UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF

Biotechnology

GRADUATE SCHOOL OF

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Biotechnology

XXXII cycle

TITLE OF THE DOCTORAL THESIS

Enhanced genotypability for a more accurate variant calling in targeted resequencing

S.S.D. BIO/18

Coordinator: Prof. Matteo Ballottari

Tutor: Prof. Massimo Delledonne

Doctoral Student: Barbara Iadarola

Enhanced genotypability for a more accurate variant calling in targeted resequencing – Barbara Iadarola
Tesi di Dottorato
Verona, 08 Giugno 2020

ABSTRACT

The analysis of Next-Generation Sequencing (NGS) data for the identification of DNA genetic variants presents several bioinformatics challenges. The main requirements of the analysis are the accuracy and the reproducibility of results, as their clinical interpretation may be influenced by many variables, from the sample processing to the adopted bioinformatics algorithms. Targeted resequencing, which aim is the enrichment of genomic regions to identify genetic variants possibly associated to clinical diseases, bases the quality of its data on the depth and uniformity of coverage, for the differentiation between true and false positives findings. Many variant callers have been developed to reach the best accuracy considering these metrics, but they can't work in regions of the genome where short reads cannot align uniquely (uncallable regions). The misalignment of reads on the reference genome can arise when reads are too short to overcome repetitious regions of the genome, causing the software to assign a low-quality score to the read pairs of the same fragment. A limitation of this process is that variant callers are not able to call variants in these regions, unless the quality of one of the two read mates could increase. Moreover, current metrics are not able to define with accuracy these regions, lacking in providing this information to the final customer. For this reason, a more accurate metric is needed to clearly report the uncallable genomic regions, with the prospect to improve the data analysis to possibly investigate them. This work aimed to improve the callability (genotypability) of the target regions for a more accurate data analysis and to provide a high-quality variant calling.

Different experiments have been conducted to prove the relevance of genotypability for the evaluation of targeted resequencing performance. Firstly, this metric showed that increasing the depth of sequencing to rescue variants is not necessary at thresholds where genotypability reaches saturation (70X). To improve this metric and to evaluate the accuracy and reproducibility of results on different enrichment technologies for WES sample processing, the genotypability was evaluated on four exome platforms using three different DNA fragment lengths (short: ~200, medium: ~350, long: ~500 bp). Results showed that mapping quality could

successfully increase on all platforms extending the fragment, hence increasing the distance between the read pairs. The genotypability of many genes, including several ones associated to a clinical phenotype, could strongly improve. Moreover, longer libraries increased uniformity of coverage for platforms that have not been completely optimized for short fragments, further improving their genotypability. Given the relevance of the quality of data derived, especially from the extension of the short fragments to the medium ones, a deeper investigation was performed to identify a potential threshold of fragment length above which the improvement in genotypability was significant. On the enrichment platform producing the higher enrichment uniformity (Twist), the fragments above 230 bp could obtain a meaningful improvement of genotypability (almost 1%) and a high uniformity of coverage of the target. Interestingly, the extension of the DNA fragment showed a greater influence on genotypability in respect on the solely uniformity of coverage.

The enhancement of genotypability for a more accurate bioinformatics analysis of the target regions provided at limited costs (less sequencing) the investigation of regions of the genome previously defined as uncallable by current NGS methodologies.

CONTENT

ABSTRACT ... 3

CONTENT ... 5

INTRODUCTION ... 8

- Targeted resequencing ... 8
- Metrics for the evaluation of quality for WES enrichment technologies ... 11
- Data analysis: current challenges ... 15
- DNA fragment length ... 18
- The PANINI project ... 20

AIM OF THE THESIS ... 21

MATERIALS AND METHODS ... 22

- Sample processing ... 22
  - Genotypability evaluation on a single individual ... 22
  - Genotypability evaluation on different enrichment technologies and DNA fragment lengths ... 22
  - Genotypability evaluation on a variable set of DNA fragments ... 22
- Bioinformatics pipeline ... 23
  - Preprocessing of raw reads and sequence alignment ... 23
  - Metrics collection ... 26
  - Genotypability metric ... 27
  - Variant calling ... 31
- Datasets ... 33

RESULTS ... 35

- The genotypability metric and the depth of coverage ... 35
- The pipeline for the performance evaluation of WES ... 38
- WES performances on 3 different DNA fragment lengths ... 40
  - The 140X dataset ... 44
  - The 80X dataset (on design) ... 46
  - The 80X dataset (on RefSeq genes) ... 48
  - The multiple-downsampling dataset ... 53
  - Variant calling results ... 57
- WES performances on 9 different DNA fragment lengths ... 58
  - The 200X dataset ... 61

The 80X dataset (on design and RefSeq genes) 63

DISCUSSION 66

REFERENCES.. 70

ACKNOWLEDGEMENTS 73

INTRODUCTION

Targeted resequencing

Whole exome sequencing (WES) coupled with Next-generation Sequencing (NGS) platforms is a methodology that allows to capture and sequence the protein-coding regions of the genome with unprecedented efficiency [1]. Despite Whole-Genome Sequencing (WGS) is considered as the most comprehensive strategy for the analysis of the human genome, it still presents unaffordable costs for many research laboratories. For this reason, WES is becoming a standard, more economic approach for the analysis of disease-causing genetic variations [2][3][4]. Despite the limited regions covered by WES (about 1% of the entire genome [5]), this method can perform a deeper sequencing (higher coverage levels of the target regions) and hence produce a large quantity of data (sequenced reads) that needs to be analysed through specific bioinformatics pipelines [4][6][7][8].

The Illumina sequencing technology can produce millions of short sequence information (reads) in a single run. The DNA fragments produced for the sequencing library can be read to yield single-end reads (only one end of the fragment is sequenced) or paired-end reads (both ends of the fragment are sequenced) [9]. After reads are generated, they are aligned to a known reference genome sequence (i.e. human). Alignment algorithms perform better using the paired-end read information, since they exploit the known distance between the read pairs to produce a more precise mapping to the genome (Figure 1). In some cases, the sequenced DNA fragment could be shorter than the sum of the lengths of two read pairs, producing an overlap (Figure 2). This could lead to alignment issues and for this reason short DNA fragments are usually sequenced using the 75 paired-end mode (only 75 bp from both ends of the fragment are sequenced). However, the shorter the read, the more difficult will be its alignment to the genome.

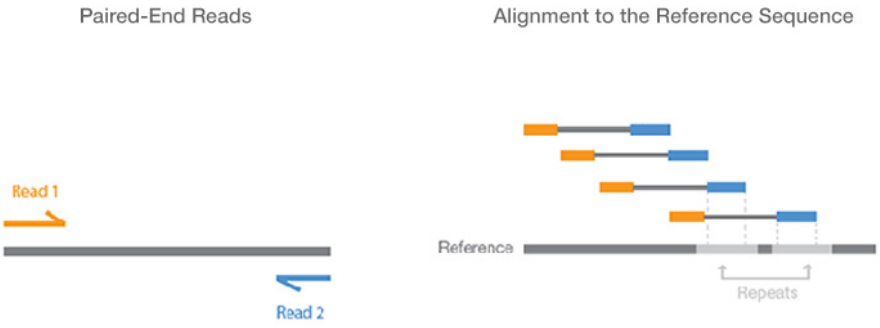


Figure 1. ***Paired-End vs. Single-Read Sequencing (Illumina).***

Retrieved 10/12/2019, from https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html?langsel=/it/

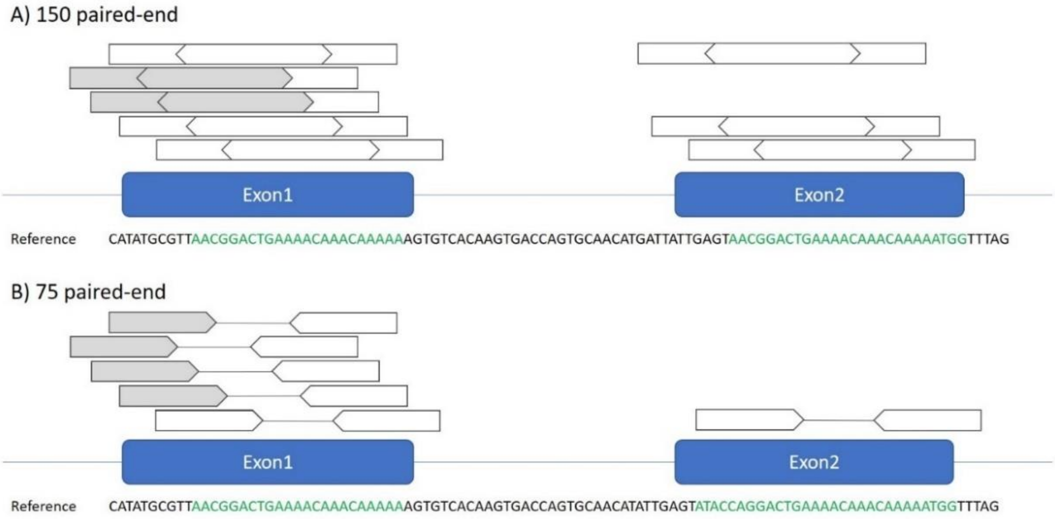

A) 150 paired-end
Exon1
Exon2
Reference
CATATGCGTTAACGGACTGAAAACAAACAAAAAAGTGTCACAAGTGACCAGTGCAACATGATTATTGAGTAACGGACTGAAAACAAACAAAAATGGTTTAG
B) 75 paired-end
Exon1
Exon2
Reference
CATATGCGTTAACGGACTGAAAACAAACAAAAAAGTGTCACAAGTGACCAGTGCAACATATTGAGTATACCAGGACTGAAAACAAACAAAAATGGTTTAG

Figure 2. ***Overlapping of sequenced read pairs.***

Reads must be therefore long enough to be aligned unambiguously to a known reference sequence [10] and the depth of coverage represents the number of times a base in the reference is covered by an aligned read from a sequencing experiment [11] (Figure 3).

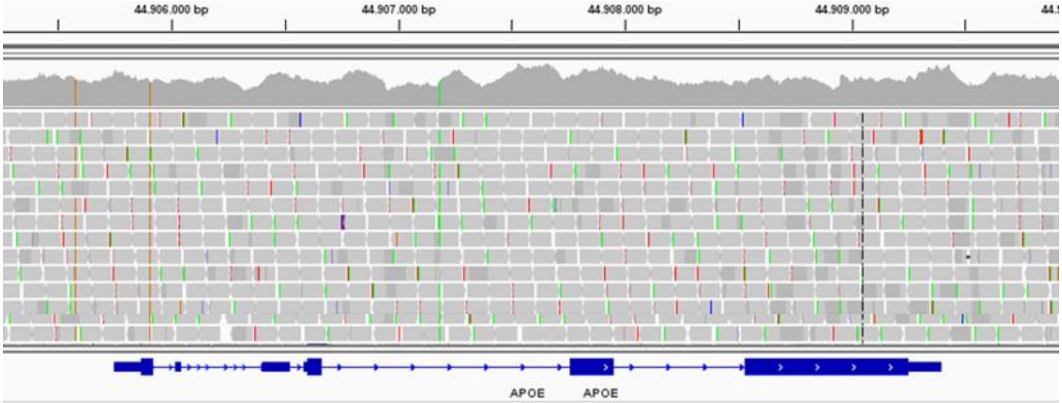


Figure 3. ***Graphical visualization of a BAM file.***

The picture shows in grey colour the aligned reads present in the sequence alignment file (BAM). Gaussian curves on the top of the picture show the coverage values of the region.

Variant calling is then performed after sequence alignment. Variations at the level of single nucleotides in the genome can be identified using different software, developed considering diverse algorithms and filtering strategies, thus leading to different outputs [12]. Lastly, the annotation of the identified variants is necessary to evaluate their biological potential consequences and hence their association to a variety of diseases [1].

Metrics for the evaluation of quality for WES enrichment technologies

Commercially available WES enrichment platforms for sample processing are designed with the aim to target selected regions of interest (ROI) through sequence enrichment, which is generally accomplished through probe-target hybridization. This methodology can directly capture large regions of interest (such as the human exome, ~30Mb) from a NGS library using complementary oligonucleotides in solution or array, with limited costs [13]. Currently, exome enrichment methods offered by vendors [14] differ in terms of enrichment efficiencies, targeted regions (Figure 4) and DNA input requirements [15]. Their performance is generally evaluated according to the depth and uniformity of coverage [16], because a minimum site coverage of more than 10-fold [17][18][19] is generally required to identify germline variants [11]. However, the average depth of coverage of the ROI is not indicative of the coverage of each gene/exon analysed, as some of them could be captured differently (or not captured at all, as shown in Figure 5) by the kits' probes.

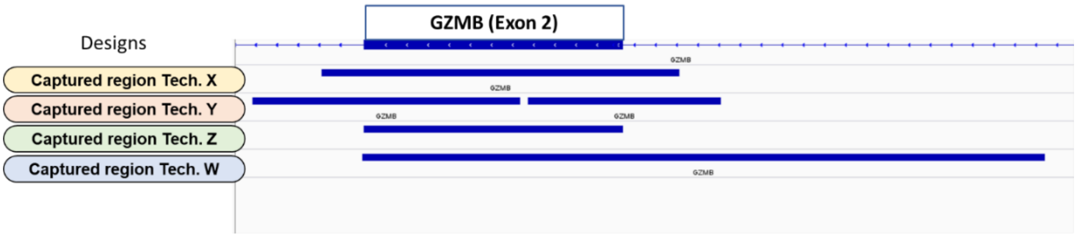


Figure 4. ***Differences in BED coordinates.***

Genomic coordinates of exon 2 of the GZMB gene reported in the BED files provided by different enrichment platforms suppliers.

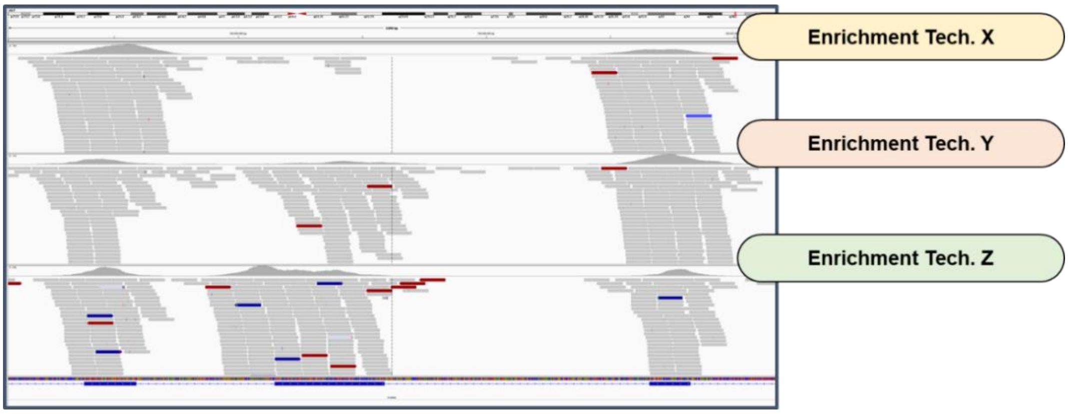


Figure 5. ***Differences in regions covered of the same gene.***

Three exons of the same gene are covered differently using diverse enrichment technologies.

Therefore, different enrichment technologies produce different coverage levels on the same set of genes (Figure 6). Considering these dissimilarities, the choice of the more appropriate platform to use for a clinical investigation of a candidate set of genes is generally based on the genes' optimal overall coverage obtained across the different kits.

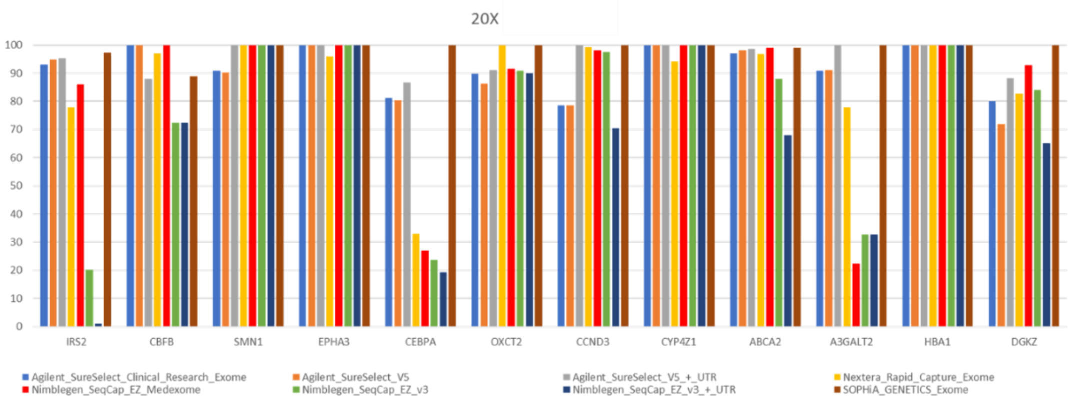


Figure 6. ***The difference in coverage levels for a set of genes enriched through different enrichment platforms.***

Percentage of each gene covered at least by 20 reads for 8 different enrichment platforms.

