DUP_X_52986	8.22E-04
WAS	DUP_X_53132	3.16E-04
ZC4H2	DUP_X_53432	1.24E-04
	DUP_X_53420	1.26E-04
	DUP_X_53434	6.32E-05

Table 11 List of OMIM selected genes having CNVs in gnomAD SV v2.1. Database ID for each variant and the related allele frequency are reported.

From this list, containing genes out of the 82 selected having a variant in gnomAD, we computed the overall cumulative allele frequency separately for deletions and duplications.

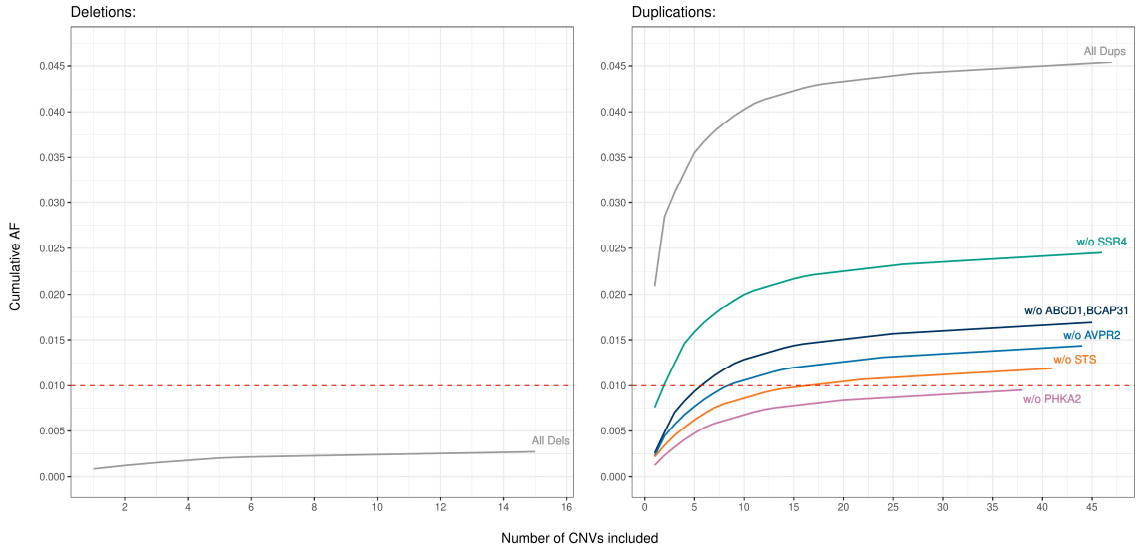


Figure 14 Cumulative gnomAD deletions and duplications allele frequencies. Threshold set at 1% AF.

After plotting the cumulative AF, we could evaluate visually which genes carried the most frequent CNVs.

For deletions, the cumulative AF was lower than the 1% threshold we set, and no further analysis was required.

For duplications, we dropped sequentially genes having CNVs with the highest AF until we reached the threshold. Thus, we removed *SSR4*, *ABCD1* and *BCAP31* (encompassing the same frequent CNV), *AVPR2*, *STS* and *PHKA2*, obtaining a final list of 75 genes.

Feature Importance, minimum number of samples and target regions

For all features in the XLR dataset, we evaluated their importance using the Permutation Feature Importance approach and Boruta algorithm.

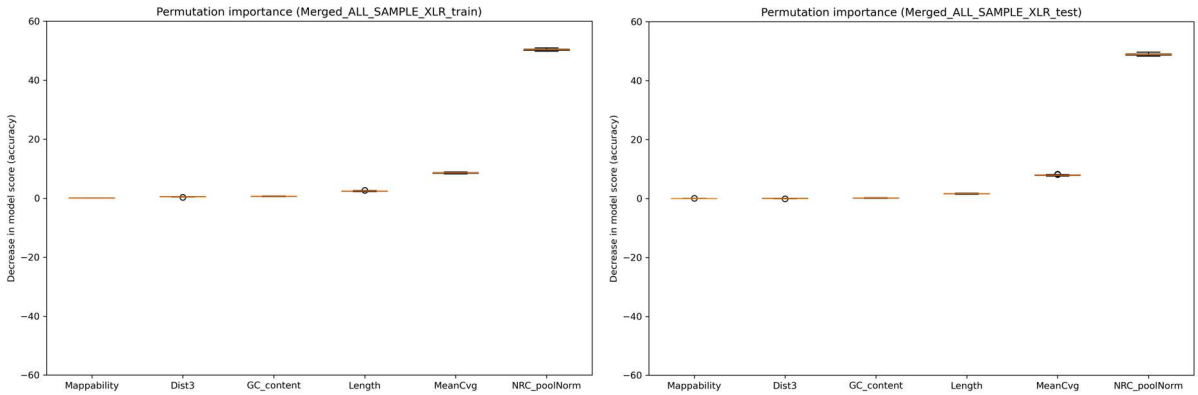


Figure 15 Permutation Feature Importance distribution for all iterations. Results for the training set on the left and on XLR test fraction on the right.

Using the Permutation Feature importance, the highest accuracy loss was obtained when removing NRC_poolNorm (Fig. 14). This feature, together with Length and MeanCvg, are the most important features for the model. The other 3 features (Mappability, Dist3 and GC content) were not considered informative by the procedure. Moreover, the value of importance remained almost constant for all training iterations of the Random Forest.

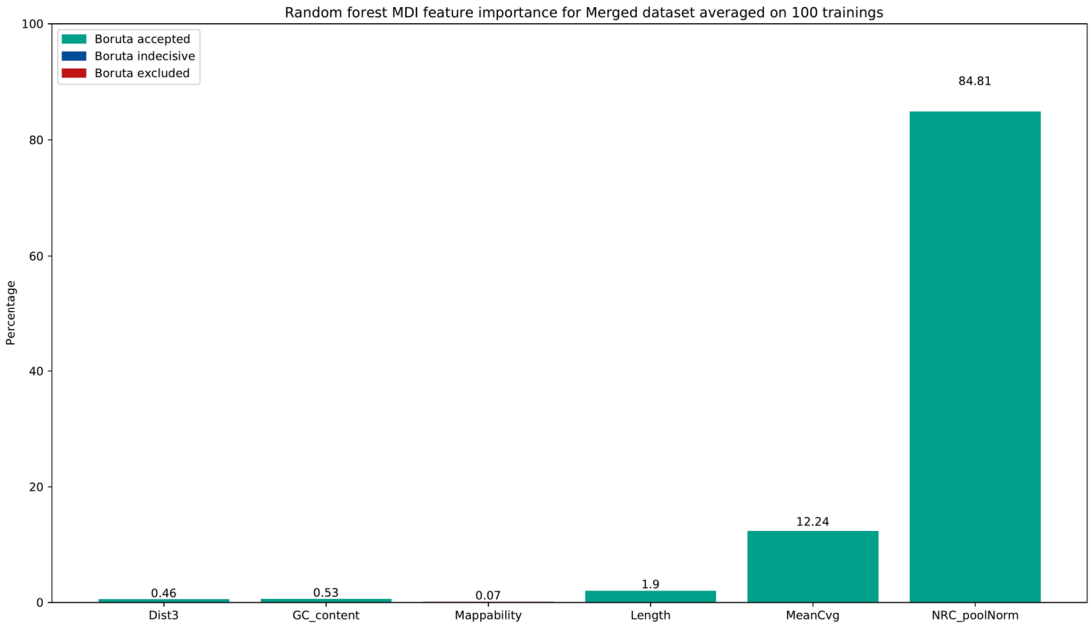


Figure 16 Boruta feature selection with MDI feature importance bar plots. On green, all features accepted by Boruta. No features were considered indecisive.

Results obtained using Boruta and Random Forest MDI are also consistent with PFI (Fig. 15). Using the Boruta method, all features were accepted without being indecisive, apart for Mappability, that was rejected. Moreover, Random Forest importance values agrees with PFI, giving the highest importance to NRC_poolNorm, followed by MeanCvg and Length.

Given these results, we removed Mappability from the feature space. This result was expected, as we saw in the boxplot distribution for all kits a mappability score almost always equal to 1. Since the XLR dataset contains a pruned set of disease-causing genes, it is quite rare to find these genes in low-complexity regions of the genome; thus, using this feature could be deleterious for our model, as the genome contains lots of regions with low mappability.

Finally, we evaluated, using the Merged dataset, the minimum number of samples and target regions required to get a good prediction accuracy.

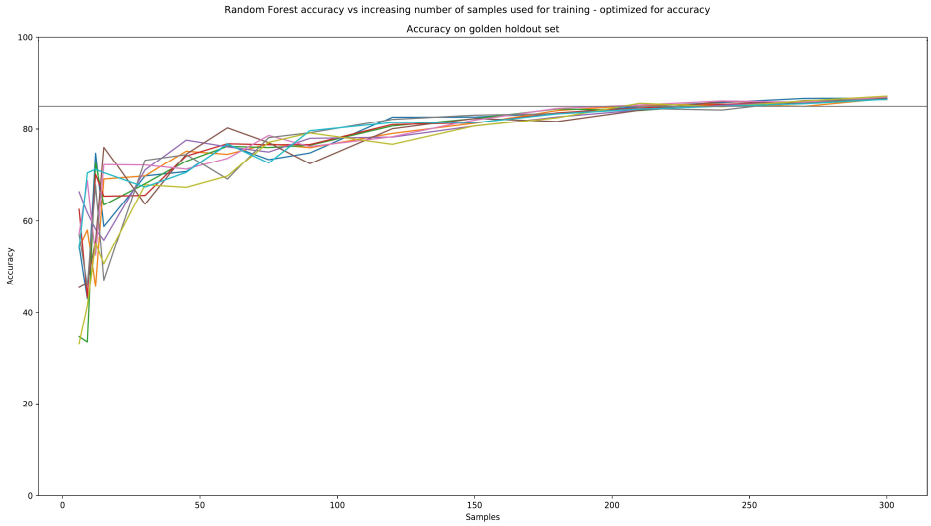


Figure 17 Random Forest accuracy for each subset of individuals, repeated 10 times.

As shown in the plot (Fig. 16), a low number of samples gives a high variability in accuracy for each iteration. This is mostly due to the fact that, by randomly sampling a small fraction of individuals, it is easier to obtain unbalanced datasets. Sampling fraction and accuracy show a positive correlation, increasing at the same time the robustness of the model.

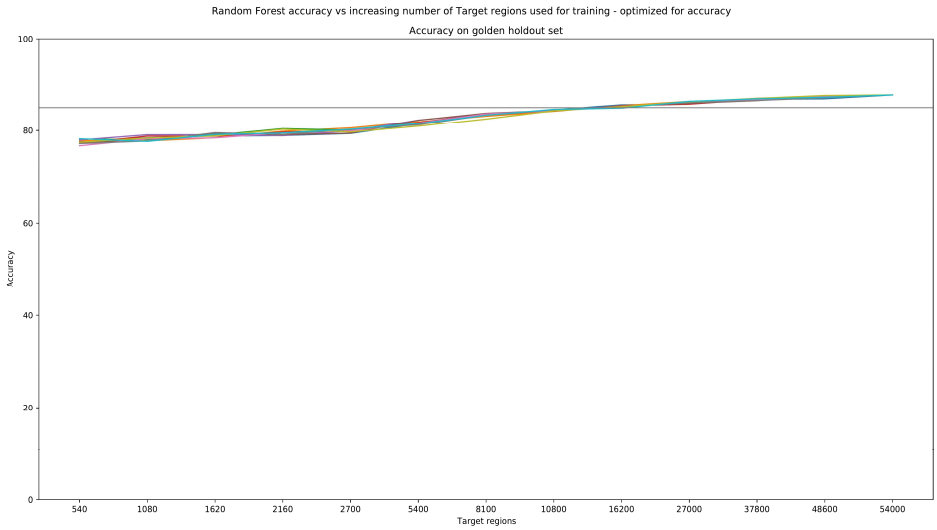


Figure 18 Random Forest accuracy for each subset of target regions, considering all individuals. Test repeated 10 times.

When using the Merged XLR dataset with all its individuals, sampling the number of target regions from a minimum fraction to the entire set does not have the same impact seen for the number of samples. By these results, it is plausible that a very small fraction of target regions is required for training, as long as they come from a large cohort of individuals.

Training results

Selected models and hyperparameters

After 10-fold cross-validation, we obtained for each model tested the best set of hyperparameters.

Model	Best hyperparameters
DecisionTreeClassifier()	criterion : gini max_depth : 10 min_samples_split : 50 min_samples_leaf : 50
RandomForestClassifier()	criterion : gini n_estimators : 100 max_depth : 10 max_features : 3 min_samples_leaf : 100 min_samples_split : 100
SVC()	C : 20 kernel : rbf cache_size : 1000
KNeighborsClassifier()	n_neighbors : 10 weights : distance
MLPClassifier()	hidden_layer_sizes : 8, 8, 8 activation : tanh solver : sgd learning_rate : adaptive learning_rate_init : 1e-4 early_stopping : false n_iter_no_change : 10 tool : 0.0035 max_iter : 200

Table 12 Set of best hyperparameters for each models using GridSearch selection.

When training the autoencoder, we obtained 1.84 mean reconstruction error. Mean AE_{XLR} + 2σ value was used as threshold in our test datasets, to remove all target regions with an AE exceeding this value (Fig. 18).

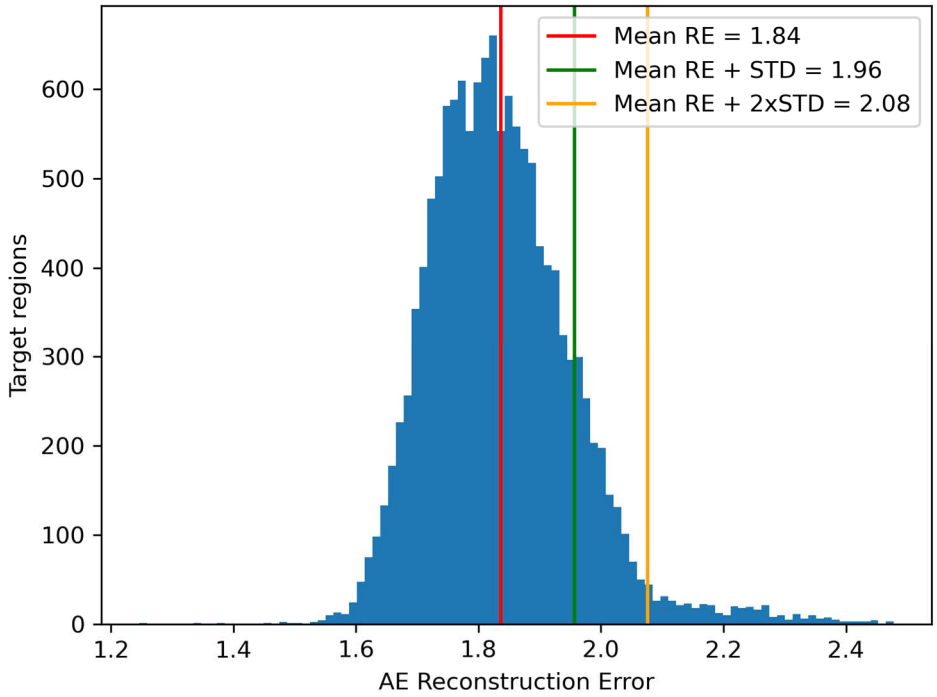


Figure 19 Histogram distribution of the Reconstruction error obtained when training an Autoencoder using the XLR dataset.

Results on the XLR test fraction

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	84.78	85.63	84.78	84.95
Random Forest	82.35	84.18	82.34	82.66
Support Vector Machine	76.31	80.42	76.30	76.88
K-NN	73.97	78.43	73.95	74.59
Multi-Layer Perceptron	72.91	77.88	72.89	73.55

Table 13 Macro-averaged metrics for XLR test fraction.

All trained models obtained a good accuracy and F1 score in making predictions on the XLR test fraction, with Decision Tree obtaining the best scores.

Results on NSD dataset

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	83.44	84.43	83.44	83.63
Random Forest	81.07	82.99	81.07	81.39
Support Vector Machine	74.68	79.15	71.68	75.29
K-NN	70.73	75.85	70.73	71.39
Multi-Layer Perceptron	70.75	76.45	70.75	71.41

Table 14 Macro-averaged metrics for noSegDup test set.

For the NSD dataset we obtained scores similar to the ones obtained in the XLR dataset. Since the training set contains very clean regions compared to this set, the very slightly decrease in performance can be expected.

Results on SD dataset

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	78.53	79.49	78.53	78.74
Random Forest	76.52	78.60	76.52	76.90
Support Vector Machine	70.72	75.44	70.72	71.37
K-NN	65.62	71.14	65.62	66.31
Multi-Layer Perceptron	66.03	72.52	66.03	66.66

Table 15 Macro-averaged metrics for SegDup test set.

This is the set containing more problematic regions. As shown in the table, all models have a drop in performance.

Results on the 5 enrichment kits separately

Finally, we tested the trained models on each kit test set separately.

	Model	DT	RF	SVM	K-NN	MLP
NSD	BGI	93.46	92.92	89.06	81.84	83.74
	MedExome	81.17	74.92	63.00	63.96	60.77
	Nextera	84.78	85.85	79.27	74.66	72.82
	SureSelect	77.00	74.90	72.13	66.47	69.36
	Twist	81.46	77.15	68.40	68.42	67.44
SD	BGI	87.76	87.68	82.56	75.70	75.71
	MedExome	76.07	70.19	61.65	59.29	59.16
	Nextera	83.61	84.65	76.87	71.67	69.18
	SureSelect	70.02	68.92	66.86	61.07	64.78
	Twist	76.19	72.32	64.39	61.98	61.47

Table 16 Macro-averaged F1 scores obtained using the erged dataset on each kit-specific SD and NSD dataset.

Benchmark

1000 Genomes: NS and PS

SureSelect v2 target was first lifted to GRCh38 coordinates, losing 0.15% regions. We proceeded with the workflow obtaining our target reference dataset and NS.

Target	Bait Size (Mb)	Reference Build	Number of regions
Agilent SureSelect All Exon V2	46	hg19	194680
Agilent SureSelect All Exon V2 – after liftover	45.9	GRCh38	193873
Target reference dataset	44	GRCh38	186200
NS	36.5	GRCh38	150634

Table 17 Description of the number of target regions lost at each passage of the workflow

To construct the PS, we first had to make some evaluations for variants derived from DGV's array studies. After intersecting the selected studies, we found a higher intersection between studies with the highest resolution (Campbell et al., 2011, Conrad et al., 2009 and McCarroll et al., 2008).

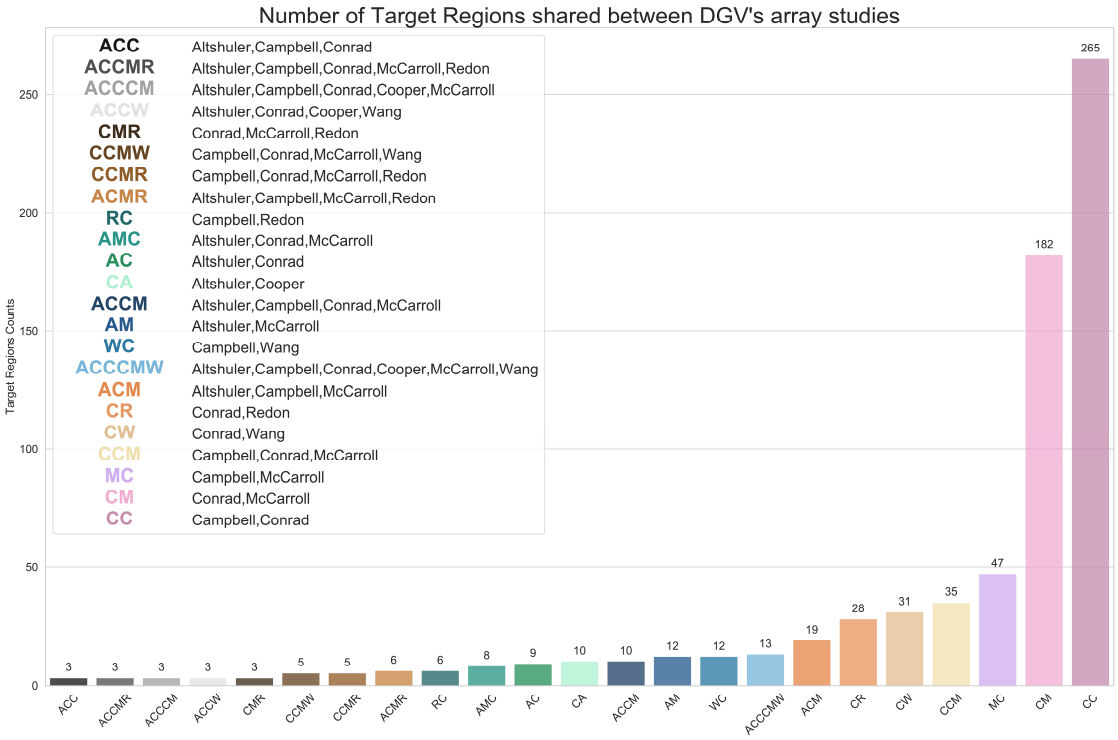


Figure 20 Distribution of target regions containing DGV's CNV shared between all studies.