The optimal coverage for the achievement of useful data is obtained when the entire length of the target regions reaches the desired coverage at the expense of the off-target rate, which is referred to as the sequencing data mapping near (~250 bp) and outside the target region (Figure 7). Most of the off-target sequencing is probe panel-specific and is usually a result of indiscriminate hybridization. On-target and off-target rates are both considered in combination with the uniformity of enrichment (FOLD 80 penalty value) to define the efficiency of the targeted resequencing.

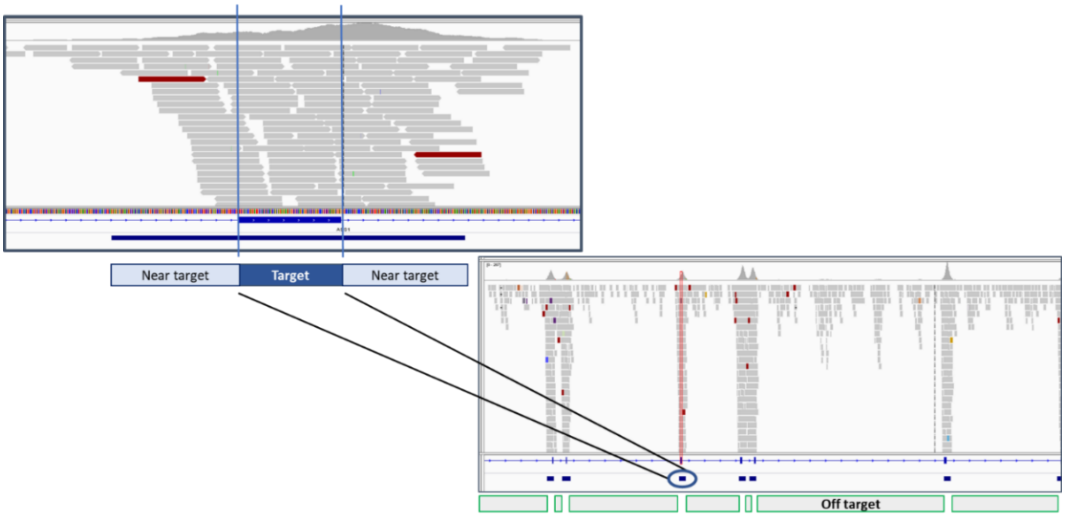


Figure 7. ***Definition of on/near/off target.***

The uniformity of coverage describes the read distribution along target regions of the genome and it’s calculated by the FOLD 80 penalty value, which indicates “the fold of additional sequencing required to ensure that 80% of the target bases achieve the desired average coverage”. FOLD 80 penalty is calculated as the average coverage on target divided by the coverage at the 80th percentile, which is the coverage value that lies at the 80% line of an ordered set of coverage values representing each sequenced base of the target. Uniformity can be improved reducing the coverage of over-sequenced targets and increasing the coverage of the target with lower sequencing, so that the amount of sequencing needed to obtain high-confidence data will be reduced [20] (Figure 8). Therefore, small improvements in uniformity can have a much larger impact on increasing the efficiency of the targeted resequencing.

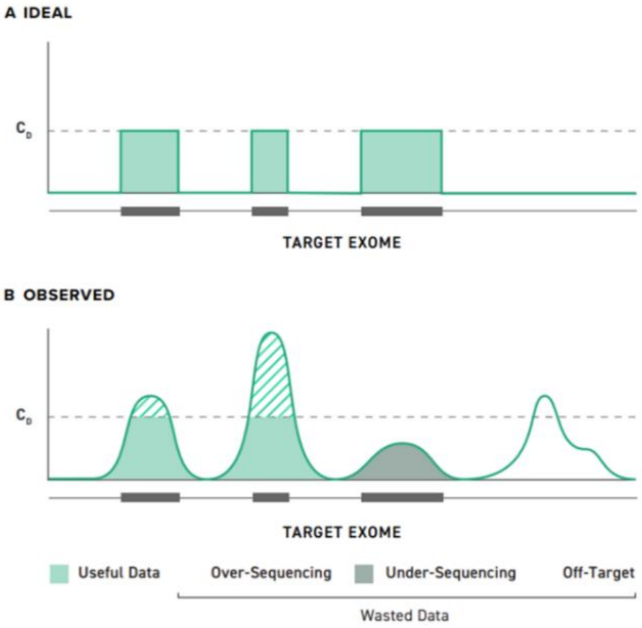


Figure 8. ***Uniformity of coverage (Twist [20]).***

Data analysis: current challenges

Challenges of current bioinformatic pipelines for WES data analysis are diverse, from the read alignment to the variant calling. Read alignment could negatively affect the identification of variants if the reads are not correctly assigned to their position along the reference genome. This problem could arise if the reads map to multiple locations on the reference sequence, and various strategies have been adopted to solve it [13]:

- Discard the reads mapping not uniquely to the genome (this can cause an omission of up to 30% of mappable reads)
- The best-match approach maps the reads choosing the location with the fewest mismatches (in case of more than one best match, all locations or a random selection is provided)
- Report all alignments until a maximum number consented

However, the second and the third approach could lead to a misalignment of reads, especially in repetitious regions of the genome. Sequence aligners assign quality scores to read pairs according to the uniqueness of the alignment (probability the read is not mapped randomly), so reads mapping to duplicated regions gain a low quality. However, if one of the two read mates can be mapped unambiguously they both may gain a high quality score [8][21][22].

The use of the solely depth of coverage as main quality parameter for the WES performance presents therefore some limitations. Indeed, high coverage levels do not always correspond to a high quality of read alignment, as shown in Figure 9. If the target region is repeated along the genome, the quality of the reads aligned there is low. In case a variant is present in this region, the software cannot provide a high confidence of call.

For this reason, depth and uniformity of coverage cannot be considered as the main parameters for the evaluation of WES performances.

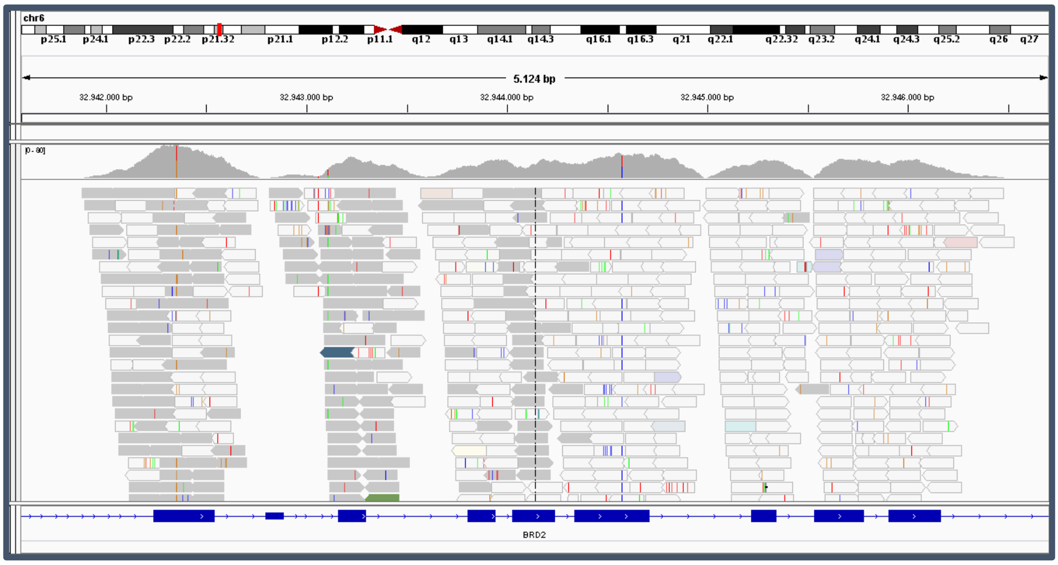


Figure 9. ***Region of the genome highly covered but with low mapping quality.***

Graphic visualization of a BAM file. Colour of reads indicates the mappinq quality: white = low quality mapping; grey = high quality mapping.

Genotypability is a metric introduced in this work which reports the “callable” and “uncallable” regions of the target, through the use of the gVCF (Figure 10). The pipeline chosen for the analysis of WES data integrates the use of the gVCF file for the analysis of the base calling at the level of the entire genome. While current pipelines use the VCF to store the high-confidence variant sites present in the analysed individual’s genome, the gVCF contains also the invariant sites passing the quality filters, allowing the distinction between a variant “not called” because not present in the individual’s genome and a variant “not called” because the site coverage and the quality of the alignment in that position do not satisfy the requirements of base calling.

This value can be calculated for any region of interest (target design, RefSeq genes, a locus) and in this work it is used to evaluate how the DNA fragment length could improve the sequence alignment of genomic regions which do not satisfy the requirements of minimum read depth and mapping quality.

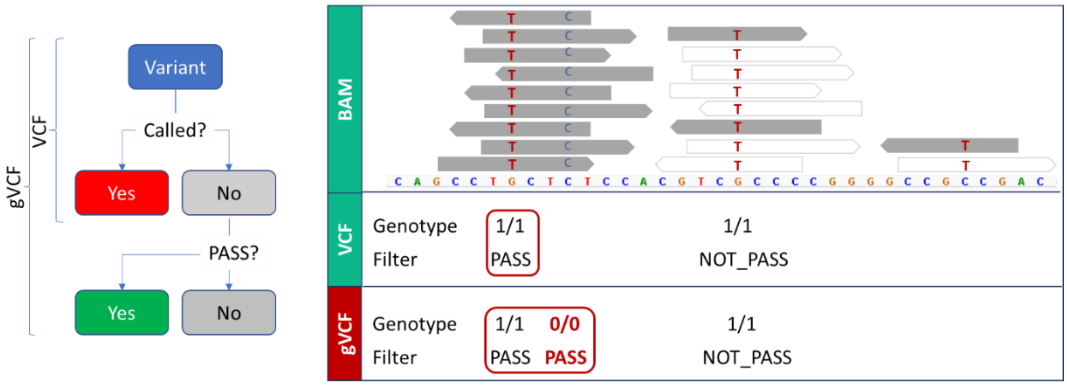
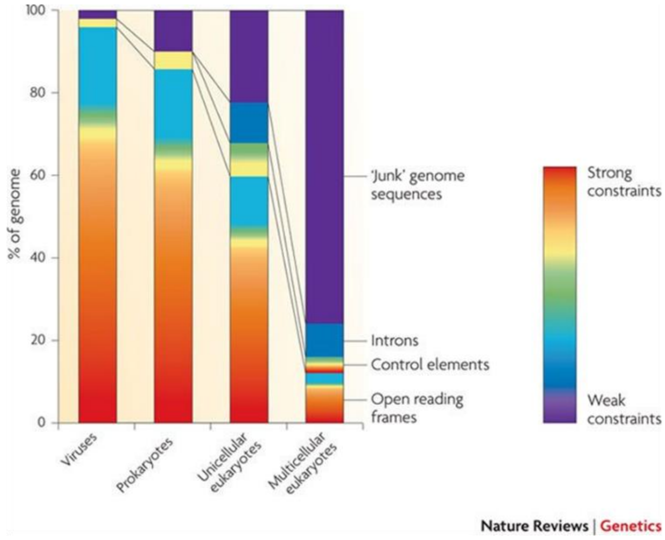


Figure 10. ***Differences between a VCF and a gVCF.***

gVCF reports not only the variant sites in the genome, but also the invariant one which can satisfy the requirements of minimum mapping quality and depth of coverage.

Through the use of gVCF it is possible to rescue homozygous reference variants for the data analysis, but unravelling the variants present in repetitious regions of the genome is still difficult using short reads. For this reason, an approach to improve this metric is needed.

DNA fragment length



Koonin et al. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*

Figure 11. ***Constraints in genome evolution.***

WES library preparation protocols set the DNA fragment size to the average exon length, which is 170 bp in the human genome [21][22][23]. Short (< 100 bp) paired-end reads are generated to avoid the overlap of read pairs, but this fragment length is often shorter than duplicated regions. Furthermore, library preparation protocols often start from very low quantities of material (nanograms to picograms) [24], limiting the amount of DNA and consequently the number of unique fragments that can be produced. For this reason, 2 × 75 sequencing requires double the number of fragments to produce the expected depth of coverage that can be achieved by 2 × 150 sequencing. More amplification is therefore necessary, producing more PCR duplicates that must be removed during downstream data analysis, thus limiting the depth of coverage at target regions [25].

Considering the challenges due to the difficult alignment of short reads to repetitious regions of the genome, this approach aims to increase the standard DNA fragment size to allow longer fragments to extend beyond exonic regions to reach introns, which are under less selection pressure than protein coding sequences but still retain conserved polymorphisms [23] (Figure 11). This means that introns are still evolutionary conserved (as they are important in regulating gene expression), but with a greater variability in respect of exons. Therefore, reads that cannot uniquely align in repetitious exonic regions could better align on flanking intronic regions, that may not conserve the same repetition. In this way, the higher quality of mapping obtained for the read mapping outside of repetitious regions can be transferred to its mate, allowing the identification of variants that could be otherwise discarded, due to poor-mapping (Figure 12).

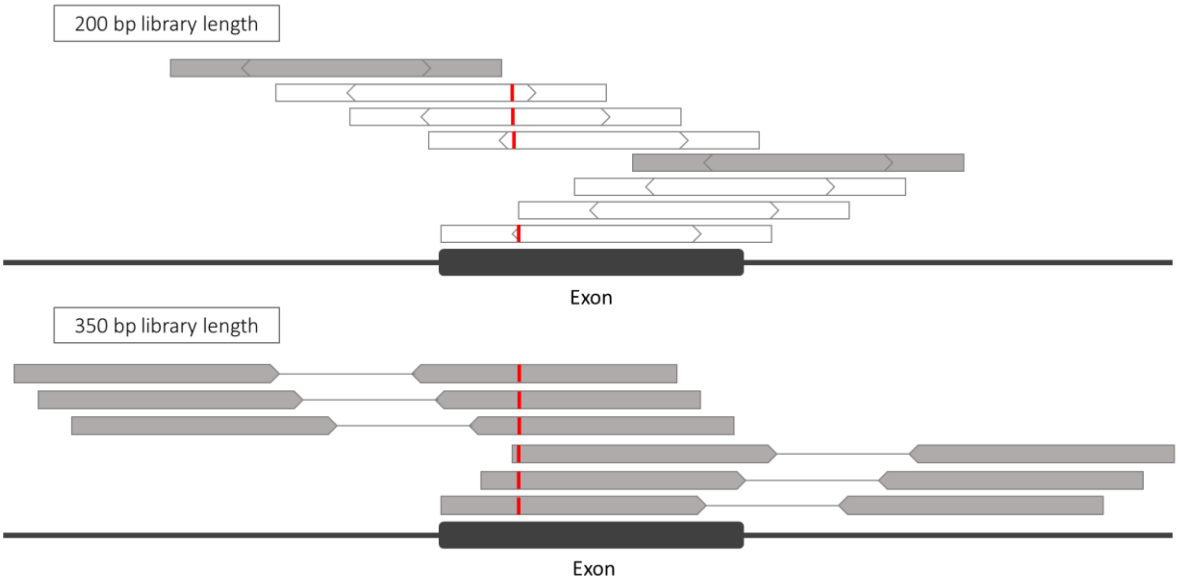


Figure 12. ***Alignment of read pairs and extension of the DNA fragment.***

Transfer of high-quality of read alignment between read pairs through the extension of the DNA fragment size.

The PANINI project

Considering the implications of targeted analysis for diagnosis and therapies, the European project "Physical Activity and Nutrition INfluences In ageing" (PANINI) aimed to develop a policy document to promote healthy ageing in Europe [26]. In particular, the project addressed the need to identify the genetic markers responsible for ageing diseases and nutritional responses, to drive personalized treatments to older adults. The application of a bioinformatics pipeline such as the one here developed could provide the more accurate analysis for this type of clinical employment.

AIM OF THE THESIS

Variant calling on human DNA-seq samples presents limitations due to possible misalignments of short reads on the reference genome and incomplete quality metrics to define the accuracy of bioinformatics data. The aim of the thesis was to improve the accuracy of data produced by targeted resequencing workflows through the enhancement of the genotypability of the target regions, for a more precise variant calling and a more accurate investigation of the uncallable regions of the genome.

MATERIALS AND METHODS

Sample processing

Genotypability evaluation on a single individual

The WES analysis was performed on data derived from an individual processed by the wet-lab using the Human Core Exome Kit + RefSeq V1 enrichment platform (Twist), producing a DNA fragment size based on the manufacturers' recommendations.

Genotypability evaluation on different enrichment technologies and DNA fragment lengths

The WES analysis was performed on data derived from three unrelated individuals. Samples were processed using four different enrichment platforms: xGen Exome Research Panel V1 (IDT), SeqCap EZ MedExome (Roche), SureSelect Human All Exon V6 (Agilent), and the Human Core Exome Kit + RefSeq V1 (Twist). The wet-lab produced three different DNA fragment lengths for each sample: short fragments based on the manufacturers' recommendations (IDT = 150 bp, Roche, Agilent and Twist = 200 bp), medium fragments (expected length ~350 bp), and long fragments (expected length ~500 bp).