We removed Cambpell et al., 2011 since it was a custom study and did not target all the genome and Cooper et al., 2008 due to the method of validation (a variant was considered validated if seen from other studies). At the end, we selected the target regions in common between Conrad et al., 2009 and McCarroll et al., 2008 to have a selection of the two main technologies used (Oligo aCGH and SNP-array) with the best resolution, validation method and number of shared regions.

For real CNVs evaluation, we used the native NA12878 BAM file. Meanwhile, for synthetic variants we selected and split dbVar's pathogenic CNVS in 5 lists, obtaining these size-separated number of variants:

dbVar CNVs	1 region	2 – 5 regions	6 – 10 regions	11 – 50 regions	>50 regions
List 1	265/265	229/700	78/593	60/1248	4/281
List 2	45/45	47/144	43/339	24/522	5/492
List 3	25/25	27/81	14/104	15/260	4/291
List 4	15/15	17/63	9/71	9/164	1/65
List 5	47/47	41/114	3/18	1/35	5/430

Table 18 Number of dbVar CNVs for each list. Values reported as: "number of CNVs"/"number of targeted regions within the CNVs"

Each list of variants was added synthetically to NA12878 original file, generating 5 new BAM files. The overall number of variants for each PS is reported below.

NA12878 PS	N Total	N Deletions	N Duplications
Array based dataset	24/232	10/64	14/168
Short reads based dataset	142/1001	77/353	65/648
In Silico dataset – 1 regions	397/397	360/360	37/37
In Silico dataset – 2 to 5 regions	361/1102	312/944	49/158
In Silico dataset – 6 to 10 regions	147/1125	133/1017	14/108
In Silico dataset – 11 to 50 regions	109/2229	93/1869	16/360
In Silico dataset – >50 regions	19/1559	15/1239	4/320

Table 19 Number of CNVs for each dataset - total and stratified for deletions and duplications. Values reported as: "number of CNVs"/"number of targeted regions within the CNVs"

1000 Genomes: TrainX XLR dataset and model selection on autosomes

1000 Genomes samples were added to the merged dataset, affecting mostly coverage and NRC_poolNorm ranges.

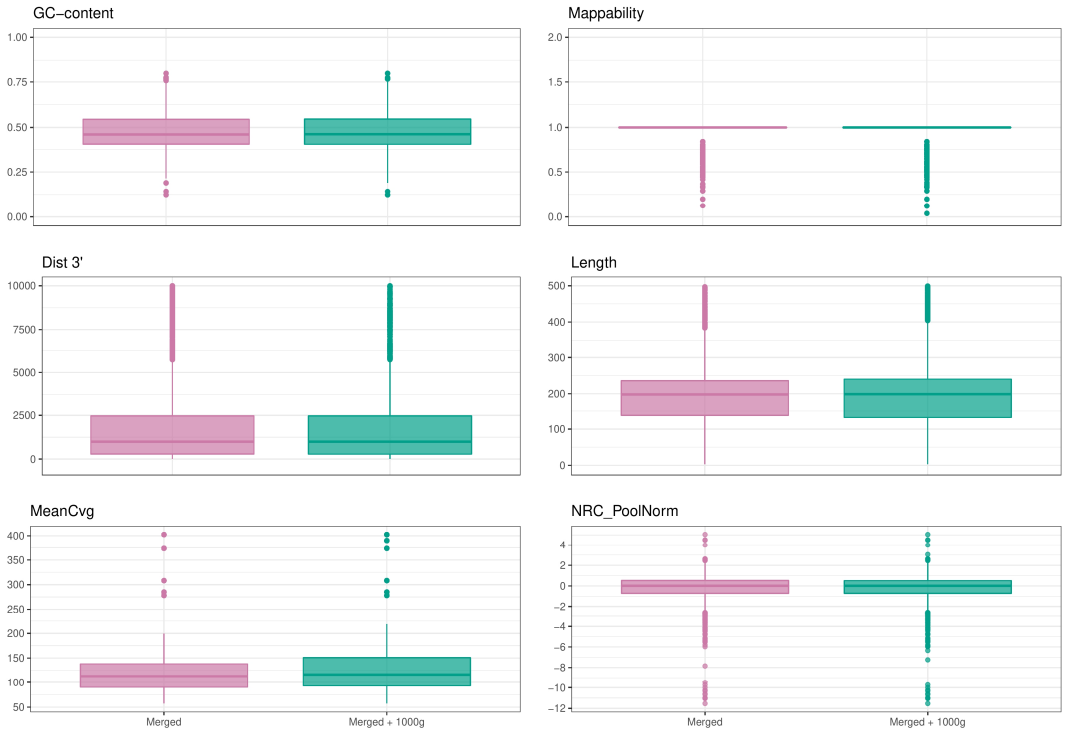


Figure 21 Box-plot distribution of all features in the XLR dataset – initial Merged dataset and Merged dataset with 1000 Genomes samples included.

To select the model to use for CNV calling using HMM, we picked the subset showing good scores on X chromosome test sets (RF, SVC and MLP), tested them on NA12878, and evaluated their performance using NA12878 real PS, computing the overall number of target regions called wild-type and altered, without CNV type separation.

Experiment		Accuracy	BA	TPR	TNR	FPR	FNR
RF	Arrays	87.51	65.98	**44.40**	87.57	12.43	**55.60**
	Short Reads	87.28	58.79	29.97	87.61	12.39	70.03
MLP	Arrays	65.40	52.11	38.79	65.43	34.57	61.21
	Short Reads	65.37	51.00	**36.46**	65.53	34.47	**63.54**
SVC	Arrays	**94.50**	**67.97**	41.38	**94.57**	**5.43**	58.62
	Short Reads	**94.20**	**60.48**	26.37	**94.59**	**5.41**	73.63

Table 20 Performance metrics for RF, MLP and SVC models, trained on the Merged dataset and used to make predictions on NA12878.

We found that, when testing the models on the autosomes, SVC is the one that gets less false positive calls and is better in generalizing on new data. Considering these results, we decided to only use our trained SVC model to make predictions on the autosomes. Despite having obtained best results with RF and DT during training, the number of false positive calls obtained in this test make it clear that those models were partially overfitting the data.

1000 Genomes: Benchmark results

Performance with size separation.

We first considered callers' performance on subsets of our PS, grouped by ranges of target regions included in each CNV. For the synthetic set, results were averaged for the 5 lists created. When using EXCAVATOR2, since in the original paper it is suggested to try different window sizes and choose calls from windows giving the most similar variability between in- and off-target signals, we selected calls made using a 10kb window size. Whereas for TrainX, we tested the performance with and without the autoencoder and filtering calls having an HMM posterior alt probability <70%.

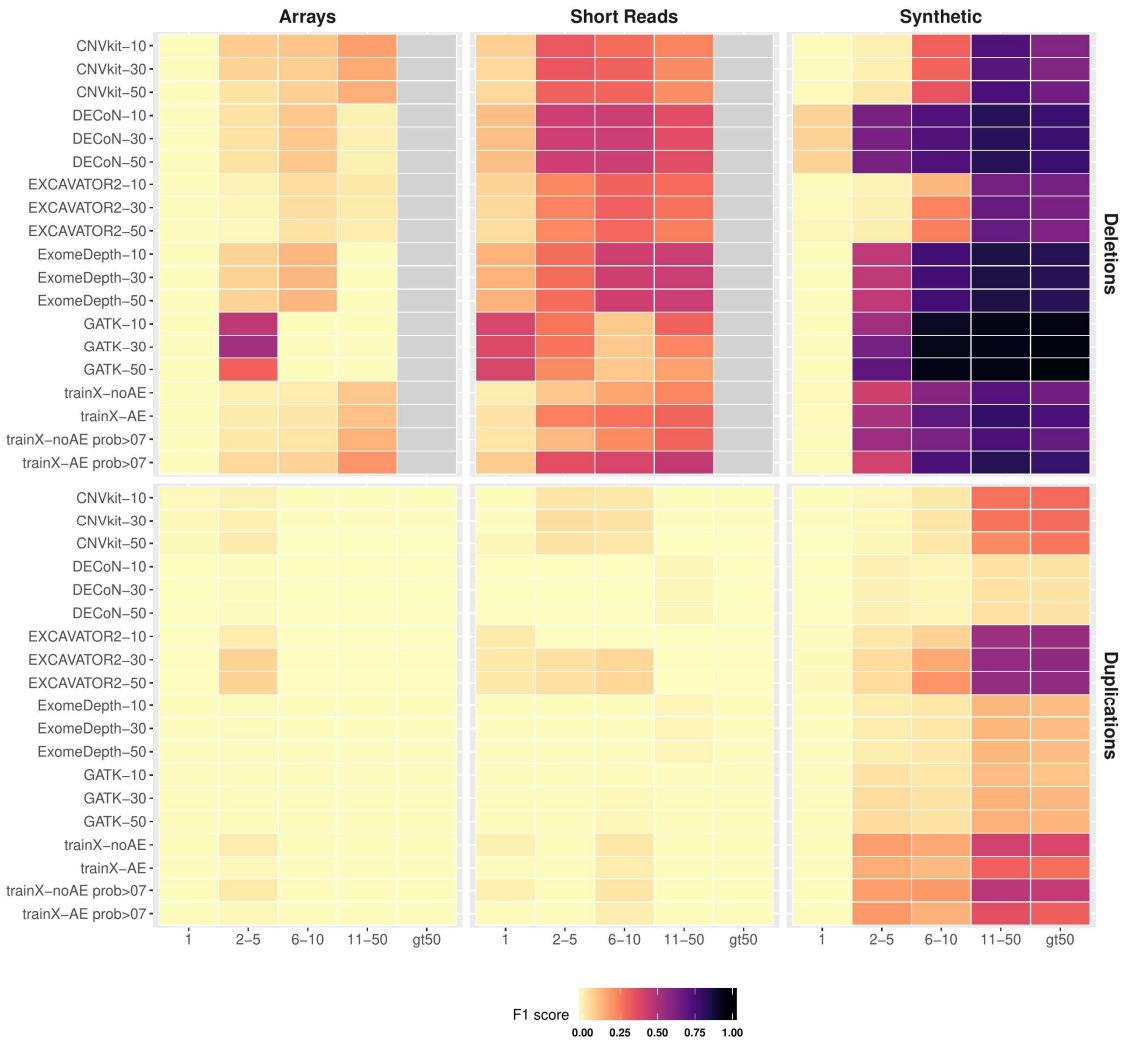


Figure 22 F1- scores heatmap showing CNV callers performance in calling single target regions of the positive sets; stratified by deletions and duplications for each range of sizes. Grey boxes: size ranges that do not contain positive CNVs.

Overall performance.

After considering each caller performance based on specific ranges of sizes, we further evaluated the overall performance of each tool without size distinction.

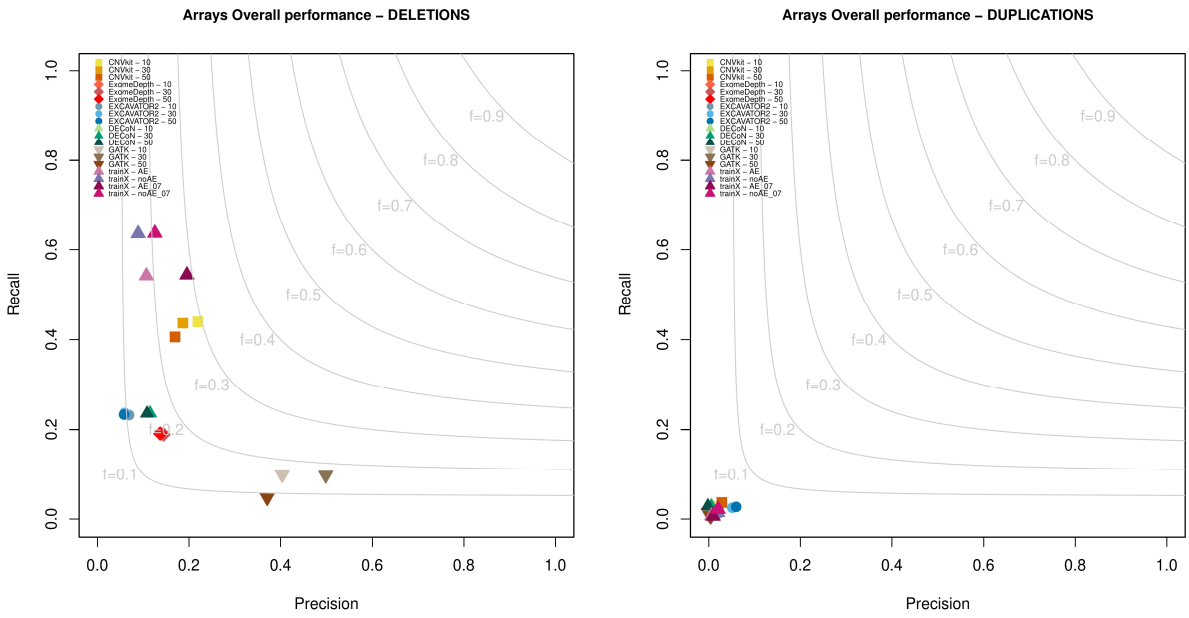


Figure 23 Precision-Recall plot for NA12878 array PS.

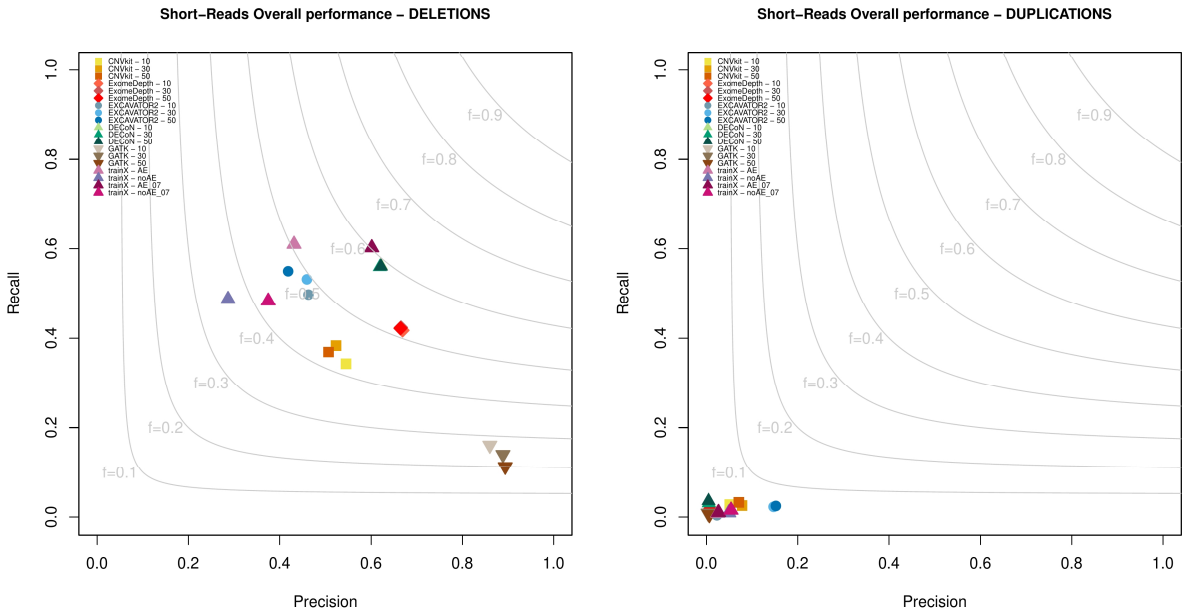


Figure 24 Precision-Recall plot for NA12878 short-reads PS.

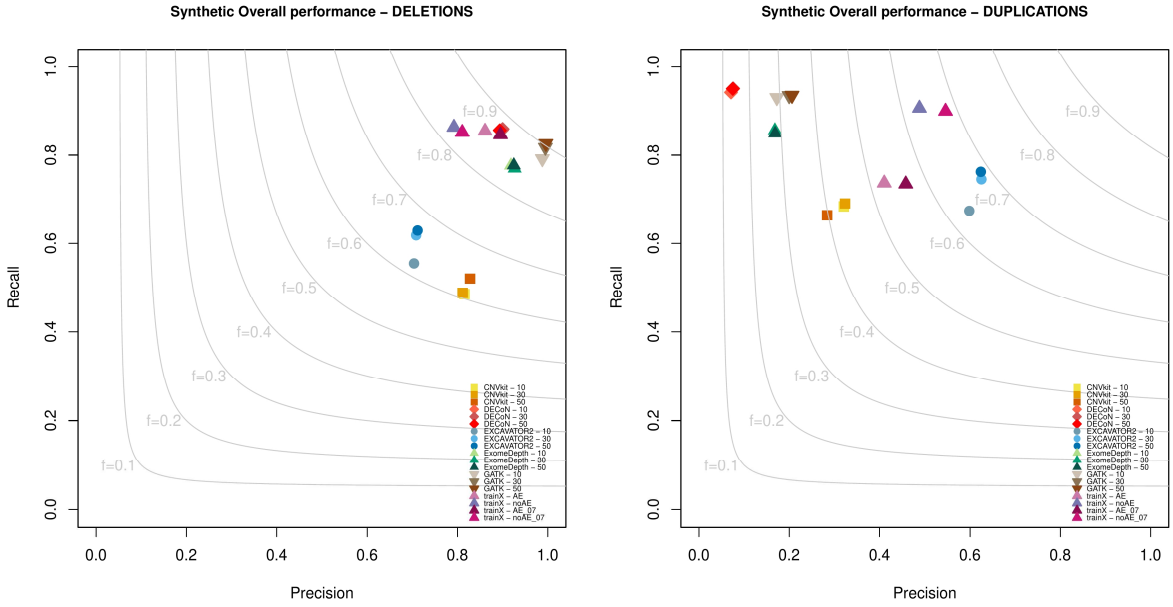


Figure 25 Precision-Recall plot for NA12878 synthetic PS. Results averaged on the 5 lists created.

All callers showed best performances with the short-reads and synthetic datasets; considering that both sets are based on detection on short-reads data, this could explain the better results. Moreover, while for the synthetic dataset we precisely know which CNV is a deletion or a duplication, this is not always true for the DGV datasets, as we saw that some CNVs have discordant CNV types between each study for the same sample, making these results less reliable in terms of recall.

Regarding the number of samples used to create the pool of controls, mostly all callers showed an improvement in performance when using a bigger cohort. ExomeDepth and DECoN however did not follow the same trend, confirming the maximum number of 10 controls suggested in their documentation. With TrainX, a probability threshold helps increasing the precision, removing a part of FP calls. Also, duplications do not seem to benefit the use of an autoencoder to filter outliers, probably because duplications are in general more difficult to model.

Epi25: NS and PS

Nextera Rapid Capture Exomes target was first lifted to GRCh38 coordinates, losing 0.13% regions. As for 1000 genomes, we removed chrX, chrY and alternative contigs to obtain our target reference dataset.

Target	Bait Size (Mb)	Reference Build	Number of regions
Nextera Rapid Capture Exomes V1.2	45.33	hg19	214035
Nextera Rapid Capture Exomes V1.2 – after liftover	45.2	GRCh38	213378
Target reference dataset	43.3	GRCh38	205388

Table 21 Overall number of target regions included at each step

After filtering the outlier samples (N=51) from both SNP-array datasets and TrainX call sets, we obtained a final number of 251 samples analysed.