Genotypability evaluation on a variable set of DNA fragments

The WES analysis was performed on data derived from 27 individuals processed by the wet-lab using the Human Core Exome Kit + RefSeq V1 enrichment platform (Twist), which produced a variable set of DNA fragments sizes from ~200 to ~350 bp.

Bioinformatics pipeline

Preprocessing of raw reads and sequence alignment

All individuals were sequenced on an Illumina instrument in 75 bp paired-end mode for the short libraries (~200 bp) and in 150 bp paired-end mode for the other DNA fragment lengths. All samples were analysed performing a preprocessing of raw reads and the alignment to the reference genome sequence.

The preprocessing pipeline was based on a set of available tools as described below (Figure 13).

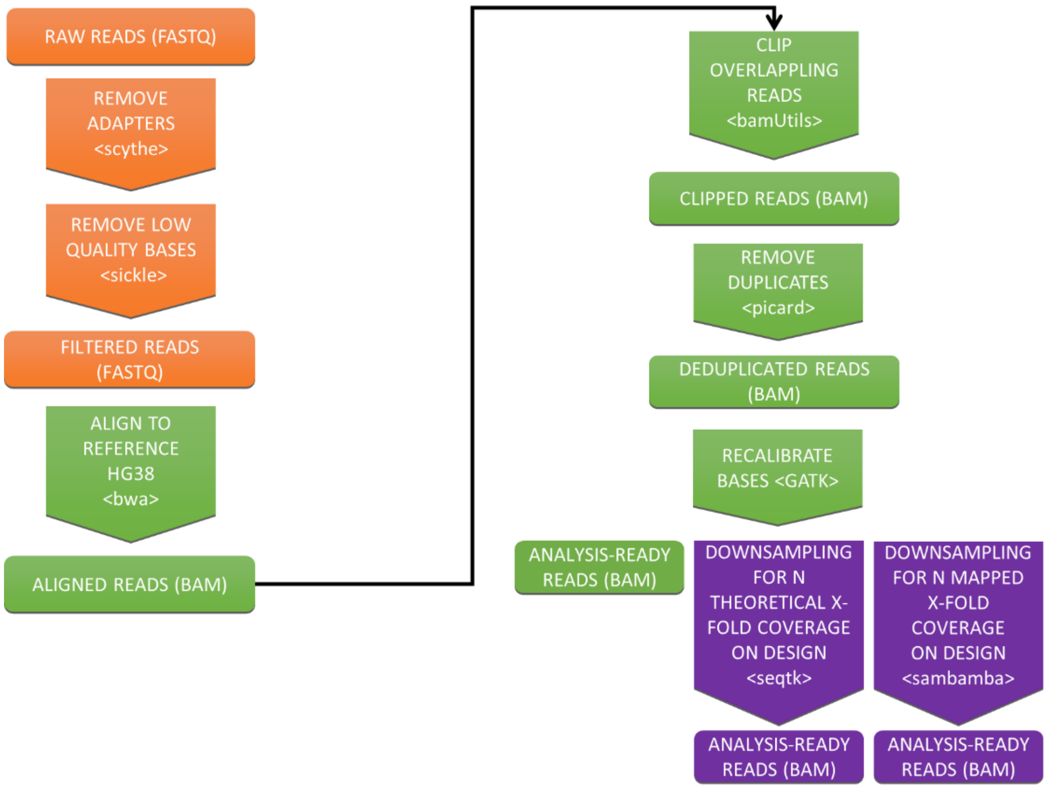

RAW READS (FASTQ)
REMOVE ADAPTERS <scythe>
REMOVE LOW QUALITY BASES <sickle>
FILTERED READS (FASTQ)
ALIGN TO REFERENCE HG38 <bwa>
ALIGNED READS (BAM)
CLIP OVERLAPPLING READS <bamUtils>
CLIPPED READS (BAM)
REMOVE DUPLICATES <picard>
DEDUPLICATED READS (BAM)
RECALIBRATE BASES <GATK>
ANALYSIS-READY READS (BAM)
DOWNSAMPLING FOR N THEORETICAL X-FOLD COVERAGE ON DESIGN <seqtk>
DOWNSAMPLING FOR N MAPPED X-FOLD COVERAGE ON DESIGN <sambamba>
ANALYSIS-READY READS (BAM)
ANALYSIS-READY READS (BAM)

Figure 13. ***Pipeline for preprocessing of raw reads and sequence alignment.***

Initial FASTQ files were quality controlled using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low quality nucleotides have been trimmed using sickle v1.33

(https://github.com/najoshi/sickle) and adaptors were removed using scythe v0.991 (https://github.com/vsbuffalo/scythe).

```
#FastQC

fastqc sample.read1.fastq.gz sample.read2.fastq.gz -o fastqc/

#Trimming

sickle pe -g -t sanger \
      -f <( scythe -a adapters.file -q sanger sample.read1.fastq.gz } ) \
      -r <( scythe -a adapters.file -q sanger sample.read2.fastq.gz } ) \
      -o trimmed1.fastq.gz -p trimmed2.fastq.gz -s /dev/null;
```

Reads were then aligned to the reference human genome sequence (GRCh38/hg38) using BWA-MEM v0.7.15 (https://arxiv.org/abs/1303.3997). BWA is a fast and memory-efficient read aligner widely used for WES. The SAM output file was converted into a sorted BAM file using SAMtools. Overlapping regions of the BAM file were clipped using BamUtil v1.4.14 to avoid counting multiple reads representing the same fragment. The BAM files were processed by local realignment around insertion–deletion sites, duplicate marking and recalibration using Genome Analysis Toolkit v4.0.2.1 [27].

```
#Alignment

my $RG = '"@RG'
     . qq|\tID:$ID\tPU:lane\tLB:$NAME\tSM:$NAME\tCN:CGF-ddlab\tPL:ILLUMINA"|;

bwa mem -R $RG -t 16 Homo_sapiens_assembly38.fasta trimmed*fastq.gz |
samtools sort --threads 4 -m 5G - -o start_sorted.bam

sambamba index --nthreads= 20 start_sorted.bam

#Clipping

bam clipOverlap --in start_sorted.bam --out start_sorted.clipped.bam

#Duplicate marking

java -jar gatk.jar MarkDuplicates -I start_sorted.clipped.bam -O alignment.rg.bam -M
duplicates.txt --REMOVE_DUPLICATES true --VALIDATION_STRINGENCY SILENT --
CREATE_INDEX true

#Base recalibrator

java -jar gatk.jar BaseRecalibrator -I alignment.rg.bam -R
Homo_sapiens_assembly38.fasta --use-original-qualities --knownSites dbsnp.vcf --
knownSites mills.vcf -O recal_data.table

java -jar gatk.jar ApplyBQSR -R Homo_sapiens_assembly38.fasta -I alignment.rg.bam
-bsqr recal_data.table -O alignment.rg.recalibrated.bam --static-quantized-quals 10 --
static-quantized-quals 20 --static-quantized-quals 30 --add-output-sam-program-
record --create-output-bam-md5 --use-original-qualities
```

Downsampling for a N theoretical X-fold coverage on the target design was calculated subsampling the required number of fragments (calculated as: (N * design length) / (read length * 2)) using seqtk (https://github.com/lh3/seqtk). Downsampling on the mapped coverage was generated by sub-sampling the full dataset using sambamba v0.6.7 – https://github.com/biod/sambamba –).

```
#Downsampling for theoretical coverage

seqtk sample -s100 sample.read1.fastq.gz $number_of_fragments
seqtk sample -s100 sample.read2.fastq.gz $number_of_fragments

#Downsampling for mapped coverage

my $ratio = $target_mapped_coverage / $real_mapped_coverage
sambamba view -h -t 30 -s $ratio -f bam alignment.rg.recalibrated.bam -o
downsampled.bam
```

Metrics collection

Insert sizes were calculated after read alignment, measuring the distance of the two mates mapped on the genome using CollectInsertSize by Picard v2.17.10 (http://broadinstitute.github.io/picard/). CollectHsMetrics by Picard was used to calculate fold enrichment and FOLD 80 penalty values to determine enrichment quality. For each sample, the near target length was defined as the “average length of the DNA fragments” padding the on-target region. All WES performance parameters were calculated both on the design of each platform and on the standard dataset of RefSeq genes.

```
#Insert size collection

java -jar gatk.jar CollectInsertSizeMetrics -I alignment.rg.recalibrated.bam -H
alignment.rg.recalibrated.hist.pdf -O alignment.rg.recalibrated.output -AS true --
VALIDATION_STRINGENCY SILENT

#Mapping statistics

samtools flagstat alignment.rg.recalibrated.bam > flagstat_recal

#ON/NEAR/OFF target statistics on design and RefSeq

java -jar GenomeAnalysisTK.jar CollectHsMetrics --INPUT
alignment.rg.recalibrated.bam --OUTPUT design.HsMetrics.txt -R
Homo_sapiens_assembly38.fasta \
        --BAIT_INTERVALS design.bed.interval --TARGET_INTERVALS design.bed.interval
\
        --PER_TARGET_COVERAGE design.PER_TARGET_COVERAGE.txt --
PER_BASE_COVERAGE design.PER_BASE_COVERAGE.txt --
VALIDATION_STRINGENCY=SILENT \
        --NEAR_DISTANCE insert_length

java -jar GenomeAnalysisTK.jar CollectHsMetrics --INPUT
alignment.rg.recalibrated.bam --OUTPUT RefSeq.HsMetrics.txt -R
Homo_sapiens_assembly38.fasta \
        --BAIT_INTERVALS RefSeq.bed.interval --TARGET_INTERVALS RefSeq.bed.interval
\
        --PER_TARGET_COVERAGE RefSeq.PER_TARGET_COVERAGE.txt --
PER_BASE_COVERAGE RefSeq.PER_BASE_COVERAGE.txt --
VALIDATION_STRINGENCY=SILENT \
        --NEAR_DISTANCE insert_length
```

Genotypability metric

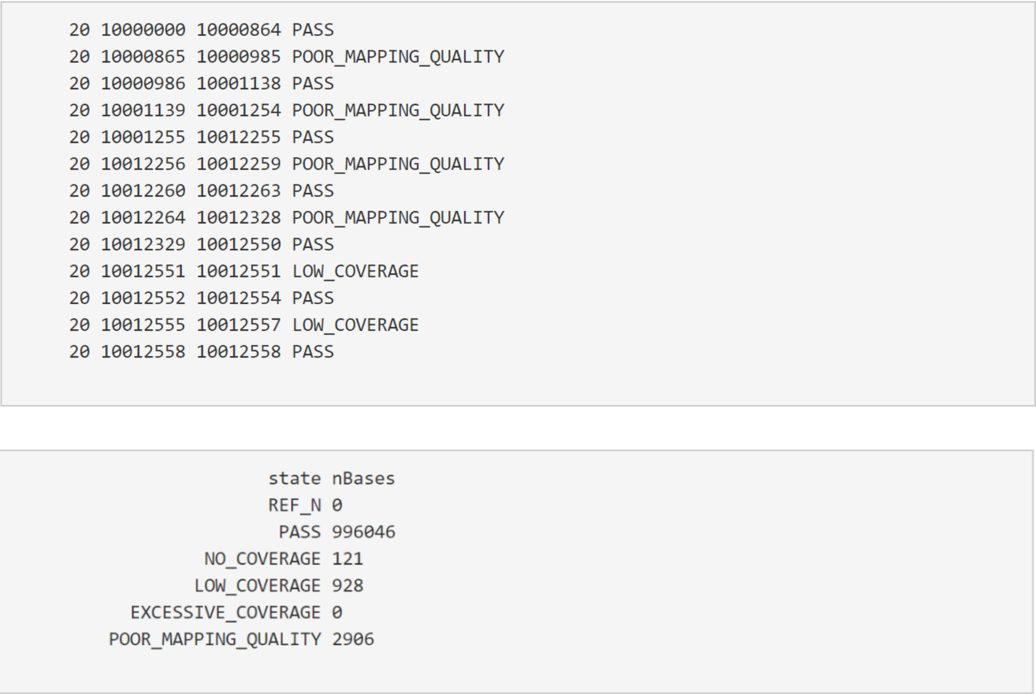
I then used *CallableLoci* in GATK v3.8 to identify callable regions of the target (genotypability), with minimum read depths of 3 and 10.

```
#CallableLoci

java -jar GenomeAnalysisTK.jar -T CallableLoci -R Homo_sapiens_assembly38.fasta -I
alignment.rg.recalibrated.bam -summary callable_table.txt -o callable_status.bed

java -jar GenomeAnalysisTK.jar -T CallableLoci -R Homo_sapiens_assembly38.fasta -I
alignment.rg.recalibrated.bam -minDepth 10 -summary callable_table.txt -o
callable_status_DP10.bed
```

CallableLoci produces a BED file with the callable status covering each base and a summary table of callable status per count of all examined bases (Figure 14).

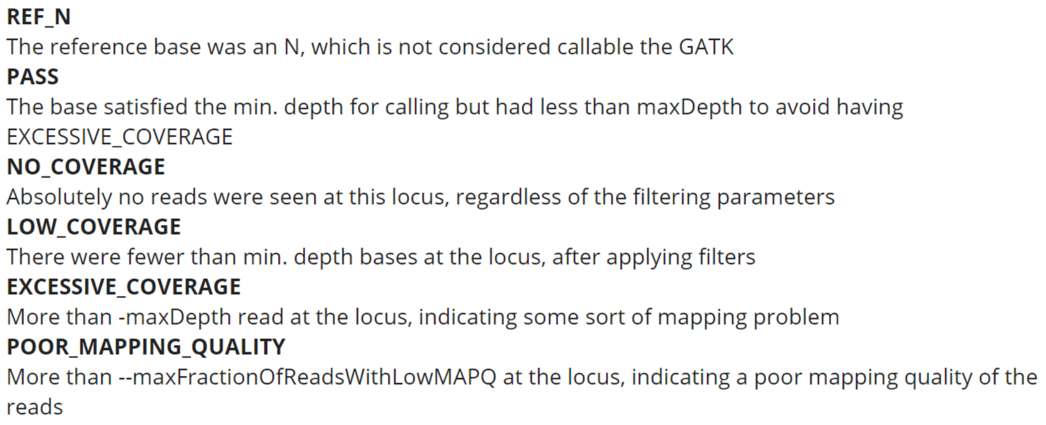


```
20 10000000 10000864 PASS
20 10000865 10000985 POOR_MAPPING_QUALITY
20 10000986 10001138 PASS
20 10001139 10001254 POOR_MAPPING_QUALITY
20 10001255 10012255 PASS
20 10012256 10012259 POOR_MAPPING_QUALITY
20 10012260 10012263 PASS
20 10012264 10012328 POOR_MAPPING_QUALITY
20 10012329 10012550 PASS
20 10012551 10012551 LOW_COVERAGE
20 10012552 10012554 PASS
20 10012555 10012557 LOW_COVERAGE
20 10012558 10012558 PASS
```

```
               state nBases
               REF_N 0
                PASS 996046
         NO_COVERAGE 121
        LOW_COVERAGE 928
  EXCESSIVE_COVERAGE 0
POOR_MAPPING_QUALITY 2906
```

Figure 14. ***BED file e summary table produced by CallableLoci.***

The callable states of the genomic intervals are summarised in Figure 15.