Epi25 NS and PS are different for all samples, but a general overview of the PS dimensionality is given below.

PS	N patients	N CNVs	N Deletions	N Duplications
High quality set	40	46/1244	25/379	21/865
Medium quality set	10	10/111	8/74	2/37
Low quality set	30	34/137	16/70	17/67

Table 22 Number of samples and variants for each SNP-array derived PS; CNVs reported as: "number of CNVs" / "number of targeted regions within the CNVs".

Epi25: TrainX XLR dataset

Epi25 training samples were added to the merged dataset.The merged dataset contained samples from the 5 enrichment kits and 1000 Genomes samples used before (see Methods: 1000 Genomes: TrainX XLR dataset) .

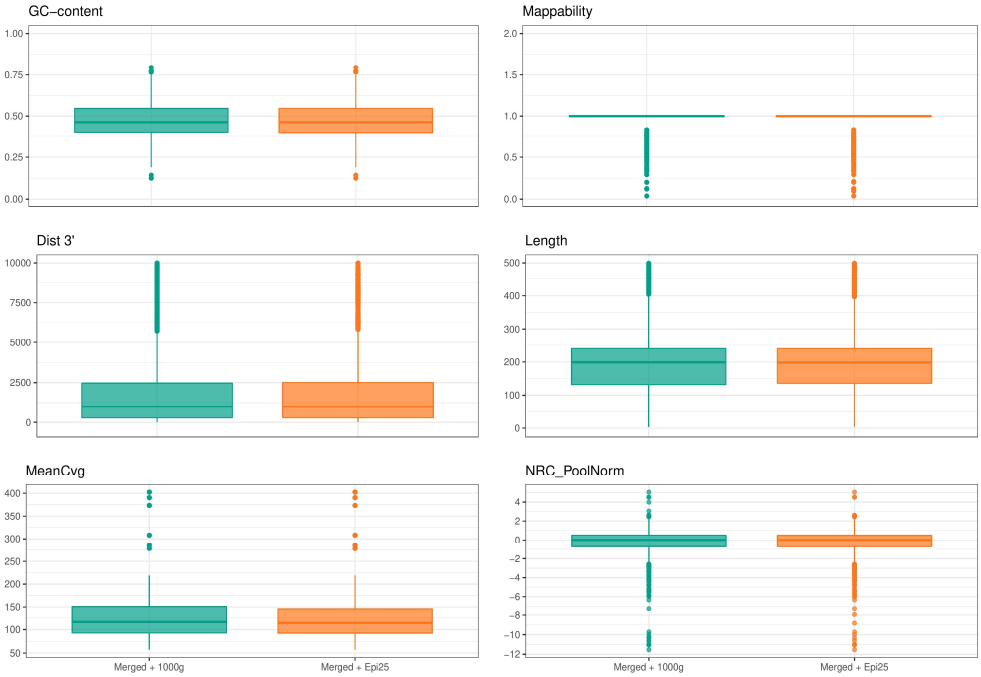
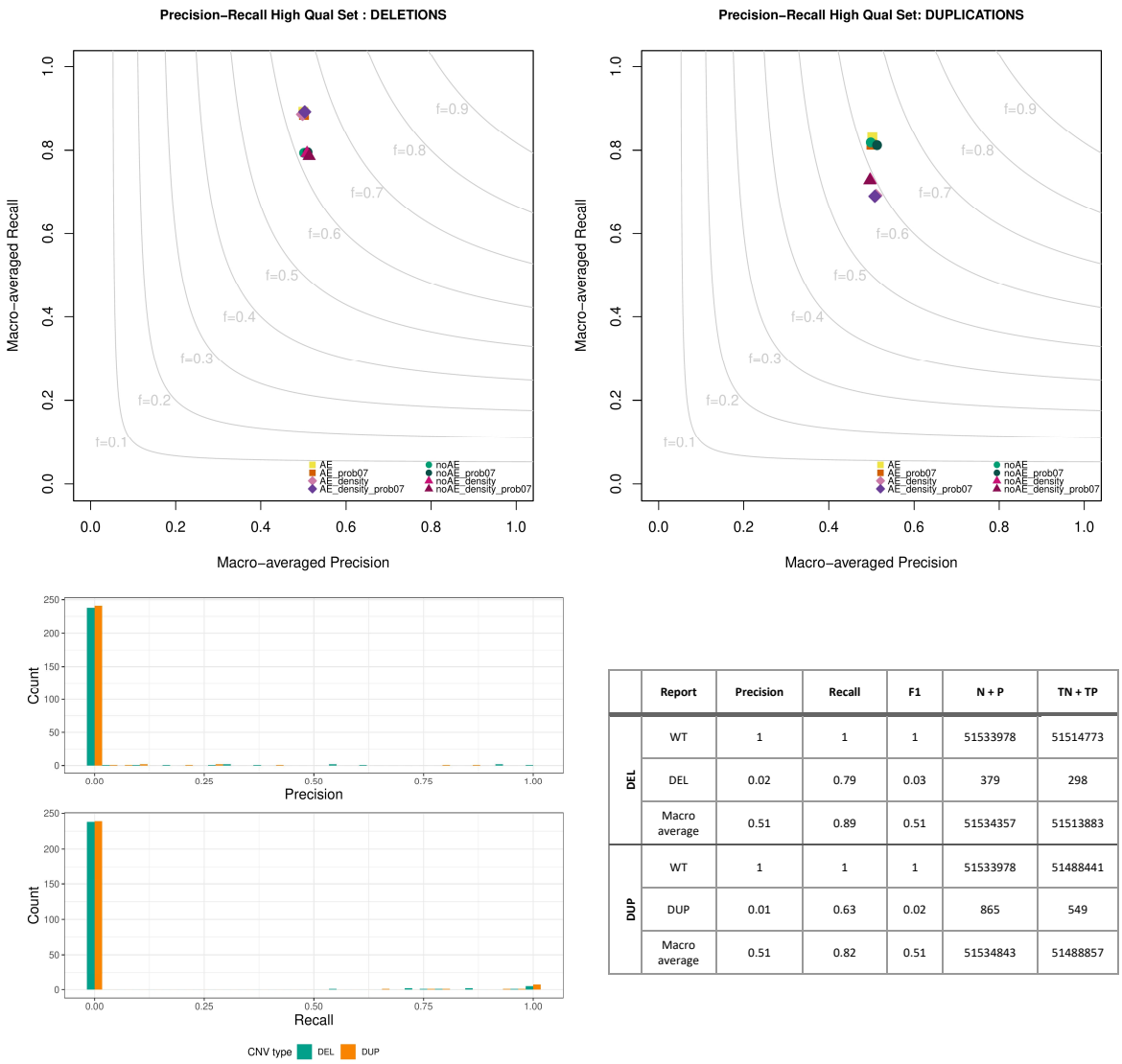


Figure 26 Box-plot distribution of all features in the XLR dataset – Merged dataset with 1000 Genomes samples and Merged dataset with 1000 Genomes and Epi25 samples included.

Adding these samples to the merged dataset mostly affects the distribution of the MeanCvg and NRC_poolNorm, as expected. We then proceeded in training the SVC model and calling CNVs in our samples.

Epi25: Results and validation

For Epi25 samples, we evaluated how many of the High Quality PS CNVs were detected using TrainX. We computed each performance metric for all 251 samples and evaluated the macro-average Precision and Recall resulting after summing all positives, negatives, TP, TN, FP and FN target regions.



	Report	Precision	Recall	F1	N + P	TN + TP
DEL	WT	1	1	1	51533978	51514773
	DEL	0.02	0.79	0.03	379	298
	Macro average	0.51	0.89	0.51	51534357	51513883
DUP	WT	1	1	1	51533978	51488441
	DUP	0.01	0.63	0.02	865	549
	Macro average	0.51	0.82	0.51	51534843	51488857

Figure 27 Precision and Recall for the high-quality positive set (N=251). First plot: showing macro averaged results for each filtered/not-filtered dataset tested. Bar plot on bottom left: Precision and Recall obtained for each sample when using Autoencoder and probability threshold at 70%. Table on bottom right: scikit Classification Report for results with Autoencoder and probability threshold at 70%, showing how the macro average results have been obtained (stratified for CNV type).

As shown in the barplot (Fig. 26), most of the samples showed zero precision and recall, since they did not have any variants in the High Quality PS.

From the original high-quality SNP-array data (Niestroj et al., 2020) we previously identified 2 clinical CNVs (1 deletion and 1 duplication) that were included in the High Quality set used. We confirmed with our method both clinical CNVs (1 1.5MB deletion and 1 10.2Mb duplication). Regarding the medium and low quality set, we confirmed 8 CNVs: 2 deletions and 1 duplication in the medium quality set and 5 deletions in the low quality set. Out of these 8 CNVs, we identified a deletion derived from the low

quality set containing 9 SNPs that encompassed *PKD1* and *TSC2* and perfectly correlated with the clinical phenotype of the patient, consisting in the coexistence of tuberous sclerosis and cystic kidney disorders which was compatible with a contiguous *PKD1/TSC2* deletion syndrome. Notably, in the SNP-array data, this deletion belonged to the low-quality dataset and as such could have been missed, while it was identified with TrainX at a high call probability (0.99).

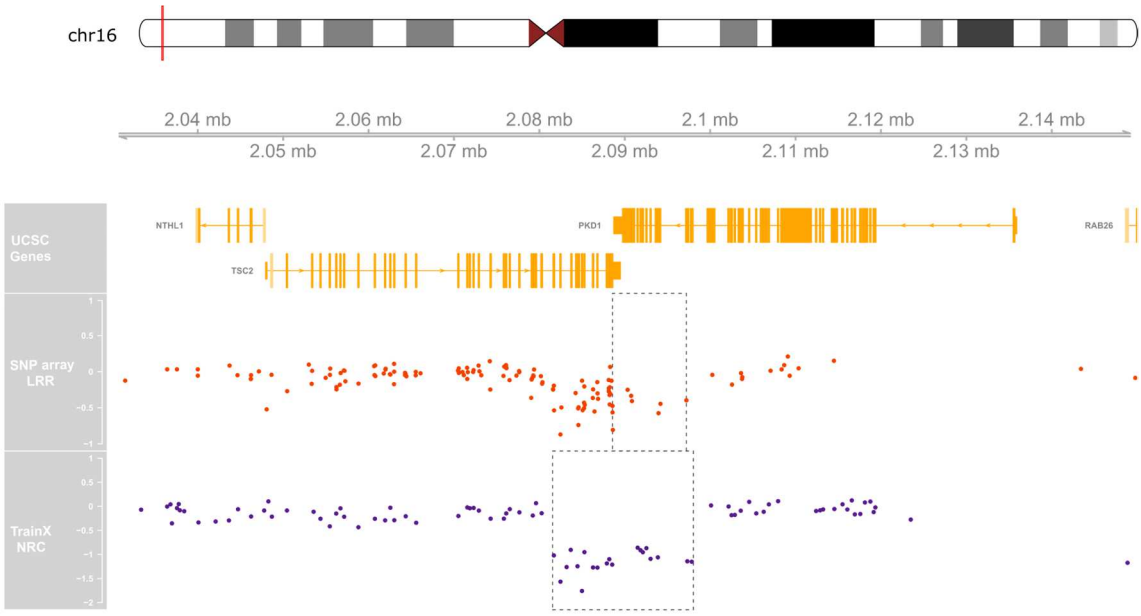


Figure 28 Signal distribution for patient with TSC2-PKD1 deletion. Log2 ratio signal across the region where we detected the deletion for both SNP-array and TrainX. Dashed boxes point to the breakpoints detected by each technique. As shown, SNP-array have a noisy log2ratio signal for this deletion.

We validated and confirmed this heterozygous deletion using MLPA as orthogonal technique, finding the breakpoints at Exon 31 of *TSC2* and Exon 31 of *PKD1.* When looking at the breakpoints of the deletion detected by TrainX, we found the exact extension detected using MLPA.

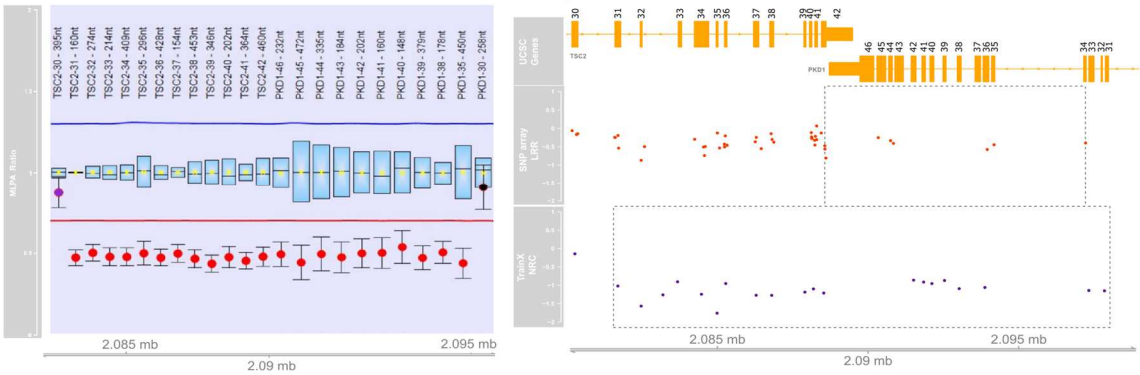


Figure 29 MLPA validation of TSC2-PKD1 deletion. The deletion breakpoints start at exon 31 of TSC2 (Forward Strand) and end at exon 31 of PKD1 (Reverse Strand). On top of the ratios, labels refer to the analysed gene's exon number and length (base pairs). On the right, a zoomed version of Fig. 27, showing the same deleted exons.

Discussion

In this thesis we introduced TrainX, an original machine learning (ML) method for CNV calling in WES data. The most distinctive feature of TrainX is that it exploits the naturally occurring hemizygous and disomic non-pseudoautosomal X chromosome sequences in males and females, respectively, in order to train its underlying model to recognize deletion and duplication states of exonic chromosomal regions. TrainX derives an indicator of genome copy number from in target WES data using Normalized Read Counts (NRC) computed with EXCAVATOR2, our previously developed tool. Apart from using NRC, TrainX methodology is completely new and is the only one that solely relies on a ML classifier to detect genomic imbalances in WES. Although there are few publications that use ML classification methods in NGS data, the majority refers to WGS and mostly rely on integrating results from several variant callers to re-classify CNVs as real or false (Pounraja et al., 2019; Zhuang et al., 2020). The only single comparable tool published to date that use machine learning directly to raw WES data is CNV-RF (Onsongo et al., 2016). This caller however is based on a Read Depth and segmentation approach and use a trained Random Forest classifier as a final step, to predict false positives from the final set of calls obtained. Moreover, what makes TrainX different from other ML tools is the amount of training data available, as the non pseudo-autosomal X chromosomes from either affected or healthy individuals can be used, sidestepping the unbalanced dataset problem that usually affects supervised machine learning trainings when using genomic data.

To create TrainX feature space, we could select only a small number of features, since the XLR dataset contains a good set of regions that have usually high intolerance to biological variability. However, we found out that 2 of the selected features, namely NRC_poolNorm and MeanCvg are sufficiently strong predictors. When plotting the relationship between these 2 features, it is evident that there is no clear linear separation between the 3 classes. Given the complexity of the data, we tried to reduce overfitting by selecting a smaller training fraction and adding gaussian noise to the dataset. Moreover, we also trained to the same training fraction an autoencoder to remove from the test sets predicted target regions gauged as outliers. Taking all these

improvements together, we obtained good performance in both chromosome X and autosomes when using a Support Vector Machine Classifier. The reduced number of false positives could be explained by the use of the RBF kernel, that probably finds a more regular separating hyperplane compared to the other models tested.

To evaluate TrainX ability of calling regions really affected by a CNV and compare its performance to that of other tools, we carried out an extensive benchmark focused on the minimum number of target regions required by each tool to correctly identify true heterozygous deletions and duplications. What we observed, when evaluating the synthetic CNVs, is that each tool has different behaviours when considering the two CNV types separately. For 1 copy deletions >1 target region, GATK4 gCNV obtained the best F1 scores. A similar performance was also seen in a recent publication (Testard et al., 2021), where they run GATK4 in a retrospective cohort and detected all the clinical CNVs previously identified with CMA, including 1 single exon deletion and 1 exon duplication (CN=0 and CN=3 respectively). However, in this study the authors did not consider precision and recall, nor any other quantitative measure, as metrics to evaluate performance of the chosen tools and did not operate any distinction between CNV types, making the present benchmark a helpful addition to what has been already published for whoever is in search of studies that compare WES-focused CNV callers.

Among the tested tools, DECoN is the only caller able to consistently detect single target region deletions, confirming the purpose of its development. However, it still shows very low recall and precision, and it is important to underscore that this very low performance may be related to the fact that this caller was primarily focused on gene panel analysis, which is usually limited in the target extent, and therefore is less impacted by a poor precision performance.

When considering 3-copies duplications, all tools that showed an overall good performance in detecting deletions demonstrate a considerably drop in performance, while EXCAVATOR2, which did not outstand in detection of 1-copy deletions, showed the best scores for duplications, especially for those larger than 11 target regions.

Comparing TrainX performance against all these tools, there is an evident difference in terms of stability across CNV types. TrainX is the only tool showing similar accuracy and

score stability for all calls including >1 target region, making the evaluation of the entire spectrum of CNV types and sizes more feasible using a single approach. Considering that diagnostic CNVs range from intragenic to large chromosomal events, this property of TrainX could prove to be really valuable in the clinical setting, where having a single tool instead of being forced to combine different solutions is an important practical advance. Moreover, all callers having good scores with deletions and a drastic drop for duplication (i.e., GATK4, DECoN and ExomeDepth) rely mostly on statistical models for CNV detection, suggesting that it is more difficult to statistically model duplications compared to deletions. Using a ML approach seems to consistently overcome this problem. When also taking into account the overall performance while benchmarking against the array and short reads positive sets, TrainX always show the best scores. In general, none of these tools could detect single target regions events, suggesting that 2-3 target regions could be a reasonable limit for CNV detection using any of the tested tools.

TrainX performance stability across CNV types have been confirmed using the Epi25 dataset. With an average recall of 80%, we were able to identify all previously detected diagnostic CNVs, and we were also able to identify a previously unnoticed pathogenic deletion in one of the patients; following best practices adopted for microarrays, this CNV was not previously considered since it included very few SNPs and was filtered out in the first steps of QC.

Nevertheless, from the results we collected several considerations on fixes and improvements that need to be tested.

The main limitation of TrainX rely on its inner nature, namely the fact we build our algorithm around the X chromosome sequences. Since we use male samples as model for deletions, it is not possible to call on X chromosome for them, as it would be automatically classified all as deleted except for pseudo-autosomal regions. Thus, further evaluations need to be done to try weighting the results when considering X chromosome in male samples. Also, since duplications are modelled less efficiently than deletions as, for example, for 1000Genomes benchmark we saw a better performance without the autoencoder, a good idea could be to create merged samples with a collection of merged BAM files from both males + females and females +

females, in order to enhance the distribution of the duplication events modelled for training.

When re-evaluating the feature distribution of our merged datasets, taking also in consideration the results, we highlighted two main aspects that need to be reconsidered. First of all, when creating the merged datasets we select, for all individuals considered, a random number of samples. Even if the number of samples for individuals should be more or less similar, a new weighted sampling approach would be considered. Secondly, we observed that individuals having a MeanCvg falling outside the whiskers of the merged distribution show an elevated number of false positive calls. Looking at Figure 12, samples with high MeanCvg have a different NRC_poolNorm distribution, more shifted to the right side of the tail, compared to the other set of samples. Since, as for now, the merged datasets include few samples with a very high coverage, these samples are just outliers and do not add value to help the model to learn their characteristics. Considering this limitation, we are planning to use an additional autoencoder to detect and remove individuals added to the analysis that have features too divergent from those included in the merged datasets. As the idea behind the merged dataset is to include new samples for each new batch to analyse, this behaviour should be limited over time; considering that WES cost is constantly dropping, high coverage sequencing is becoming the standard procedure in every clinical and research setting.