REF_N
The reference base was an N, which is not considered callable the GATK
PASS
The base satisfied the min. depth for calling but had less than maxDepth to avoid having EXCESSIVE_COVERAGE
NO_COVERAGE
Absolutely no reads were seen at this locus, regardless of the filtering parameters
LOW_COVERAGE
There were fewer than min. depth bases at the locus, after applying filters
EXCESSIVE_COVERAGE
More than -maxDepth read at the locus, indicating some sort of mapping problem
POOR_MAPPING_QUALITY
More than --maxFractionOfReadsWithLowMAPQ at the locus, indicating a poor mapping quality of the reads

Figure 15. ***Callable states of CallableLoci.***

bedtools coverage was used to calculate the coverage of the target regions at several coverage levels (1X, 5X, 10X, 20X, 30X) and the genotypability of the target at read depths of 3 (% PASS) and 10 (% PASS RD>10). Through a specific script (*geneCoverage.pl*), the information of the depth of coverage and the number of bases at that depth from the design.alignment.rg.recalibrated.capture.hist.coverage.gz and the RefSeq.alignment.rg.recalibrated.capture.hist.coverage.gz files were calculated for:

- each region of the target design/RefSeq;
- each gene of the target design/RefSeq;
- all the target design/RefSeq.

```
#Region coverage calculations

# region coverage for design

bedtools coverage -hist
        -abam alignment.rg.recalibrated.bam \
        -b design.bed | gzip >
design.alignment.rg.recalibrated.capture.hist.coverage.gz

bedtools coverage -hist
        -a callable.bed \
        -b design.bed > design.alignment.rg.recalibrated-callable.bed

bedtools coverage -hist
        -a callable_DP10.bed \
        -b design.bed > design.alignment.rg.recalibrated-callable_DP10.bed

# region coverage for RefSeq

bedtools coverage -hist
        -abam alignment.rg.recalibrated.bam \
        -b RefSeq.bed | gzip >
RefSeq.alignment.rg.recalibrated.capture.hist.coverage.gz

bedtools coverage -hist
        -a callable.bed \
        -b RefSeq.bed > RefSeq.alignment.rg.recalibrated-callable.bed

bedtools coverage -hist
        -a callable_DP10.bed \
        -b RefSeq.bed > RefSeq.alignment.rg.recalibrated-callable_DP10.bed
```

Then, the output files design.alignment.rg.recalibrated-callable.bed and RefSeq.alignment.rg.recalibrated-callable.bed were used for the calculation of the genotypability of the target. The number of CALLABLE bases covered for each region of the target design/RefSeq was extracted from the two files and transformed into a percentage value.

```
#Genotypability and coverage statistics

#For design

geneCoverage.pl
            design.alignment.rg.recalibrated.capture.hist.coverage.gz \
            design.alignment.rg.recalibrated-callable.bed \
            design.alignment.rg.recalibrated-callable_DP10.bed

#For RefSeq

geneCoverage.pl
            RefSeq.alignment.rg.recalibrated.capture.hist.coverage.gz \
            RefSeq.alignment.rg.recalibrated-callable.bed \
            RefSeq.alignment.rg.recalibrated-callable_DP10.bed
```

Variant calling

Variant calling was performed producing gVCF files through the GATK HaplotypeCaller v4.1.2.0 software. It calls germline Single Nucleotide Variations and indels via a local re-assembly of haplotypes.

```
#Variant calling

java -jar gatk.jar HaplotypeCaller -R Homo_sapiens_assembly38.fasta -I
alignment.rg.recalibrated.bam --dbsnp dbsnp.vcf -ERC GVCF --output snps.raw.g.vcf -
-standard-min-confidence-threshold-for-calling 30.0 --force-active true
```

Variant recalibration was performed to assign a well-calibrated probability to each variant call in a call set. This enabled the generation of highly accurate call sets by filtering based on this single estimate for the accuracy of each call.

```
#Variant recalibration

java -jar gatk.jar GenotypeGVCFs  -R Homo_sapiens_assembly38.fasta -V
snps.raw.g.vcf -G StandardAnnotation -O complete.raw.variants.vcf

java -jar gatk.jar VariantRecalibrator -V complete.raw.variants.vcf -O
INDEL.recalibration --tranches-file INDEL.tranches --trust-all-polymorphic \
      -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
       -an QD -an DP -an FS -an SOR -an MQRankSum -an ReadPosRankSum \
       -mode INDEL --max-gaussians 4 -R $REF \
       -resource:mills,known=false,training=true,truth=true,prior=12 mills.vcf \
       -resource:axiomPoly,known=false,training=true,truth=false,prior=10 axiom.vcf \
       -resource:dbsnp,known=true,training=false,truth=false,prior=2 dbsnp.vcf

java -jar gatk.jar VariantRecalibrator -V complete.raw.variants.vcf -O
SNPS.recalibration --tranches-file SNPS.tranches --trust-all-polymorphic \
       -tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
       -an DP -an QD -an FS -an SOR -an MQ -an MQRankSum -an ReadPosRankSum -
mode SNP --max-gaussians 6 \
       -resource:hapmap,known=false,training=true,truth=true,prior=15 hapmap.vcf \
       -resource:omni,known=false,training=true,truth=true,prior=12 omni.vcf \
       -resource:1000G,known=false,training=true,truth=false,prior=10 phase1.vcf \
       -resource:dbsnp,known=true,training=false,truth=false,prior=7 dbsnp.vcf

java -jar gatk.jar ApplyVQSR -O indel.recalibrated.vcf -V complete.raw.variants.vcf --
recal-file INDEL.recalibration --tranches-file INDEL.tranches \
       --truth-sensitivity-filter-level 99 --create-output-variant-index true -mode INDEL

java -jar gatk.jar ApplyVQSR -O variants.recalibrated.vcf -V indel.recalibrated.vcf --
recal-file SNPS.recalibration --tranches-file SNPS.tranches \
       --truth-sensitivity-filter-level 99 --create-output-variant-index true -mode SNP
```

As a final step, a hard-filtering was performed on variant calls based on certain criteria. In particular, for SNPs:

- QD (Quality by Depth) was set to < 2.0
- MQ (RMS Mapping Quality) was set to < 40
- FS (Fisher Strand) was set to > 60.0
- SOR (Strand Odds Ratio) was set to > 3.0
- MQRankSum (MappingQualityRankSumTest) was set to < -12.5
- ReadPosRankSum was set to < -8.0

as specified by the GATK Best Practices.

While, for the indels, parameters were set as following:

- QD (Quality by Depth) was set to < 2.0
- FS (Fisher Strand) was set to > 200.0
- ReadPosRankSum was set to < -20.0

as specified by the GATK Best Practices.

```
#Variant filtering

java -jar gatk.jar SelectVariants --select-type-to-include SNP --output raw_snps.vcf -V
complete.raw.variants.vcf

java -jar gatk.jar SelectVariants --select-type-to-exclude SNP --output raw_indels.vcf -
V complete.raw.variants.vcf

java -jar gatk.jar VariantFiltration -R Homo_sapiens_assembly38.fasta -V
raw_snps.vcf \
          --filter-expression "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 ||
MQRankSum < -12.5 || ReadPosRankSum < -8.0" \
          --filter-name "Broad_SNP_filter" -O  raw_filtered_snps.vcf

java -jar gatk.jar VariantFiltration -R Homo_sapiens_assembly38.fasta -V
raw_indels.vcf \
          --filter-expression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
          --filter-name "Broad_indel_Filter" -O raw_filtered_indels.vcf

java -jar gatk.jar MergeVcfs -I raw_filtered_snps.vcf -I raw_filtered_indels.vcf -O
variants.filtered.vcf

java -jar gatk.jar SelectVariants -R Homo_sapiens_assembly38.fasta --variant
variants.filtered.vcf.gz --exclude-filtered -O variants.selected.vcf
```

Datasets

The RefSeq database (release 82) was downloaded from the UCSC Genome Table Browser (http://genome.ucsc.edu/). Online Mendelian Inheritance in Man (OMIM) genes associated with a clinical phenotype were downloaded from the OMIM website (https://www.omim.org/, release 15-05-2018).

RESULTS

The genotypability metric and the depth of coverage

I initially performed an evaluation to assess the relevance of the genotypability metric in respect of the solely depth of coverage. Whole Exome Sequencing was performed on a representative sample (VRXX) processed by the wet-lab using the Twist Human Exome Core plus RefSeq v.1 Reagent kit. The expected mapped coverage for this experiment was of 200X, meaning that the entire region of interest could be read on average 200 times. From the initial set of sequenced reads, I produced downsampled BAM files (with an average X-fold coverage of 10–190) on the target design, to evaluate the percentage of the ROI covered by a minimum number of reads (1,5,10,20,30) at different coverage levels (10-200X). The same evaluation was performed for the genotypability (callability) of the target.

Results showed that the percentage of the target covered by at least 10 reads, augmented significantly between 10-140X and then reached saturation (Table 1). In a similar way, the %20X value increased continuously until reaching saturation at 170X, whereas the %30X could not reach saturation even at 200X mapped coverage. Considering these results, in order to reach a substantial coverage along all the target regions, the sequencing depth should be greatly increased. However, the genotypability of the target calculated using the standard requirements of the GATK workflow (read depth >3) reached saturation at 70X mapped coverage (Figure 16). Genotypability at a minimum read depth of 10, the coverage threshold suggested by clinical guidelines for the variant identification in genetics laboratories, instead reached saturation at higher coverage levels (130X).

These results showed that increasing the sequencing coverage above 70X could not lead to a better callability of the target regions, if considering the standard requirements of GATK. Thus, coverage levels commonly considered too low for variants identification instead could be potentially adequate for the bioinformatics data analysis. For clinical settings, further sequencing coverage could be required based on the laboratory's needs.

Mapped Coverage	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10
10	99.39	86.23	49.22	4.48	0.59	87.44	45.96
20	99.86	97.46	87.66	48.19	13.78	94.23	83.53
30	99.90	99.19	95.38	77.46	47.97	95.18	91.10
40	99.91	99.64	97.81	88.34	70.72	95.43	93.50
50	99.92	99.80	98.88	93.16	82.58	95.49	94.54
60	99.92	99.85	99.38	95.65	88.67	95.51	95.03
70	99.92	99.88	99.62	97.11	92.12	95.53	95.26
80	99.92	99.89	99.74	98.04	94.30	95.53	95.38
90	99.92	99.90	99.81	98.63	95.77	95.53	95.44
100	99.92	99.90	99.84	99.03	96.78	95.54	95.48
110	99.92	99.90	99.86	99.30	97.53	95.54	95.50
120	99.92	99.91	99.87	99.49	98.11	95.54	95.51
130	99.92	99.91	99.88	99.60	98.53	95.54	95.52
140	99.92	99.91	99.89	99.69	98.85	95.54	95.52
150	99.92	99.91	99.89	99.74	99.09	95.54	95.53
160	99.92	99.91	99.90	99.78	99.28	95.54	95.53
170	99.92	99.91	99.90	99.81	99.41	95.54	95.53
180	99.92	99.91	99.90	99.83	99.52	95.54	95.53
190	99.92	99.91	99.90	99.84	99.60	95.54	95.54
200	99.92	99.91	99.91	99.86	99.70	95.54	95.54

Table 1. ***Downsampled mapped coverage.***

Metrics were calculated on 10–200X downsampled sets, including mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, and percentage of callable bases on the target for standard read depth (>3) and read depth > 10.

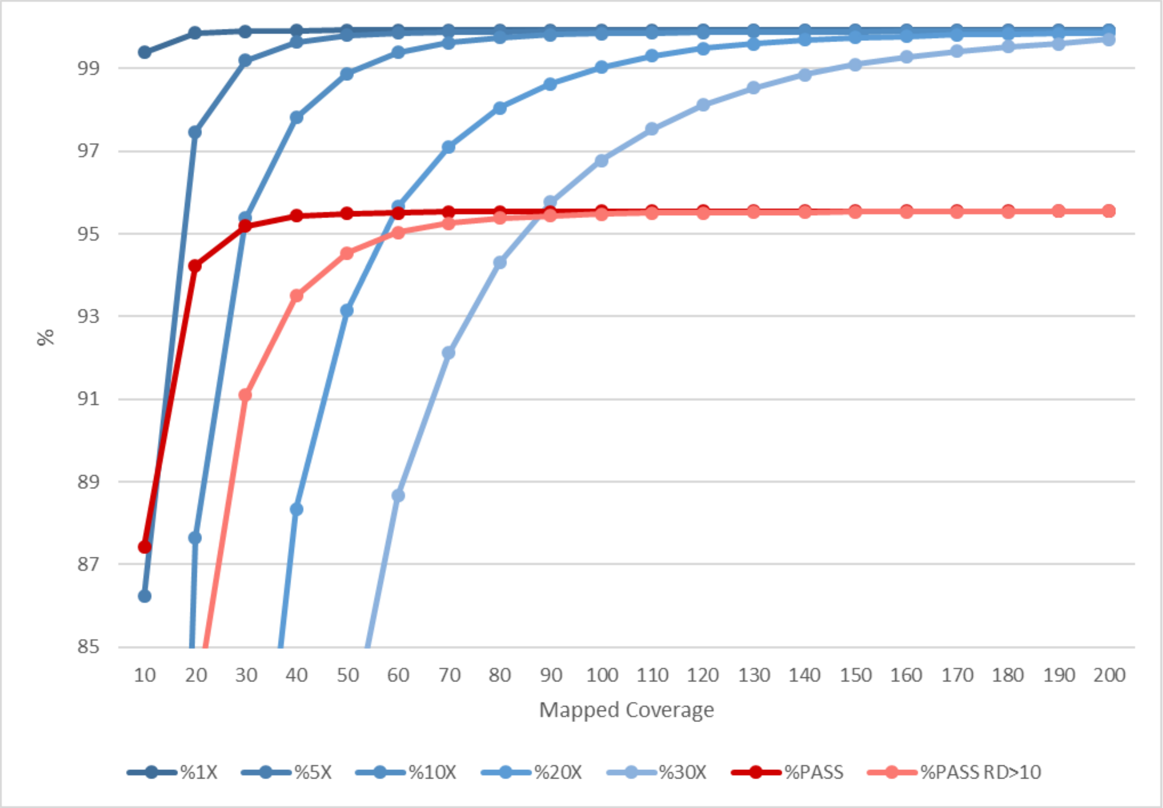


Figure 16. ***Tendency of percentage of covered target and genotypability at different coverage levels.***

Percentage of the target covered by at least 1, 5, 10, 20 and 30 reads (blue lines) and percentage of callable bases on the target for standard read depth (>3) and read depth > 10 (red lines) is shown.

The pipeline for the performance evaluation of WES

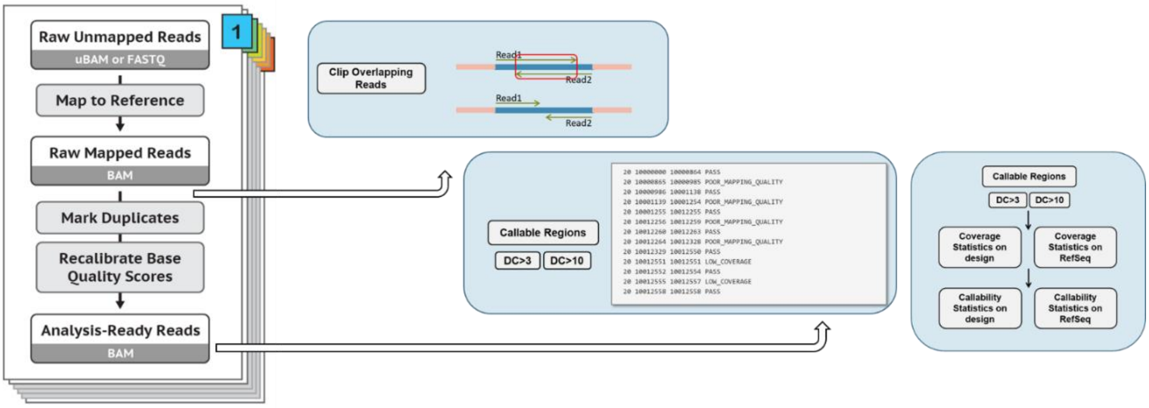
The in-house bioinformatics pipeline used for the data analysis was developed considering the relevance of the genotypability metric revealed by the previous experiment. I integrated the *Genome Analysis Toolkit Best Practices Workflow* (for Germline short variant discovery) with two available Quality Control Tools, for the analysis of WES performance (Figure 17). The first tool, *BamUtil v1.4.14* through the option *clipOverlap*, performed the clipping of overlapping read pairs from the BAM file to avoid counting multiple reads representing the same fragment. The addition of this step to the standard pipeline (before marking the duplicates in the BAM file and recalibrating the base quality scores) was necessary to prevent the identification of false positives in the downstream analysis. Statistics on the overlaps were produced to monitor the number of overlapping pairs and the average number of reference bases overlapped. Then, I used the *CallableLoci* tool (*GATK v3.8)* on the analysis-ready BAM file to identify the regions of the target considered as "callable" (the sites for which the requirements of minimum read depth and minimum quality of mapping were satisfied). This tool considers the coverage at each locus and emits an interval BED file that partitions the genome into different callable states. Only PASS states were kept for the calculation of genotypability. The BED file was produced for two minimum read depths: 3 and 10. When the BED file was produced by *CallableLoci*, it was filtered for CALLABLE regions only, followed by an analysis of regions coverage. I used the option *bedtools coverage (v2.19.1)* to compute both the depth and breadth of coverage of the target regions (the target design and the RefSeq genes).

I then collected all the statistics produced for the evaluation of WES performance:

- percentage of the target (for both the design and RefSeq genes) covered by at least 1, 5, 10, 20, 30 reads (%1X, %5X, %10X, %20X, %30X);
- the genotypability of the target at read depths of 3 (% PASS) and 10 (% PASS RD>10);
- the average insert size;
- the number of mapped deduplicated reads;

- the percentage of duplicates;
- the percentage of on/near/off target bases sequenced;
- the fold enrichment;
- the fold-80 penalty value
- the number of fragments produced by each experiment.

All these values were used for the evaluation of the effects of the DNA fragment extension on the different enrichment platforms.



```
20 10000000 10000864 PASS
20 10000865 10000985 POOR_MAPPING_QUALITY
20 10000986 10001138 PASS
20 10001139 10001254 POOR_MAPPING_QUALITY
20 10001255 10012255 PASS
20 10012256 10012259 POOR_MAPPING_QUALITY
20 10012260 10012263 PASS
20 10012264 10012328 POOR_MAPPING_QUALITY
20 10012329 10012550 PASS
20 10012551 10012551 LOW_COVERAGE
20 10012552 10012554 PASS
20 10012555 10012557 LOW_COVERAGE
20 10012558 10012558 PASS
```

Figure 17. ***Integration of quality control steps with the standard GATK pipeline.*** *DC=Depth of Coverage.*

WES performances on 3 different DNA fragment lengths

I assessed the performance of short (~200 bp), medium (~350 bp) and long (~500 bp) DNA fragments on four major commercial exome enrichment platforms produced by IDT, Roche, Agilent and Twist. For each platform, the wet-lab generated the libraries from the genomic DNA of three unrelated individuals (NA12891, NA12982 and VR00), enriched according to the manufacturers' instructions and sequenced on an Illumina HiSeq3000 instrument.