Finally, we recognize that calling and/or post-calling steps should be improved still to increase accuracy and in particular to mitigate the presence of false positive signals. As an example, we observed that some of the calls made are very large (even up to 1-2 Mb) but contain few sparse target regions; We are currently testing a way of filtering these regions out, based on the low density of observation points within the call, as this is the likely mark of a false positive. The approach used in the Epi25 benchmark, filtering out calls having a very low density, does not seem to give a huge improvement in terms of performance. An alternative to this method could be using an HMM algorithm that also considers the distance between target regions, as for now we have assumed equal distance between exons. As a general next goal, we are seeking all the potential improvements to the method that could further raise its precision and

therefore make it an asset for the clinical context, representing a more reliable and wide-scope CNV detection method than current tools.

In conclusion, WES is going to be adopted as first-tier genetic test in a vast range of clinics. CNVs represent one of the most important sources of clinical genomic variation. Providing the genomic diagnostic environment with a CNV detection approach that proves to be reliable to the entire range of coding CNV types and sizes is essential, but currently available tools still suffer from evident biases that prevent their ready use along the clinical practice. Our TrainX method, introducing an original ML solution in this field, establishes itself as a way to reduce this gap towards higher accuracy and wider applicability.

References

Arts P, Simons A, AlZahrani MS, Yilmaz E, AlIdrissi E, van Aerde KJ, Alenezi N, AlGhamdi HA, AlJubab HA, Al-Hussaini AA, AlManjomi F, Alsaad AB, Alsaleem B, Andijani AA, Asery A, Ballourah W, Bleeker-Rovers CP, van Deuren M, van der Flier M, Gerkes EH, Gilissen C, Habazi MK, Hehir-Kwa JY, Henriet SS, Hoppenreijs EP, Hortillosa S, Kerkhofs CH, Keski-Filppula R, Lelieveld SH, Lone K, MacKenzie MA, Mensenkamp AR, Moilanen J, Nelen M, Ten Oever J, Potjewijd J, van Paassen P, Schuurs-Hoeijmakers JHM, Simon A, Stokowy T, van de Vorst M, Vreeburg M, Wagner A, van Well GTJ, Zafeiropoulou D, Zonneveld-Huijssoon E, Veltman JA, van Zelst-Stams WAG, Faqeih EA, van de Veerdonk FL, Netea MG, Hoischen A. Exome sequencing in routine diagnostics: a generic test for 254 patients with primary immunodeficiencies. Genome Med. 2019 Jun 17;11(1):38. doi: 10.1186/s13073-019-0649-3.

Babadi M, Lee SK, Smirnov AN. gCNV: accurate germline copy-number variant discovery from sequencing read-depth data (published in 2019 on https://github.com/broadinstitute/).

Balachandran P, Beck CR. Structural variant identification and characterization. Chromosome Res. 2020 Mar;28(1):31-47. doi: 10.1007/s10577-019-09623-z.

Biederman B, Bowen P. Balanced translocations involving chromosome 12: report of a case and possible evidence for position effect. Ann Genet. 1976 Dec;19(4):257-60.

Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle E, Hamosh A, Shendure J, Lupski JR, Lifton RP, Valle D, Nickerson DA; Centers for Mendelian Genomics, Bamshad MJ. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015 Aug 6;97(2):199-215. doi: 10.1016/j.ajhg.2015.06.009.

D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res. 2016 Nov 16;44(20):e154. doi: 10.1093/nar/gkw695.

Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011 Dec 15;27(24):3423-4. doi: 10.1093/bioinformatics/btr539.

English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014 Jun 10;15:180. doi: 10.1186/1471-2105-15-180.

Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, Uddin I, Wylie H, Strydom A, Lunter G, Rahman N. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016 Nov 25;1:20. doi: 10.12688/wellcomeopenres.10069.1.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, Hultman CM, Sklar P, Purcell SM. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012 Oct 5;91(4):597-607. doi: 10.1016/j.ajhg.2012.08.005.

Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, Muller J. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018 Oct 15;34(20):3572-3574. doi: 10.1093/bioinformatics/bty304.

Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012 May;20(5):490-7. doi: 10.1038/ejhg.2011.258.

Gordeeva V, Sharova E, Babalyan K, Sultanov R, Govorun VM, Arapidi G. Benchmarking germline CNV calling tools from exome sequencing data. Sci Rep. 2021 Jul 13;11(1):14416. doi: 10.1038/s41598-021-93878-2.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009 Aug;10(8):551-64. doi: 10.1038/nrg2593. PMID: 19597530; PMCID: PMC2864001.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, Renkens I, Coe BP, Deelen P, de Ligt J, Lameijer EW, van Dijk F, Hormozdiari F; Genome of the Netherlands Consortium, Uitterlinden AG, van Duijn CM, Eichler EE, de Bakker PI, Swertz MA, Wijmenga C, van Ommen GB, Slagboom PE, Boomsma DI, Schönhuth A, Ye K, Guryev V. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. Nat Commun. 2016 Oct 6;7:12989. doi: 10.1038/ncomms12989.

Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020 Mar;21(3):171-189. doi: 10.1038/s41576-019-0180-9.

Kadalayil L, Rafiq S, Rose-Zerilli MJ, Pengelly RJ, Parker H, Oscier D, Strefford JC, Tapper WJ, Gibson J, Ennis S, Collins A. Exome sequence read depth methods for identifying copy number changes. Brief Bioinform. 2015 May;16(3):380-92. doi: 10.1093/bib/bbu027.

Klau J, Abou Jamra R, Radtke M, Oppermann H, Lemke JR, Beblo S, Popp B. Exome first approach to reduce diagnostic costs and time - retrospective analysis of 111 individuals with rare neurodevelopmental disorders. Eur J Hum Genet. 2022 Jan;30(1):117-125. doi: 10.1038/s41431-021-00981-z.

Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020 Oct 26;12(1):91. doi: 10.1186/s13073-020-00791-w.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019 Jun 3;20(1):117. doi: 10.1186/s13059-019-1720-5.

Kursa, M.B. and Rudnicki, W.R. 2010. Feature Selection with the Boruta Package. Journal of Statistical Software. 36, 11 (Sep. 2010), 1–13. doi:https://doi.org/10.18637/jss.v036.i11.

Lassmann T, Francis RW, Weeks A, Tang D, Jamieson SE, Broley S, Dawkins HJS, Dreyer L, Goldblatt J, Groza T, Kamien B, Kiraly-Borri C, McKenzie F, Murphy L, Pachter N, Pathak G, Poulton C, Samanek A, Skoss R, Slee J, Townshend S, Ward M, Baynam GS, Blackwell JM. A flexible computational pipeline for research analyses of unsolved clinical exome cases. NPJ Genom Med. 2020 Dec 10;5(1):54. doi: 10.1038/s41525-020-00161-w.

Le Scouarnec S, Gribble SM. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. Heredity (Edinb). 2012 Jan;108(1):75-85. doi: 10.1038/hdy.2011.100.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91. doi: 10.1038/nature19057.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

Lupski JR, Wise CA, Kuwano A, Pentao L, Parke JT, Glaze DG, Ledbetter DH, Greenberg F, Patel PI. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. Nat Genet. 1992 Apr;1(1):29-33. doi: 10.1038/ng0492-29.

Lupski JR. Genomic rearrangements and sporadic disease. Nat Genet. 2007 Jul;39(7 Suppl):S43-7. doi: 10.1038/ng2084.

Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for DNA copy number variants detection. Bioinformatics. 2012 Feb 15;28(4):470-8. doi: 10.1093/bioinformatics/btr707.

Marchuk DS, Crooks K, Strande N, Kaiser-Rogers K, Milko LV, Brandt A, Arreola A, Tilley CR, Bizon C, Vora NL, Wilhelmsen KC, Evans JP, Berg JS. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. PLoS One. 2018 Dec 17;13(12):e0209185. doi: 10.1371/journal.pone.0209185.

Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, Gujral M, Brandler WM, Malhotra D, Wang Z, Fajarado KVF, Maile MS, Ripke S, Agartz I, Albus M, Alexander M, Amin F, Atkins J, Bacanu SA, Belliveau RA Jr, Bergen SE, Bertalan M, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Bulik-Sullivan B, Byerley W, Cahn W, Cai G, Cairns MJ, Campion D, Cantor RM, Carr VJ, Carrera N, Catts SV, Chambert KD, Cheng W, Cloninger CR, Cohen D, Cormican P, Craddock N, Crespo-Facorro B, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, Del Favero J, DeLisi LE, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farh KH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedman JI, Forstner AJ, Fromer M, Genovese G, Georgieva L, Gershon ES, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Gratten J, de Haan L, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Huang H, Ikeda M, Joa I, Kähler AK, Kahn RS, Kalaydjieva L, Karjalainen J, Kavanagh D, Keller MC, Kelly BJ, Kennedy JL, Kim Y, Knowles JA, Konte B, Laurent C, Lee P, Lee SH, Legge SE, Lerer B, Levy DL, Liang KY, Lieberman J, Lönnqvist J, Loughland CM, Magnusson PKE, Maher BS, Maier W, Mallet J, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Müller-Myhsok B, Murphy KC, Murray RM, Myin-Germeys I, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, Oh SY, Olincy A, Olsen L, O'Neill FA, Van Os J, Pantelis C,

Papadimitriou GN, Parkhomenko E, Pato MT, Paunio T; Psychosis Endophenotypes International Consortium, Perkins DO, Pers TH, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Savitz A, Schall U, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Silverman JM, Smoller JW, Söderman E, Spencer CCA, Stahl EA, Strengman E, Strohmaier J, Stroup TS, Suvisaari J, Svrakic DM, Szatkiewicz JP, Thirumalai S, Tooney PA, Veijola J, Visscher PM, Waddington J, Walsh D, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wormley BK, Wray NR, Wu JQ, Zai CC, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Cichon S, Collier DA, Corvin A, Daly MJ, Darvasi A, Domenici E, Esko T, Gejman PV, Gill M, Gurling H, Hultman CM, Iwata N, Jablensky AV, Jönsson EG, Kendler KS, Kirov G, Knight J, Levinson DF, Li QS, McCarroll SA, McQuillin A, Moran JL, Mowry BJ, Nöthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sklar P, St Clair D, Walters JTR, Werge T, Sullivan PF, O'Donovan MC, Scherer SW, Neale BM, Sebat J; CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017 Jan;49(1):27-35. doi: 10.1038/ng.3725.

Marshall CR, Bick D, Belmont JW, Taylor SL, Ashley E, Dimmock D, Jobanputra V, Kearney HM, Kulkarni S, Rehm H; Medical Genome Initiative. The Medical Genome Initiative: moving whole-genome sequencing for rare disease diagnosis to the clinic. Genome Med. 2020 May 27;12(1):48. doi: 10.1186/s13073-020-00748-z.

Mergnac JP, Wiedemann A, Chery C, Ravel JM, Namour F, Guéant JL, Feillet F, Oussalah A. Diagnostic yield of clinical exome sequencing as a first-tier genetic test for the diagnosis of genetic disorders in pediatric patients: results from a referral center study. Hum Genet. 2021 Sep 8. doi: 10.1007/s00439-021-02358-0.

Moreno-Cabrera JM, Del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, Serra E, Capellà G, Lázaro C, Gel B. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur J Hum Genet. 2020 Dec;28(12):1645-1655. doi: 10.1038/s41431-020-0675-z.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012 Jul 15;28(14):1919-20. doi: 10.1093/bioinformatics/bts277. Epub 2012 May 9. PMID: 22576172; PMCID: PMC3389768.

Niestroj LM, Perez-Palma E, Howrigan DP, Zhou Y, Cheng F, Saarentaus E, Nürnberg P, Stevelink R, Daly MJ, Palotie A, Lal D; Epi25 Collaborative. Epilepsy subtype-specific copy number burden observed in a genome-wide study of 17 458 subjects. Brain. 2020 Jul 1;143(7):2106-2118. doi: 10.1093/brain/awaa171.

Onsongo G, Baughn LB, Bower M, Henzler C, Schomaker M, Silverstein KA, Thyagarajan B. CNV-RF Is a Random Forest-Based Copy Number Variation Detection Method Using Next-Generation Sequencing. J Mol Diagn. 2016 Nov;18(6):872-881. doi: 10.1016/j.jmoldx.2016.07.001.

Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018 Mar 1;34(5):867-868. doi: 10.1093/bioinformatics/btx699.

Pellegrino E, Jacques C, Beaufils N, Nanni I, Carlioz A, Metellus P, Ouafik L. Machine learning random forest for predicting oncosomatic variant NGS analysis. Sci Rep. 2021 Nov 8;11(1):21820. doi: 10.1038/s41598-021-01253-y.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, Nejentsev S. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012 Nov 1;28(21):2747-54. doi: 10.1093/bioinformatics/bts526.

Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019 Jul;29(7):1134-1143. doi: 10.1101/gr.245928.118.

Quentin Testard, Xavier Vanhoye, Kevin Yauy, Marie-Emmanuelle Naud, Gaelle Vieville, Francis Rousseau, Benjamin Dauriat, Valentine Marquet, Sylvie Bourthoumieu, David

Genevieve, Vincent Gatinois, Constance Wells, Marjolaine Willems, Christine Coubes, Lucile Pinson, Rodolphe Dard, Aude Tessier, Bérénice Hervé, François Vialard, Ines Harzallah, Renaud Touraine, Benjamin Cogné, Wallid Deb, Thomas Besnard, OIivier Pichon, Béatrice Laudier, Laurent Mesnard, Alice Doreille, Tiffany Busa, Chantal Missirian, Véronique Satre, Charles Coutton, Tristan Celse, Radu Harbuz, Laure Raymond, Jean-François Taly, Julien Thevenon. Exome sequencing as a first-tier test for copy number variant detection: retrospective evaluation and prospective screening in 2418 cases. medRxiv 2021.10.14.21264732; doi: https://doi.org/10.1101/2021.10.14.21264732

Rajagopalan R, Murrell JR, Luo M, Conlin LK. A highly sensitive and specific workflow for detecting rare copy-number variants from exome sequencing data. Genome Med. 2020 Jan 30;12(1):14. doi: 10.1186/s13073-020-0712-0.

Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, Vertino-Bell A, Smaoui N, Neidich J, Monaghan KG, McKnight D, Bai R, Suchy S, Friedman B, Tahiliani J, Pineda-Alvarez D, Richard G, Brandt T, Haverfield E, Chung WK, Bale S. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016 Jul;18(7):696-704. doi: 10.1038/gim.2015.148.

Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. Mutat Res Rev Mutat Res. 2019 Jan-Mar;779:114-125. doi: 10.1016/j.mrrev.2019.02.005.

Semeraro R, Orlandini V, Magi A. Xome-Blender: A novel cancer genome simulator. PLoS One. 2018 Apr 5;13(4):e0194472. doi: 10.1371/journal.pone.0194472.

Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, Jäger M, Hochheiser H, Washington NL, McMurry JA, Haendel MA, Mungall CJ, Lewis SE, Groza T, Valentini G, Robinson PN. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016 Sep 1;99(3):595-606. doi: 10.1016/j.ajhg.2016.07.005.

Speicher MR, Carter NP. The new cytogenetics: blurring the boundaries with molecular biology. Nat Rev Genet. 2005 Oct;6(10):782-92. doi: 10.1038/nrg1692.

Spielmann M, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018 Jul;19(7):453-467. doi: 10.1038/s41576-018-0007-0.

Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55. doi: 10.1146/annurev-med-100708-204735.

Stranneheim H, Lagerstedt-Robinson K, Magnusson M, Kvarnung M, Nilsson D, Lesko N, Engvall M, Anderlid BM, Arnell H, Johansson CB, Barbaro M, Björck E, Bruhn H, Eisfeldt J, Freyer C, Grigelioniene G, Gustavsson P, Hammarsjö A, Hellström-Pigg M, Iwarsson E, Jemt A, Laaksonen M, Enoksson SL, Malmgren H, Naess K, Nordenskjöld M, Oscarson M, Pettersson M, Rasi C, Rosenbaum A, Sahlin E, Sardh E, Stödberg T, Tesi B, Tham E, Thonberg H, Töhönen V, von Döbeln U, Vassiliou D, Vonlanthen S, Wikström AC, Wincent J, Winqvist O, Wredenberg A, Ygberg S, Zetterström RH, Marits P, Soller MJ, Nordgren A, Wirta V, Lindstrand A, Wedell A. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. Genome Med. 2021 Mar 17;13(1):40. doi: 10.1186/s13073-021-00855-5.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stütz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA; 1000 Genomes Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81. doi: 10.1038/nature15394.

Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016 Apr 21;12(4):e1004873. doi: 10.1371/journal.pcbi.1004873.

Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, Jiang Q, Allen AS, Zhu M. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat. 2014 Jul;35(7):899-907. doi: 10.1002/humu.22537.

Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol. 2015 Jun 25;3:92. doi: 10.3389/fbioe.2015.00092.

Truty R, Paul J, Kennemer M, Lincoln SE, Olivares E, Nussbaum RL, Aradhya S. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet Med. 2019 Jan;21(1):114-123. doi: 10.1038/s41436-018-0033-5.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007 Nov;17(11):1665-74. doi: 10.1101/gr.6861907.

Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015 Mar;16(3):172-83. doi: 10.1038/nrg3871.

Zhang JX, Yordanov B, Gaunt A, Wang MX, Dai P, Chen YJ, Zhang K, Fang JZ, Dalchau N, Li J, Phillips A, Zhang DY. A deep learning model for predicting next-generation sequencing depth from DNA sequence. Nat Commun. 2021 Jul 19;12(1):4387. doi: 10.1038/s41467-021-24497-8.

Zhang L, Zhou X, Weng Z, Sidow A. *De novo* diploid genome assembly for genome-wide structural variant detection. NAR Genom Bioinform. 2019 Dec 6;2(1):lqz018. doi: 10.1093/nargab/lqz018.

Zhao L, Liu H, Yuan X, Gao K, Duan J. Comparative study of whole exome sequencing-based copy number variation detection tools. BMC Bioinformatics. 2020 Mar 5;21(1):97. doi: 10.1186/s12859-020-3421-1.

Zhuang X, Ye R, So MT, Lam WY, Karim A, Yu M, Ngo ND, Cherny SS, Tam PK, Garcia-Barcelo MM, Tang CS, Sham PC. A random forest-based framework for genotyping and accuracy assessment of copy number variations. NAR Genom Bioinform. 2020 Sep 22;2(3):lqaa071. doi: 10.1093/nargab/lqaa071.

Sitography

OMIM: https://www.omim.org/

GENCODE: https://www.gencodegenes.org/

1000 Genomes Project Data Portal: https://www.internationalgenome.org/data-portal/sample

DGV: http://dgv.tcag.ca/

dbVar data: https://github.com/ncbi/dbvar/tree/master/Structural_Variant_Sets/Nonredundant_Structural_Variants

IMGT:https://www.imgt.org/

Acknowledgments

Several people have been of essential support in the making of this thesis. First, I am grateful to Romina D'Aurizio and Elia Giuseppe Ceroni, Consiglio Nazionale delle Ricerche Area della Ricerca di Pisa, for their collaboration and contribution to the design and development of TrainX. Alberto Magi and Roberto Semeraro, Department of Experimental and Clinical Medicine, University of Florence, helped with the generation of the synthetic CNVs using their tool ExomeBlender. This study makes proficient use of WES data generated by the Epi25 Collaborative, https://epi-25.org/epi25-members, a large-scale genomic project of the Centers for Common Disease Genomics (CCDG) program, funded by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI). I would like to thank Tania Giangregorio, U.O. Genetica Medica, IRCCS Azienda Ospedaliero-Universitaria di Bologna, for her help in processing Epi25 WES, and Francesca Bisulli, University of Bologna, who is Epi25 principal investigator for the Bologna site. Furthermore, Dennis Lal and Lisa-Marie Niestroj, Cologne Center for Genomics (CCG), University of Cologne, provided useful details on how to process Epi25 microarray data. Finally, Pamela Magini and Amalia Conti, U.O. Genetica Medica, IRCCS Azienda Ospedaliero-Universitari di Bologna, helped in performing molecular validation of clinical CNVs and in the training set design.