For the initial dataset of 36 samples, statistics were calculated considering: the total number of sequenced fragments, the GC percentage, the theoretical coverage, the mapped deduplicated reads, the average insert size, the percentage of duplicates and the mapped coverage (Table 2).

ID	Sequenced fragments	GC%	Design length	Theoretical coverage (X)	Mapped deduplicated fragments	Average insert size	% Duplicates	Mapped coverage (X)
NA12891_IDT-S	41,735,851	52	38,871,205	161.05	34,798,458	170.56	14.76	78.02
NA12891_IDT-M	31,721,147	51	38,871,205	244.82	27,295,886	338.36	12.57	95.92
NA12891_IDT-L	35,056,454	52	38,871,205	270.56	29,048,970	419.16	15.62	97.28
NA12891_Roche-S	69,139,041	48	47,007,710	220.62	53,971,159	250.20	19.09	92.04
NA12891_Roche-M	30,139,859	48	47,007,710	192.35	25,187,948	352.58	14.76	73.15
NA12891_Roche-L	37,597,648	48	47,007,710	239.95	31,408,182	475.51	14.05	81.22
NA12891_Agilent-S	69,997,150	51	60,448,148	173.70	56,394,632	267.56	16.11	85.70
NA12891_Agilent-M	43,377,376	50	60,448,148	215.28	36,041,040	350.30	15.74	96.23
NA12891_Agilent-L	62,476,876	49	60,448,148	310.07	50,712,887	438.63	15.70	122.77
NA12891_Twist-S	62,145,209	52	36,715,240	253.89	53,990,842	211.43	9.24	106.91
NA12891_Twist-M	45,101,242	49	36,715,240	368.52	37,418,907	390.49	14.57	106.47
NA12891_Twist-L	56,814,853	49	36,715,240	464.23	48,349,215	389.09	12.24	136.39
NA12892_IDT-S	39,279,677	52	38,871,205	151.58	33,418,710	174.23	12.97	74.92
NA12892_IDT-M	29,737,235	51	38,871,205	229.51	25,829,978	340.75	11.40	90.66
NA12892_IDT-L	31,442,649	52	38,871,205	242.67	26,557,527	422.28	13.69	88.66
NA12892_Roche-S	63,475,661	48	47,007,710	202.55	51,637,279	263.49	15.33	86.49
NA12892_Roche-M	25,113,466	48	47,007,710	160.27	21,737,831	353.95	11.70	63.42
NA12892_Roche-L	35,647,295	48	47,007,710	227.50	29,755,564	483.46	13.90	76.69
NA12892_Agilent-S	63,367,878	50	60,448,148	157.25	50,458,064	270.99	16.64	76.08
NA12892_Agilent-M	38,379,719	50	60,448,148	190.48	32,736,504	357.00	13.32	87.30
NA12892_Agilent-L	59,863,365	49	60,448,148	297.10	49,779,709	446.38	13.80	117.87
NA12892_Twist-S	58,556,655	52	36,715,240	239.23	51,774,402	207.55	7.76	102.74
NA12892_Twist-M	53,142,446	49	36,715,240	434.23	44,835,844	360.36	13.57	131.09
NA12892_Twist-L	53,633,893	49	36,715,240	438.24	46,784,910	411.80	9.39	130.16
VR00_IDT-S	43,858,514	52	38,871,205	169.25	36,430,811	170.05	14.88	81.28
VR00_IDT-M	30,647,544	51	38,871,205	236.53	26,574,858	343.45	11.82	92.66
VR00_IDT-L	28,719,060	51	38,871,205	221.65	24,489,849	429.35	13.02	80.87
VR00_Roche-S	69,341,641	47	47,007,710	221.27	55,349,772	262.24	16.77	80.89
VR00_Roche-M	25,769,903	48	47,007,710	164.46	22,210,538	361.45	11.86	64.00
VR00_Roche-L	37,393,464	47	47,007,710	238.64	31,375,345	482.07	13.60	80.81
VR00_Agilent-S	63,423,698	51	60,448,148	157.38	51,058,660	265.11	17.83	77.59
VR00_Agilent-M	37,984,977	50	60,448,148	188.52	31,756,515	354.10	15.15	84.83
VR00_Agilent-L	34,450,041	49	60,448,148	170.97	28,964,914	439.13	13.13	69.93
VR00_Twist-S	58,092,508	52	36,715,240	237.34	51,158,376	209.79	8.01	101.11
VR00_Twist-M	49,286,947	49	36,715,240	402.72	42,067,556	360.41	12.36	121.92
VR00_Twist-L	53,025,771	48	36,715,240	433.27	46,216,132	400.91	9.92	128.25

Table 2. ***WES initial dataset.***

For each replicate, platform and DNA fragment length combination, the number of sequenced fragments, percentage GC content, theoretical coverage, number of mapped fragments without duplicates, average insert size, percentage of reads marked as duplicates and mapped coverage on the target are shown. DNA fragment lengths: S = short, M = medium, L = long.

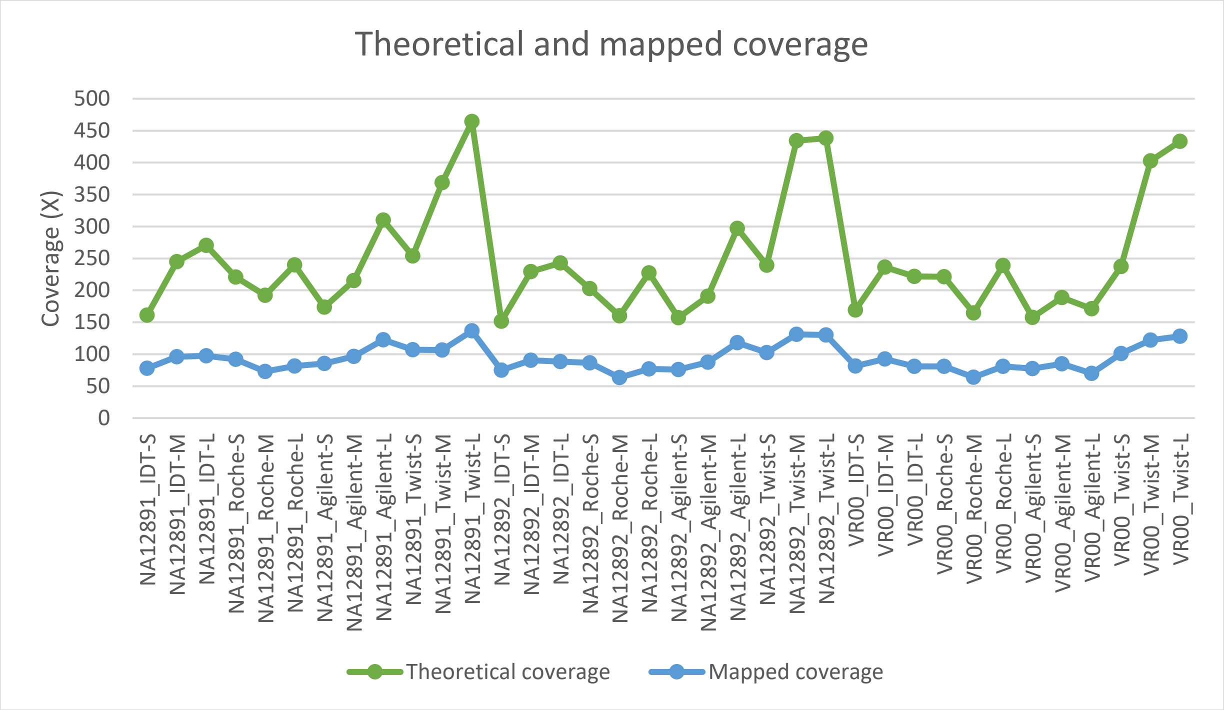


Figure 18. ***Theoretical and mapped coverage.***

For each replicate, platform and DNA fragment length, theoretical coverage and mapped coverage on the target are plotted.

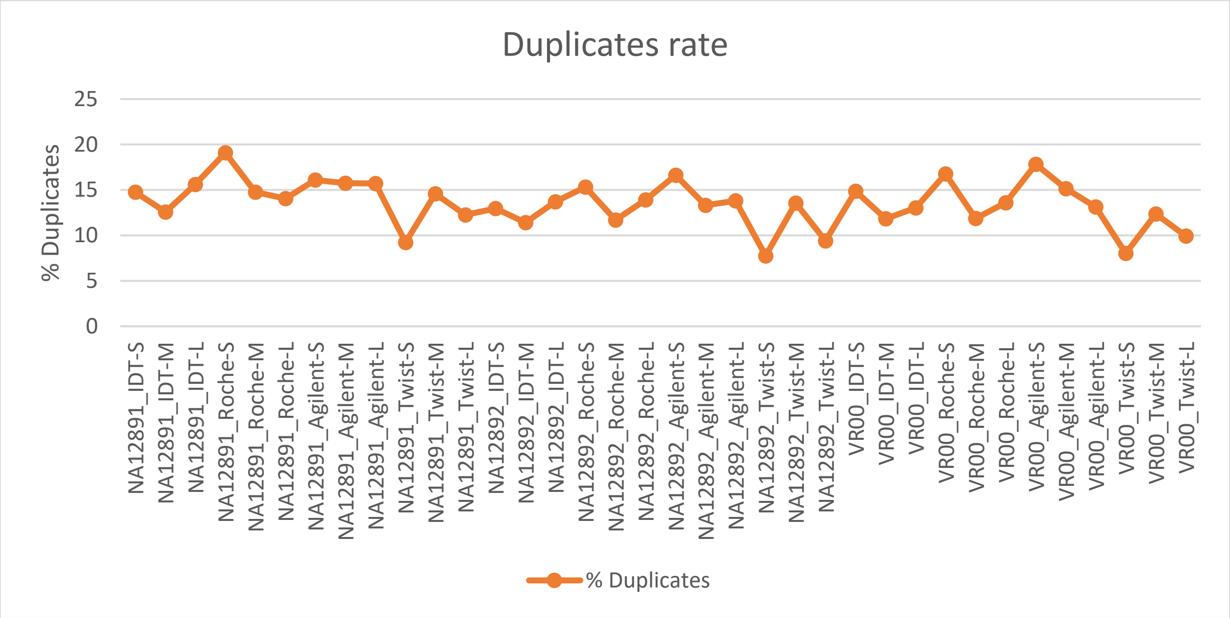


Figure 19. ***Duplicates rate.***

For each replicate, platform and DNA fragment length, the duplicates rate is plotted.

The dataset showed many differences at the level of fragments produced and consequently in the number of mapped deduplicated reads. Theoretical coverage varied from 151X to 464X, and this was reflected in the mapped coverage (63-136X) (Figure 18). The percentage of duplicates (reads sequenced from the same fragment) was also very variable (7-19%) (Figure 19). To avoid a biased comparison of statistics values, due to the difference in coverage levels and target regions considered, the dataset was firstly aggregated by the mean of the three independent experiments (VR00, NA12891, NA12892), and then subdivided in normalized datasets (Figure 20). These were used to evaluate WES performances for each combination of DNA fragment length and enrichment platform at different conditions.

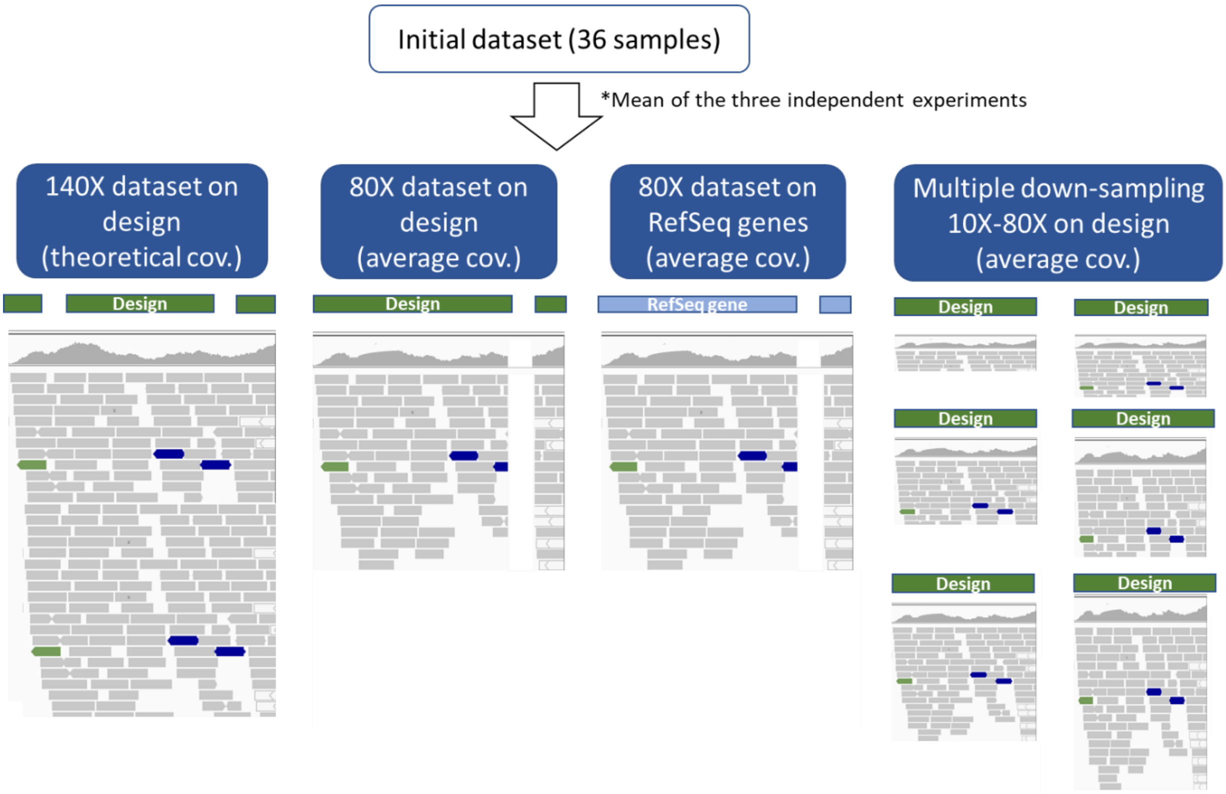

Initial dataset (36 samples)
*Mean of the three independent experiments
140X dataset on design (theoretical cov.)
80X dataset on design (average cov.)
80X dataset on RefSeq genes (average cov.)
Multiple down-sampling 10X-80X on design (average cov.)
Design
RefSeq gene

Figure 20. ***Different datasets used for the calculation of WES performances.***

The 140X dataset

From the initial dataset of sequenced reads representing each sample, I produced downsampled BAM files with a 140 theoretical X-fold coverage (the maximum theoretical coverage value obtained by all the platforms) on the target design (Table 3). The theoretical coverage is a computed coverage based on randomly subsampled sequenced reads, whose amount is set considering both the length of the reads and the length of the target region. Theoretical coverage allows the evaluation of:

- the on/near/off target rate, as the subsampled reads are randomly selected considering the entire genome
- the fold enrichment and the FOLD 80 penalty, which both provide information about the read distribution
- the percentage of duplicates, which are calculated and removed after mapping the reads on the target region

The achieved average insert sizes were firstly evaluated, as these values were used for the calculation of the near-target rate. The short and medium fragment lengths obtained were as expected, whereas the long fragments were often shorter than anticipated (398–480 bp). Then, for each combination of enrichment platform and DNA fragment length, the 140X dataset allowed the estimation of the near and off target rate obtained. The number of bases sequenced near the target augmented with the increase of the DNA fragment size, whereas the off-target rate showed different trends between the platforms. The evaluation of the number of duplicates showed that it was consistently higher using the short DNA fragments for all platforms (12-15%) except Twist, which had almost comparable values for the short and medium size (~5%). The variability in the number of duplicates and in the near/off-target rates determined a difference in the mapped coverage obtained, which could affect the evaluation of other important statistics values such as the genotypability and the enrichment uniformity. For this reason, a normalization on the mapped coverage was conducted.

ID	Average insert size	% Duplicates	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	% ON TARGET	% NEAR TARGET	% OFF TARGET	Fold enrichment	FOLD 80 penalty
IDT-S	171.61	12.61	69.38	99.82	99.77	99.61	98.14	92.96	96.81	96.66	60.36	29.42	10.22	50.01	1.60
IDT -M	340.85	7.39	57.92	99.81	99.64	98.84	92.37	79.64	97.58	96.72	48.95	40.63	10.43	40.56	1.95
IDT -L	423.60	8.60	54.28	99.80	99.32	97.01	85.53	70.18	97.39	94.83	45.15	44.10	10.75	37.41	2.28
Roche-S	258.64	11.81	60.10	99.86	99.36	98.33	93.53	83.01	96.17	95.02	51.86	18.13	30.01	35.50	1.89
Roche-M	355.99	10.60	55.76	99.83	99.27	98.03	91.86	79.02	96.90	95.53	49.33	38.52	12.14	33.77	1.90
Roche-L	480.35	8.75	50.31	99.80	99.11	97.37	87.91	71.05	96.79	94.84	42.28	36.04	21.68	28.94	2.02
Agilent-S	267.89	14.89	70.25	99.79	99.33	98.36	94.21	85.84	95.23	94.22	61.57	21.20	17.23	32.85	2.04
Agilent-M	353.80	10.88	66.15	99.75	99.33	98.49	94.48	85.58	96.27	95.40	57.04	26.17	16.79	30.44	1.96
Agilent-L	441.38	11.48	58.31	99.74	99.16	97.73	90.82	78.15	96.13	94.62	50.92	36.96	12.13	27.17	2.06
Twist-S	209.67	5.10	61.80	99.86	99.78	99.33	95.85	89.38	95.51	95.06	52.23	33.16	14.61	45.77	1.59
Twist-M	368.28	5.26	46.41	99.82	99.71	99.21	94.96	83.24	96.57	96.06	38.92	45.65	15.43	34.11	1.47
Twist-L	398.16	3.60	45.17	99.82	99.69	99.05	94.00	80.94	96.56	95.90	37.24	47.04	15.72	32.63	1.51

Table 3. ***The 140X dataset.***

For each platform and DNA fragment length combination, the 140 theoretical X-fold coverage is shown for the target design dataset (mean of the three independent experiments). The columns show the average insert size, percentage of reads marked as duplicates, mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, percentage of bases on/near/off target, fold enrichment and FOLD 80 base penalty. DNA fragment lengths: S = short, M = medium, L = long.

The 80X dataset (on design)

I produced downsampled BAM files with a 80X-fold mapped coverage (the maximum mapped coverage value obtained by all the platforms) on the target design region (Table 4). The mapped coverage was calculated as the real average coverage obtained on the ROI, excluding the duplicates. This dataset focused only on the reads mapping on the design region, as everything mapping near and off the expected captured region was not intended to be analysed. As parameters for comparison I considered: coverage of the target at different thresholds (1X, 5X, 10X, 20X, 30X), genotypability (% PASS and % PASS RD>10), the fold enrichment and the FOLD 80 values. I evaluated the enrichment uniformity of each DNA fragment/enrichment platform combination obtained using the FOLD 80 penalty value (the fold over-coverage necessary to raise 80% of bases to the mean coverage level in those targets). The increase of the DNA fragments influenced the uniformity of enrichment: longer fragments decreased the FOLD 80 value in one platform (Twist), while in two others increased (IDT and Roche). The genotypability for each platform improved for the medium and long DNA fragments for both % PASS and % PASS RD>10 values.

ID	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
IDT-S	78.04	99.83	99.77	99.67	98.72	95.26	96.81	96.71	49.99	1.60
IDT-M	80.00	99.82	99.72	99.45	97.01	90.60	97.62	97.32	40.49	1.93
IDT-L	80.29	99.82	99.63	98.89	93.99	85.11	97.55	96.73	37.28	2.28
Roche-S	80.30	99.88	99.55	98.96	96.71	91.89	96.29	95.61	35.46	1.86
Roche-M*	66.83	99.85	99.40	98.54	94.63	86.06	96.99	96.01	33.74	1.91
Roche-L	79.52	99.85	99.49	98.83	95.99	89.65	97.04	96.30	28.82	2.01
Agilent-S	77.73	99.65	99.15	98.25	94.79	88.17	97.58	96.58	32.79	2.05
Agilent-M	80.01	99.64	99.30	98.76	96.26	90.63	98.23	97.64	30.29	1.95
Agilent-L	76.42	99.66	99.27	98.55	94.99	87.51	98.20	97.39	26.82	2.05
Twist-S	80.00	99.86	99.81	99.65	97.94	94.22	95.52	95.36	45.67	1.58
Twist-M	79.96	99.84	99.78	99.70	99.08	97.11	96.58	96.51	33.52	1.42
Twist-L	80.05	99.84	99.78	99.69	98.93	96.66	96.58	96.50	32.17	1.45

Table 4. ***The 80X mapped dataset.***

For each platform and DNA fragment length combination, the 80 mapped X-fold coverage is shown for the target design dataset (mean of the three independent experiments). The columns show the mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty. DNA fragment lengths: S = short, M = medium, L = long.

** The sequencing data available for this combination did not reach 80X mapped coverage.*

The 80X dataset (on RefSeq genes)

The same calculation was performed for the RefSeq genes using the downsampled BAM files at 80X mapped coverage on the target designs. This dataset focused only on the reads mapping on the genes regions (defined by the RefSeq database), since according to the literature the clinical interest is mostly on the protein-coding part of the genome. Results of the influence of DNA fragment length on the genotypability of each gene showed similar trends to those described above (Table 5). The higher genotypability was obtained using the medium and the long DNA fragments for all platforms.

ID	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
IDT-S	78.43	99.16	99.02	98.93	98.15	94.96	95.72	95.63	50.41	1.60
IDT-M	79.36	99.22	99.03	98.73	96.21	89.55	96.59	96.27	40.31	1.95
IDT-L	79.61	99.26	98.96	98.17	93.07	83.84	96.54	95.68	37.10	2.31
Roche-S	82.94	99.80	99.48	98.99	97.31	93.82	96.21	95.67	36.62	1.74
Roche-M*	69.35	99.77	99.36	98.69	95.77	89.05	96.98	96.23	35.01	1.81
Roche-L	83.25	99.79	99.44	98.93	96.82	91.89	97.03	96.47	30.18	1.92
Agilent-S	86.86	99.76	99.44	98.91	96.60	91.58	96.28	95.73	36.63	2.01
Agilent-M	89.27	99.72	99.39	98.97	97.20	93.02	97.03	96.60	33.80	1.94
Agilent-L	86.23	99.74	99.37	98.79	96.06	90.20	96.99	96.37	30.26	2.08
Twist-S	79.50	99.81	99.75	99.60	97.90	94.18	96.18	96.04	45.39	1.58
Twist-M	79.96	99.80	99.73	99.67	99.09	97.19	97.13	97.07	33.51	1.41
Twist-L	80.08	99.81	99.73	99.65	98.94	96.74	97.14	97.06	32.18	1.44

Table 5. ***The 80X mapped dataset (RefSeq genes).***

For each platform and DNA fragment length combination, the 80 mapped X-fold coverage is shown for the target design dataset (mean of the three independent experiments). The columns show the mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty. DNA fragment lengths: S = short, M = medium, L = long.

** The sequencing data available for this combination did not reach 80X mapped coverage.*

I then focused the analysis on the number of RefSeq genes which could reach 100% genotypability in all the platforms using different DNA fragment lengths (Table 6).

In three of the four platforms, the medium length obtained more 100% callable genes, as Roche obtained the highest value using the long DNA fragments. The best result was obtained by Twist-M (17,709 genes), while the worst value was obtained by Agilent-S (16,053 genes), with a difference of 1656 genes.

Enrichment platform	**Average DNA fragment size**		
	Short (S)	**Medium (M)**	**Long (L)**
IDT	16,430	16,823	16,144
Roche	16,091	16,299	16,599
Agilent	16,053	16,869	16,547
Twist	16,812	17,709	17,706

Table 6. ***Number of RefSeq genes reaching 100% genotypability.***
Number of RefSeq genes reaching 100% genotypability at 80X mapped coverage on the target design dataset using different platforms and DNA fragment lengths.

Aggregating the results of increased genotypability for each platform, many RefSeq genes could reach 100% genotypability from short-to-medium and short-to-long fragment extension (840-1330) (Table 7 and Figure 21). Considering the genes which showed any increase in genotypability, the number was even higher (1837-2429). For a minimal number of genes, a decrease in genotypability was observed increasing the DNA fragment length.

Dataset	IDT	Roche	Agilent	Twist
RefSeq genes – up to 100% genotypability	840	1007	1330	1107
RefSeq genes – increased genotypability	1837	2247	2429	1993
OMIM genes – up to 100% genotypability	156	125	270	232
OMIM genes – increased genotypability	321	288	459	370

Table 7. ***Number of genes showing increased genotypability.***

Number of RefSeq and OMIM genes showing increased genotypability following the extension of the DNA fragment size from short to medium, or short to long, at 80X mapped coverage on each target design.

The genes associated with a clinical phenotype (derived from the OMIM database) were investigated to analyse the improvement in genotypability through the extension of the DNA fragment length. 125-270 OMIM genes could reach 100% genotypability from short-to-medium and short-to-long fragment extension, and considering the genes which showed any increase in genotypability, the number achieved was 288-459. As seen before, for a minimal number of genes there was a decrease in genotypability with the increase of the DNA fragment length.

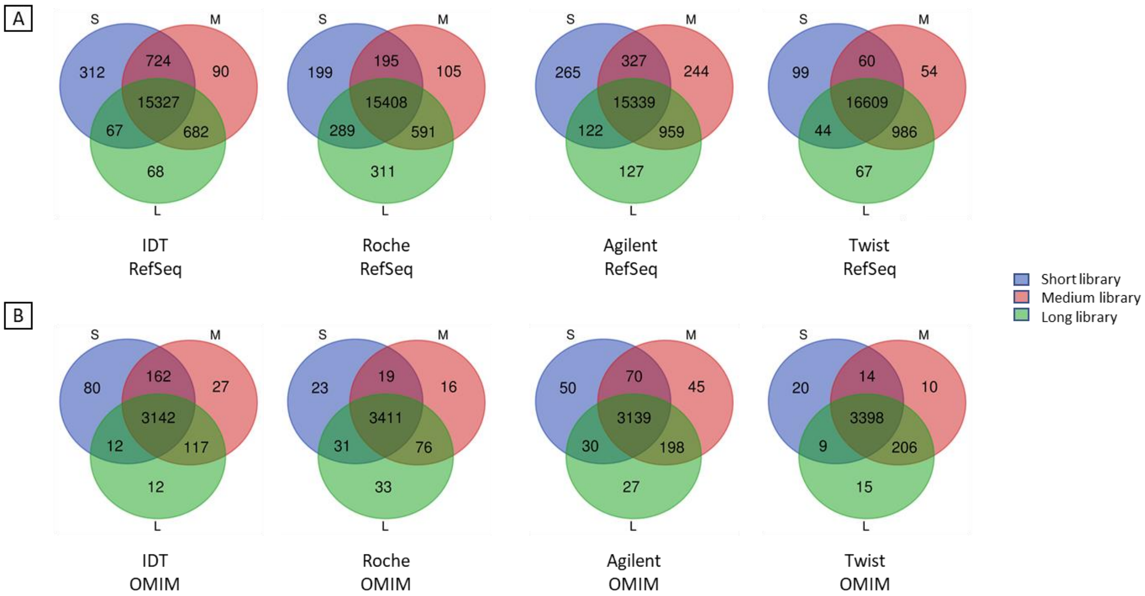


Figure 21. ***RefSeq/OMIM genes reaching 100% genotypability.***

Number of RefSeq (A) and OMIM (B) genes reaching 100% genotypability at 80X mapped coverage on each target design using different DNA fragment lengths.

I then ranked by improvement in genotypability the OMIM genes and took the top 20 considering both the improvements between the short and the medium size and between the short and the long size (Table 8). The difference between short and longer fragments showed that, at same coverage levels, the genotypability of the target region could increase up to 53%.

OMIM	% Genotypability			% Diff.	%10X Coverage		
	S	M	L		S	M	L
RPS26	47.13	100	100	52.87	100	100	100
RPL15	49.98	100	100	50.02	99.90	100	100
RPL21	60.60	100	100	39.4	100	100	100
RPSA	63.29	100	100	36.71	100	100	100
GCSH	64.56	100	97.38	35.44	100	100	100
HNRNPA1	66.84	100	100	33.16	100	100	100
CISD2	53.37	85.15	100	31.78	100	100	100
IFNL3	69.43	100	100	30.57	100	100	100
LEFTY2	74.00	100	100	26.00	100	100	100
BMPR1A	74.19	100	100	25.81	100	100	100
RPS23	75.00	100	100	25.00	100	100	100
ISCA1	75.13	100	100	24.87	100	100	100
ALG10	75.15	100	100	24.85	100	100	100
IFITM3	77.53	100	100	22.47	100	100	100
PTEN	78.55	100	100	21.45	98.49	100	100
BANF1	78.64	100	100	21.36	100	100	100
HLA-A	79.02	100	100	20.98	99.88	100	99.82
RPS28	80.79	100	100	19.21	100	100	100
RP9	78.88	97.60	100	18.72	100	100	100
CYP11B1	81.73	100	100	18.27	100	100	100

Table 8. ***Top 20 OMIM genes showing the best improvement in genotypability.***

Top 20 OMIM genes showing the best improvement in genotypability following the extension of the DNA fragment length from short to medium and short to long (Twist enrichment platform). The data represent the maximum difference in genotypability at 80X mapped coverage on the Twist design. DNA fragment lengths: S = short, M = medium, L = long.

The multiple-downsampling dataset

Finally, I evaluated the influence at different coverage levels of the DNA fragment size and enrichment uniformity on genotypability of the target. Therefore, I produced downsampled BAM files (with an average X-fold coverage of 10–80) on the corresponding target designs. Since coverage levels are considered fundamental in the evaluation of WES performances, I compared: coverage of the target at different thresholds (1X, 5X, 10X, 20X, 30X), genotypability (% PASS and % PASS RD>10), the fold enrichment and the FOLD 80 values through a variable mapped coverage.

I initially focused on the single effect of enrichment uniformity on genotypability at 10-80X mapped coverage. I performed a comparison between the platform with the best enrichment uniformity (lower FOLD 80 value), Twist, and the one with the highest FOLD 80 value, Agilent, considering a fixed DNA fragment length (medium) (Table 9). Twist with its higher uniformity (1.42-1.58) could reach saturation of genotypability at 60X mapped coverage (96.57% for % PASS and 96.40% for % PASS RD>10), while Agilent with higher FOLD 80 values (1.94-2.39) could not reach the same callability (96.31% and 96.40% for % PASS and % PASS RD>10, respectively, at 80X mapped coverage). The percentage of the target covered by at least 30 reads (%30X) reflected the higher uniformity of Twist already at 60X in respect of Agilent.

Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
Twist-M									
80	99.84	99.78	99.70	99.08	97.11	96.58	96.51	33.52	1.42
70	99.84	99.77	99.66	98.66	95.67	96.57	96.47	33.51	1.42
60	99.83	99.76	99.58	97.87	93.03	96.57	96.40	33.52	1.42
50	99.83	99.73	99.40	96.32	87.50	96.55	96.21	33.51	1.44
40	99.82	99.68	98.94	92.62	74.75	96.53	95.77	33.51	1.46
30	99.81	99.49	97.55	81.67	47.28	96.45	94.39	33.51	1.50
20	99.79	98.64	91.45	47.98	11.39	96.02	88.26	33.51	1.58
10	99.54	89.28	49.25	3.28	0.38	90.86	46.27	33.51	1.58
Agilent-M									
80	99.78	99.45	98.86	96.41	90.93	96.31	95.72	30.41	1.94
70	99.76	99.37	98.62	95.16	87.38	96.27	95.50	30.41	1.95
60	99.75	99.27	98.24	93.00	81.79	96.21	95.16	30.41	1.96
50	99.73	99.09	97.54	89.08	73.10	96.12	94.51	30.41	1.97
40	99.69	98.78	96.10	81.57	59.87	95.96	93.13	30.41	1.98
30	99.63	98.05	92.40	67.22	40.83	95.56	89.53	30.41	2.00
20	99.47	95.47	80.96	41.50	17.69	94.11	78.27	30.42	2.13
10	98.63	80.04	43.34	8.65	1.84	84.16	41.31	30.41	2.39

Table 9. ***Downsampled mapped coverage.***

Parameters were calculated on 10–80X downsampled sets, including mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty.

Then I focused on the single effect of the DNA fragment size on genotypability at 10-80X mapped coverage. I performed a comparison between two different DNA fragment lengths (short and long) on the same platform (IDT), which showed a very high variability in enrichment uniformity using different fragment lengths (1.60-1.86 for S and 2.26-2.35 for L) (Table 10). Again, the %30X reflected the FOLD 80 values based on the use of the short and the long lengths. In particular, the long fragments produced a higher number of over-represented regions at 10-20X mapped coverage in respect of the short fragments. On the contrary, at higher mapped coverage levels (40-80X) the %10X was lower for the long fragments, suggesting an uneven distribution of longer reads. With regard to genotypability, longer fragments achieved higher values already at 40X, but the % PASS RD>10 did not performed as well (91.32% for IDT-L against 95.61% for IDT-S at 40X mapped coverage). In this case, the lower enrichment uniformity of IDT-L negatively effected the genotypability of longer DNA fragments.

Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
IDT-S									
80	99.88	99.84	99.73	98.77	95.48	96.84	96.72	50.05	1.60
70	99.83	99.77	99.62	98.18	93.15	96.81	96.66	49.98	1.60
60	99.82	99.75	99.51	96.94	88.62	96.80	96.54	49.99	1.61
50	99.82	99.72	99.25	94.25	80.45	96.79	96.29	49.99	1.60
40	99.81	99.64	98.57	88.18	65.63	96.74	95.61	49.99	1.62
30	99.80	99.38	96.34	73.74	40.61	96.62	93.38	49.98	1.64
20	99.77	98.04	86.96	41.40	11.46	95.95	84.03	49.99	1.69
10	99.42	85.07	43.50	4.75	1.72	88.31	40.93	49.99	1.86
IDT-L									
80	99.82	99.63	98.89	93.98	85.10	97.56	96.73	37.28	2.28
70	99.81	99.56	98.44	91.65	80.56	97.51	96.27	37.28	2.26
60	99.80	99.44	97.67	88.17	74.47	97.45	95.50	37.28	2.28
50	99.79	99.20	96.28	82.84	66.15	97.31	94.08	37.28	2.34
40	99.77	98.65	93.54	74.46	54.62	96.99	91.32	37.29	2.34
30	99.72	97.23	87.75	60.97	38.36	96.14	85.47	37.28	2.34
20	99.54	92.90	74.56	38.95	17.37	93.33	72.26	37.28	2.35
10	98.21	74.64	40.47	8.48	2.36	79.91	38.37	37.29	2.35

Table 10. ***Downsampled mapped coverage.***

Parameters were calculated on 10–80X downsampled sets, including mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty.

Finally, I evaluated the combined effect of DNA fragment extension with enrichment uniformity on genotypability at 10-80X mapped coverage. I compared different DNA fragment lengths (short and medium) on the platform which showed the best enrichment uniformity (1.42-1.58) with longer fragments (Twist) (Table 11). Results showed that both the DNA fragments could reach saturation of genotypability already at 60X mapped coverage, but with a substantial difference of 1% more for the medium fragments for % PASS (96.57% for Twist-M against 95.50% for Twist-S) and % PASS RD>10 (96.40% for Twist-M against 95.01% for Twist-S). Therefore, the combined effect of higher enrichment uniformity and extension of DNA fragment length led to better genotypability, especially for clinically-relevant thresholds.

Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
Twist-S									
80	99.86	99.81	99.65	97.94	94.22	95.52	95.36	45.67	1.58
70	99.86	99.79	99.52	97.02	92.09	95.51	95.25	45.68	1.58
60	99.86	99.77	99.28	95.56	88.66	95.50	95.01	45.67	1.60
50	99.85	99.71	98.80	93.13	82.65	95.48	94.53	45.67	1.60
40	99.84	99.56	97.74	88.39	70.86	95.41	93.50	45.67	1.62
30	99.83	99.09	95.29	77.32	47.59	95.18	91.09	45.67	1.66
20	99.78	97.36	87.52	47.91	13.60	94.22	83.48	45.67	1.71
10	99.30	86.10	48.97	4.46	0.59	87.35	45.76	45.67	1.88
Twist-M									
80	99.84	99.78	99.70	99.08	97.11	96.58	96.51	33.52	1.42
70	99.84	99.77	99.66	98.66	95.67	96.57	96.47	33.51	1.42
60	99.83	99.76	99.58	97.87	93.03	96.57	96.40	33.52	1.42
50	99.83	99.73	99.40	96.32	87.50	96.55	96.21	33.51	1.44
40	99.82	99.68	98.94	92.62	74.75	96.53	95.77	33.51	1.46
30	99.81	99.49	97.55	81.67	47.28	96.45	94.39	33.51	1.50
20	99.79	98.64	91.45	47.98	11.39	96.02	88.26	33.51	1.58
10	99.54	89.28	49.25	3.28	0.38	90.86	46.27	33.51	1.58

Table 11. ***Downsampled mapped coverage.***

Parameters were calculated on 10–80X downsampled sets, including mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty.

Variant calling results

To assess the effects of the improvement in genotypability through the combination of the DNA fragment extension and a high enrichment uniformity, the variant calling was performed on each of the three individuals. The variant calling could evaluate the difference in the number of variants identified using short and longer DNA fragments due to the higher number of callable bases achieved in all platforms.

For each sample, I used the HaplotypeCaller software (v4.1.2.0) to identify the genetic variants in respect of the human genome reference sequence. Variants were filtered using the target design regions of Twist, which showed the best results in terms of genotypability of the target after the extension of the DNA fragments, and results were aggregated by the mean values obtained from the individuals (Table 12). Results showed an increase of >1% in both the short-to-medium and short-to-long fragment extensions. The same >1% increase with longer DNA fragments was observed for the number of variants identified in the RefSeq and OMIM genes included in the target design. These results reflected the same trend seen for the genotypability (1% increase) in genotypability achieved by increasing the length of the DNA fragments.

DNA fragment size	#variants in design	#variants in RefSeq genes	#variants in OMIM Genes
S	23,140	20,279	5,008
M	23,461	20,509	5,057
L	23,521	20,576	5,074

Table 12. ***Variants in the Twist target design (HaplotypeCaller).***
Total number of variants identified in the Twist target design, and in the corresponding RefSeq and OMIM genes, for each DNA fragment size (S = short, M = medium, L = long).

WES performances on 9 different DNA fragment lengths

The evaluation of the WES performance for 3 different DNA fragment lengths (~200, ~350 and ~500 bp) showed relevant differences in terms of genotypability of the ROI. The highest change in genotypability was identified between the short and the medium fragments, whose difference in terms of length was not trivial (150 bp). To identify the presence of a threshold above which the genotypability of the ROI could significantly improve, I isolated 27 individuals from almost 1,000 exomes processed with the Twist platform in over a year. These samples could represent nine different DNA fragment lengths (200, 230, 260, 270, 280, 290, 340, 360 and 400 bp) in replicates of three. The developed bioinformatics pipeline was then used to analyse this new set of samples.

For the initial dataset of 27 samples, statistics were calculated considering: the total number of sequenced fragments, the GC percentage, the theoretical coverage, the mapped deduplicated reads, the average insert size, the percentage of duplicates and the mapped coverage (Table 13).

The dataset showed many differences at the level of fragments produced and in the number of mapped deduplicated reads, as previously seen. Theoretical coverage varied from 198X to 464X, whereas the mapped coverage varied from 80 to 136X (Figure 22). The percentage of duplicates was also very variable (7-20%) (Figure 23). To compute an unbiased comparison of WES performances, the dataset was aggregated by the mean values obtained for each group of individuals, and then subdivided in normalized datasets using the theoretical coverage (200X) and the mapped coverage (80X) on the target design, as performed previously.

ID	Sequenced fragments	GC%	Design length	Theoretical coverage (X)	Mapped deduplicated fragments	Average insert size	% Duplicates	Mapped coverage (X)
VR00_200	58,092,508	52	36,715,240	237.34	51,158,376	209.79	8.01	101.11
NA12891_200	62,145,209	52	36,715,240	253.89	53,990,842	211.43	9.24	106.91
NA12892_200	58,556,655	52	36,715,240	239.23	51,774,402	207.55	7.76	102.74
VR00_230	48,873,712	51	36,715,240	199.67	42,612,354	222.17	9.03	82.13
NA12891_230	48,612,772	52	36,715,240	198.61	41,948,721	236.17	9.57	80.02
NA12892_230	53,294,072	52	36,715,240	217.73	45,120,197	244.94	11.22	85.21
3234V_260	38,894,499	49	36,715,240	317.81	35,348,300	261.83	8.42	93.92
2852T_260	35,460,051	49	36,715,240	289.74	32,054,884	261.88	8.91	85.38
3258V_260	55,593,344	49	36,715,240	454.25	50,254,880	263.55	8.98	132.45
376V_270	51,391,218	49	36,715,240	419.92	46,352,045	269.72	9.09	122.61
3260V_270	38,768,558	49	36,715,240	316.78	35,472,484	270.95	7.72	95.64
3233V_270	37,135,966	49	36,715,240	303.44	33,726,602	272.39	8.38	90.85
603V_280	39,062,497	49	36,715,240	319.18	35,586,582	282.06	8.08	94.62
19N0175_280	44,493,206	48	36,715,240	363.55	40,179,822	282.23	8.91	107.87
3269V_280	42,372,135	48	36,715,240	346.22	38,497,147	281.65	8.39	103.36
377V_290	45,368,031	48	36,715,240	370.70	41,185,643	288.65	8.49	110.41
19N0104_290	37,116,772	49	36,715,240	303.28	33,657,874	291.56	8.39	89.46
3778V_290	35,789,884	48	36,715,240	292.44	32,646,350	292.83	7.86	86.80
VR00_340	52,398,002	48	36,715,240	428.14	41,262,820	323.46	20.01	108.73
NA12891_340	48,348,995	48	36,715,240	395.06	38,860,722	338.35	18.23	102.00
NA12892_340	51,378,070	48	36,715,240	419.81	41,175,544	339.21	18.42	107.62
VR00_360	49,286,947	49	36,715,240	402.72	42,067,556	360.41	12.36	121.92
NA12891_360	45,101,242	49	36,715,240	368.52	37,418,907	390.49	14.57	106.47
NA12892_360	53,142,446	49	36,715,240	434.23	44,835,844	360.36	13.57	131.09
VR00_400	53,025,771	48	36,715,240	433.27	46,216,132	400.91	9.92	128.25
NA12891_400	56,814,853	49	36,715,240	464.23	48,349,215	389.09	12.24	136.39
NA12892_400	53,633,893	49	36,715,240	438.24	46,784,910	411.80	9.39	130.16

Table 13. ***WES initial dataset.***

For each replicate and DNA fragment length combination, the number of sequenced fragments, percentage GC content, theoretical coverage, number of mapped fragments without duplicates, average insert size, percentage of reads marked as duplicates and mapped coverage on the target are shown.

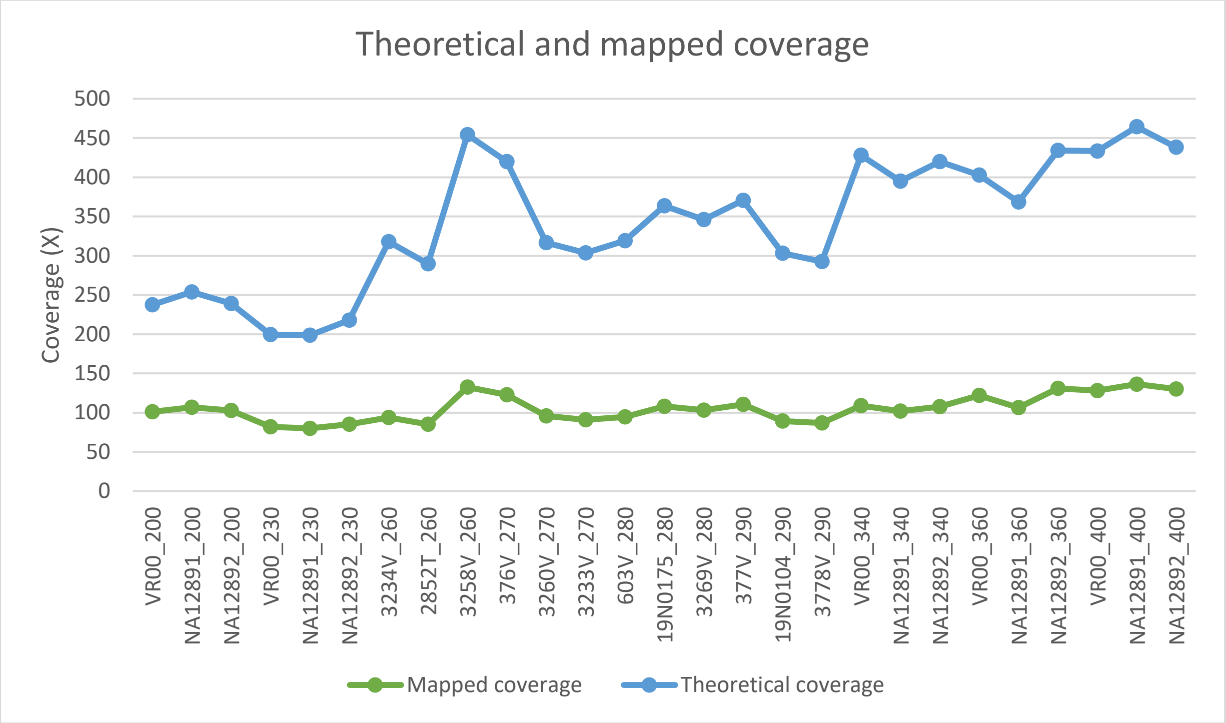


Figure 22. ***Theoretical and mapped coverage.***

For each replicate and DNA fragment length, theoretical coverage and mapped coverage on the target are plotted.

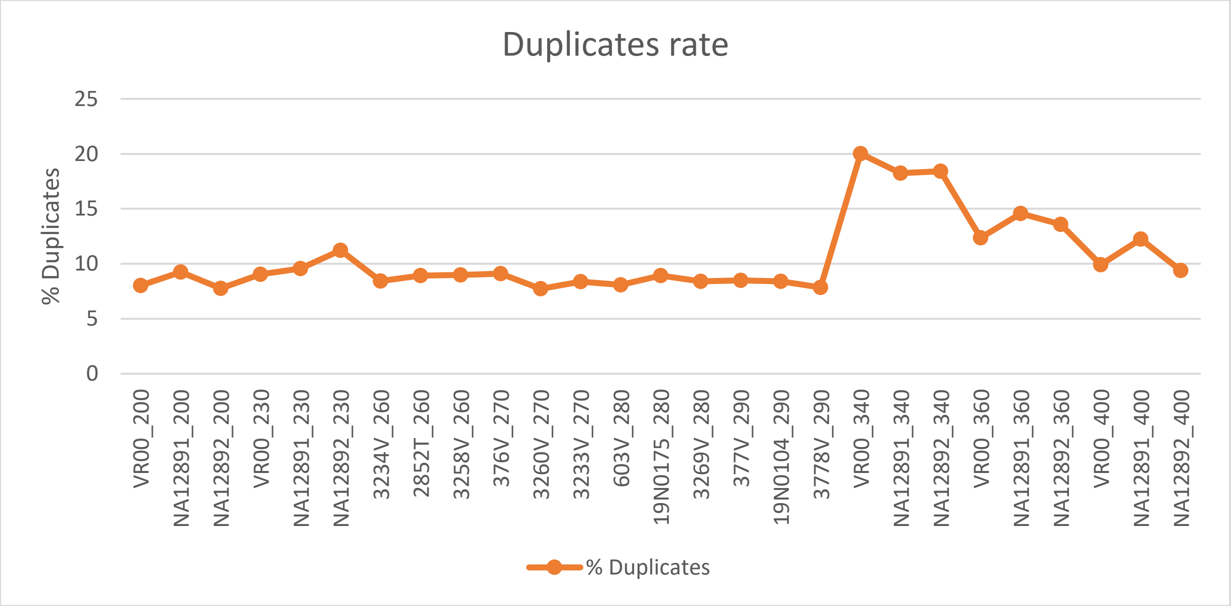


Figure 23. ***Duplicates rate.***

For each replicate and DNA fragment length, the duplicates rate is plotted.

The 200X dataset

From the initial dataset of sequenced reads representing each sample, I produced downsampled BAM files with a 200 theoretical X-fold coverage (the maximum mapped coverage value obtained by all the samples) on the target design (Table 14). As parameters for comparison I considered: the on/near/off target rate, the fold enrichment, the FOLD 80 values and the percentage of duplicates.

DNA fragment length	Average insert size	% Duplicates	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	% ON TARGET	% NEAR TARGET	% OFF TARGET	Fold enrichment	FOLD 80 penalty
200	206.28	6.29	87.30	99.86	99.82	99.70	98.36	95.22	95.50	95.39	52.36	32.96	14.68	45.89	1.58
230	231.64	9.24	80.90	99.86	99.82	99.71	98.40	95.17	95.53	95.43	49.78	35.58	14.63	43.63	1.53
260	262.34	5.31	61.15	99.80	99.72	99.64	99.32	97.15	96.38	96.33	40.20	36.52	23.28	35.23	1.36
270	270.91	5.08	61.86	99.78	99.71	99.64	99.34	97.33	96.41	96.36	39.73	37.21	23.06	34.82	1.36
280	281.83	5.11	61.81	99.78	99.69	99.61	99.25	97.03	96.42	96.35	38.70	37.43	23.87	33.92	1.38
290	290.87	5.32	61.37	99.79	99.69	99.61	99.24	96.87	96.43	96.35	37.89	38.05	24.06	33.21	1.38
340	334.18	10.16	57.26	99.83	99.74	99.63	98.96	94.90	96.54	96.44	36.35	40.44	23.20	31.86	1.38
360	366.62	7.18	64.74	99.83	99.76	99.60	98.16	94.11	96.58	96.43	38.80	45.65	15.56	34.00	1.44
400	396.33	4.94	63.50	99.83	99.75	99.56	97.78	93.14	96.59	96.39	37.16	47.03	15.81	32.57	1.48

Table 14. ***The 200X dataset.***

For each DNA fragment length, the 200 theoretical X-fold coverage is shown for the target design dataset (mean of the three independent experiments). The columns show the average insert size, mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, percentage of bases on/near/off target, fold enrichment and FOLD 80 base penalty.

I evaluated the number of sequenced bases near and off the target and the frequency of duplicates obtained. The extension of the DNA fragment length from 200 to 400 bp generally decreased the on-target rate, except for the 360 and the 400 bp length. As expected, the near-target increased from the short to longer libraries, while the off-target rate showed lower values only for very short (200-230 bp) or longer fragments (360-400 bp), reflecting a higher cross-hybridization to regions outside of the target for fragments between 260 and 340 bp. The highest frequency of duplicates was generated by the 230 and 340 bp length (9–10%), followed by 200 and 360 bp (6-7%). On the contrary, fragment lengths between 260 and 290 bp, followed by the longest fragment size (400 bp) seemed to optimize the frequency of duplicates for this platform. As previously observed, the differences in off-target and duplicates rates resulted in a variability in the mapped coverage values.

The 80X dataset (on design and RefSeq genes)

The same calculation was performed for the design target regions and the RefSeq genes, using the downsampled BAM files at 80X mapped coverage on the target designs (Table 15 and 16). As parameters for comparison I considered: coverage at different thresholds (1X, 5X, 10X, 20X, 30X), genotypability (% PASS and % PASS RD>10), the fold enrichment and the FOLD 80 values. Results confirmed the trend previously seen. Genotypability increased when optimizing uniformity of enrichment and increasing DNA fragment length, especially between the 260 and the 340 bp DNA fragment length. For the design target region, % PASS jumped from 95.52% to 96.38% using the 230 bp and 260 bp lengths, respectively. FOLD 80 decreased from 1.58 to 1.34, concomitantly. Enrichment uniformity started to decrease from the 360 bp length (1.42-1.45), which already obtained saturation in terms of genotypability.

DNA fragment length	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
200	80.72	99.86	99.82	99.71	98.42	95.21	95.55	95.45	43.50	1.53
230	80.00	99.86	99.81	99.65	97.94	94.22	95.52	95.36	45.67	1.58
260	79.99	99.81	99.73	99.68	99.54	99.13	96.38	96.34	35.13	1.34
270	79.97	99.79	99.72	99.67	99.54	99.13	96.40	96.37	34.72	1.34
280	80.00	99.79	99.71	99.65	99.48	99.01	96.43	96.38	33.79	1.36
290	80.02	99.80	99.71	99.65	99.49	99.01	96.43	96.38	33.10	1.36
340	79.99	99.84	99.78	99.71	99.50	98.88	96.53	96.47	31.59	1.37
360	79.96	99.84	99.78	99.70	99.08	97.11	96.58	96.51	33.52	1.42
400	80.05	99.84	99.78	99.69	98.93	96.66	96.58	96.50	32.17	1.45

Table 15. ***The 80X mapped dataset.***

For each DNA fragment length, the 80 mapped X-fold coverage is shown for the target design dataset (mean of the three independent experiments). The columns show the mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty.

Similar results were obtained for the RefSeq genes. Genotypability increased at lower FOLD 80 values, with a leap between the 230 bp and the 260 bp length (% PASS jumped from 95.81% to 96.58%, respectively). Above the 230 bp length, the FOLD 80 decreased (1.58-1.34), together with an increase of the genotypability. Enrichment uniformity started to decrease from the 360 bp length (1.42-1.44), which again obtained saturation in terms of genotypability.

DNA fragment length	Mapped coverage (X)	%1X	%5X	%10X	%20X	%30X	% PASS	% PASS RD>10	Fold enrichment	FOLD 80 penalty
200	80.18	99.71	99.63	99.53	98.23	95.02	95.84	95.74	43.21	1.52
230	79.45	99.71	99.62	99.46	97.74	94.00	95.81	95.65	45.36	1.58
260	79.59	99.69	99.54	99.49	99.37	98.99	96.58	96.54	34.95	1.34
270	79.55	99.68	99.54	99.49	99.37	98.99	96.60	96.56	34.54	1.34
280	79.99	99.69	99.53	99.47	99.32	98.92	96.63	96.57	33.78	1.35
290	79.84	99.70	99.54	99.47	99.33	98.89	96.64	96.58	33.02	1.37
340	79.71	99.75	99.60	99.54	99.36	98.79	96.75	96.68	31.48	1.37
360	79.83	99.72	99.61	99.53	98.94	97.02	96.81	96.73	33.46	1.42
400	79.94	99.73	99.61	99.52	98.79	96.56	96.82	96.72	32.13	1.44

Table 16. ***The 80X mapped dataset.***

For each DNA fragment length, the 80 mapped X-fold coverage is shown for the RefSeq genes dataset (mean of the three independent experiments). The columns show the mapped coverage on the target, percentage of the target covered by at least 1, 5, 10, 20 and 30 reads, percentage of callable bases on the target for standard read depth (>3) and read depth >10, fold enrichment and FOLD 80 base penalty.

The trend obtained for the FOLD 80 and the genotypability values (% PASS and % PASS RD>10) was plotted against the respective DNA fragment lengths (Figure 20). The enrichment uniformity showed a decrease (increase in the curve) above the 260 bp length, indicating that longer DNA fragments did not optimize the coverage uniformity, for both the design and RefSeq genes regions (Figure 24 A-B). However, longer DNA fragments increased the genotypability for both % PASS and % PASS RD>10 values, as uniformity of coverage was sufficient to allow good performances in base calling (Figure 24 C-D). In particular, a remarkable leap (more than 1%) of genotypability was evident between the 230 bp and the 260 bp

DNA fragment length. From the 260 bp length, results showed a less evident increase in genotypability, indicating that satisfactory base calling could be obtained already at this threshold. The decrease in enrichment uniformity observable from the 360 bp length (which was related to an increase in the genotypability of a few percentage points), pointed out that base calling was not strictly influenced by the FOLD 80 value, but more by the DNA fragment length.

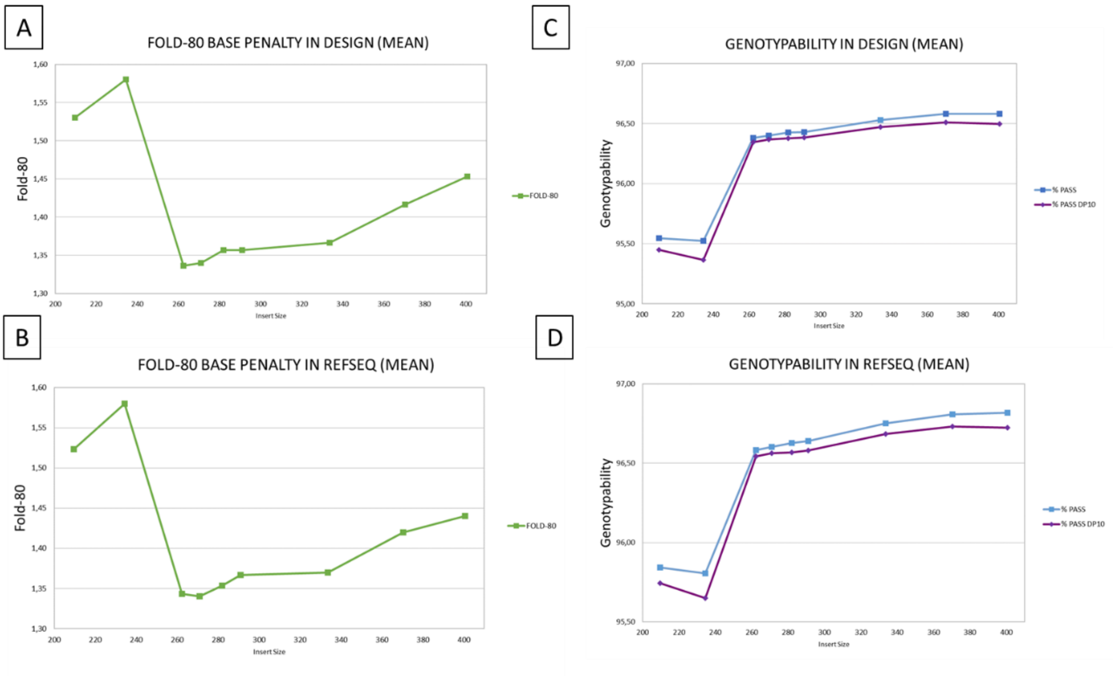


Figure 24. ***Tendency of FOLD 80 penalty and genotypability at different insert sizes.*** *The figure shows the FOLD 80 base penalty and genotypability values at different insert sizes considering the target design (A-C) and the RefSeq genes (B-D) regions.*

DISCUSSION

The current challenges for the bioinformatics data analysis of NGS data are several, from the correct alignment of the sequenced reads on the reference genome to the accurate variant calling. Software benchmarking could help in the decision of finding the most accurate pipeline to use [28][29][30], but some limitations are still present. Alignment of reads in repetitious regions of the genome yet leads to segments of the genome not investigable [8] and variant calling provides an incomplete genetic information (VCF files), as invariant sites cannot be further analysed. Moreover, data derived from samples processed using different enrichment technologies are analysed considering only the depth and uniformity of coverage [16], whereas regions highly-covered cannot always ensure a high confidence of the alignment. This work aimed to improve the bioinformatics analysis of targeted resequencing and subsequently ameliorate the current limitations of the analysis.

Considering that the solely depth of coverage could be misleading for the data analysis, a new metric was included, namely the genotypability. This value, which reflects both the depth of coverage and the quality of the alignment at a specific genomic site, was calculated integrating the *CallableLoci* tool in the *The Genome Analysis Toolkit Best Practices Workflow*. In this way, callable regions of any ROI (i.e. the design or the coding regions of the genome) could be provided, so that the consequent variant calling could produce adequate files for a potential clinical setting (gVCF files). Several standard parameters generally used for the assessment of WES performance were also integrated within the pipeline, however genotypability could show a substantial difference from other metrics, especially from the depth of coverage. From a single initial experiment sequenced at very high coverage levels (200X), results showed that whereas optimal coverage levels on the entire target could be obtained only a high mapped coverage, genotypability increased to reach saturation at coverage levels usually considered low for WES [31]. For this reason, contrary to what is expected, deeper sequencing would be useless as the callability would not improve at higher coverage levels.

While the lack of sequencing in some regions could be ameliorated increasing the depth of coverage, the genotypability clearly showed a different trend. For this reason, there was a need to improve the alignment of reads at low quality, as the base calling is directly influenced by that. The approach was then to extend the DNA fragment length so that one of the read pairs could align outside of the exonic regions to reach the introns, known to be under greater evolutionary constraints [32]. Therefore, the bioinformatics pipeline developed was applied on three biological replicates sequenced using different enrichment technologies (as they provide a variability with regards to the target region) and extending the DNA fragment to a medium (∼350 bp) and a long (∼500 bp) size.

Overall, independently from the ROI (enrichment platform), the extension of the DNA fragment size could always provide a better genotypability of the target. Thus, an improvement of the read alignment could effectively be obtained. Indeed, many genes derived from the RefSeq gene dataset at 80X mapped coverage improved their mappability, including genes of known clinical interests. This result was relevant in light of the challenges posed by the American College of Medical Genetics and Genomics (ACMG), that already stressed the importance of detecting disease-causing variants in repetitious regions of the genome [2]. Among the genes that improved their genotypability, RPS26 and RPL15 (associated with the bone marrow disorder Diamond-Blackfan anemia according to the OMIM database) obtained 100% base calling extending the DNA fragment from short to medium (starting from a callability of 47.13% for RPS26 and 49.98% for RPL15). RPSA, which is associated with the immunodeficiency disease isolated congenital asplenia, also reached 100% from 63.29% extending the DNA fragments, and similarly the tumor suppressor gene PTEN could also obtain 100% (starting from 78.55% using short fragments). This means that these genes of medical relevance contained regions of low mapping quality that could potentially harbour pathogenic variants otherwise neglected.

This result was confirmed with the variant calling performed on the replicates, which on average provided an increase in the number of variants of ∼1%, the same increase showed for the genotypability metric. This proved again an effective improvement of the read alignment in regions previously considered uncallable, but also the presence of a consistent number of variants present in repetitious genomic regions.

With regards to uniformity of coverage, the extension of the DNA fragments could not clearly determine an improvement, as previously stated [6]. Indeed, for some enrichment platforms, longer fragments could improve the coverage uniformity, while for others there was no improvement. More generally, with a low uniformity of coverage, genotypability was more dependent on the mapped coverage (higher coverage = higher genotypability), but with higher uniformity values, the genotypability reached saturation at lower coverage levels (60X). This result confirmed once more that with a more uniform coverage of the target, deep sequencing is not necessary, as genotypability could not further improve. The fold enrichment value, which provides the "efficiency of enrichment" through the evaluation of the on-target in respect of the near and off-target rates, did not strongly correlate with the genotypability of the target as well. Indeed, low fold enrichment values did not correspond to a reduction of callability.

The number of duplicates obtained for each enrichment platform and DNA fragment size combination confirmed as well the importance of the extension of the DNA fragment length for the reduction of the sequencing depth, related to the reduction of the number of fragments that need to be produced. Indeed, as the use of 2 x 75 bp reads requires double the amount of sequencing of the 2 x 150 bp reads, the problem of duplicates could be greatly reduced.

Overall, the most relevant change in genotypability was observed between the short (~200) and the medium (~350) fragments. This indicated that short fragments could successfully improve the read alignment when extended, but the minimum extension required to generally overcome repetitious regions was still not known. For this reason, nine more different DNA fragment lengths were investigated between the short and the medium one. The increase in genotypability between the different lengths was continuous, but an evident leap was present between the 230 and the 260 bp length. The same leap was evident in the uniformity of coverage (from 1.58 to 1.34), showing once again that higher enrichment uniformity lead to better genotypability. However, while uniformity slightly started to decrease using DNA fragments above 260 bp, the genotypability could still improve. This indicated that uniformity of coverage had an influence on genotypability of the region, but the major influencing factor was the extension of the DNA fragment length. The 1% increase in genotypability was obtained immediately above the 230 bp, pointing out that this threshold should be used as a minimum requirement for the library preparation of samples analysed with the specified enrichment platform. Genotypability could still slightly improve, leaving to the single laboratory the choice of the more appropriate DNA length to use. Interestingly, the number of bases sequenced on-target decreased when increasing the fragment size between 260 to 340 bp, but genotypability was not affected by that. This proved that fold enrichment and uniformity of coverage are still incomplete metrics for the evaluation of WES performances.

Exome sequencing costs could also be reduced through the extension of the DNA fragment length. While short DNA fragments generally allows to limit the costs of the analysis, longer fragments could improve the quality of the read alignment, producing a higher uniformity of coverage and hence reducing the amount of sequencing needed to sufficiently cover the entire target region. In this way, the overall costs could be reduced and DNA fragment extension revealed to be less costly than the increase of the sequencing depth.

In this thesis work, the performance of WES was evaluated through a metric that considered not only the depth and uniformity of coverage of the region investigated, but also the quality of the read alignment. Genotypability confirmed to be a more informative parameter in the evaluation of WES, and this could be improved extending the DNA fragment length. Although the combination of DNA fragment size and enrichment platform showed an influence on the base calling, this one improved in all cases, even despite slightly worsening of performance of the uniformity of coverage. The use of this approach in a clinical setting could provide to clinicians the best options from the sample processing to the variant calling, even for repetitious genomic regions. The identification of more variants in regions difficult to align could provide new insights into human diseases and their associations to variants with a biological consequence.

REFERENCES

[1] G. Goh and M. Choi, “Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases,” *Genomics Inform.*, vol. 10, no. 4, p. 214, 2012.

[2] B. Quintáns, A. Ordóñez-Ugalde, P. Cacheiro, A. Carracedo, and M. J. Sobrido, “Medical genomics: The intricate path from genetic variant identification to clinical interpretation,” *Appl. Transl. Genomics*, vol. 3, no. 3, pp. 60–67, 2014.

[3] C. Di Resta, S. Galbiati, P. Carrera, and M. Ferrari, “Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities,” *Electron. J. Int. Fed. Clin. Chem. Lab. Med.*, vol. 29, no. 1, pp. 4–14, 2018.

[4] M. Kumaran, U. Subramanian, and B. Devarajan, “Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019.

[5] Z. Wang, X. Liu, B. Z. Yang, and J. Gelernter, “The role and challenges of exome sequencing in studies of human diseases,” *Front. Genet.*, vol. 4, no. August, pp. 1–8, 2013.

[6] B. Rabbani, M. Tekin, and N. Mahdieh, “The promise of whole-exome sequencing in medical genetics,” *J. Hum. Genet.*, vol. 59, no. 1, pp. 5–15, 2014.

[7] Y. Sun *et al.*, “Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome?,” *Hum. Mutat.*, vol. 36, no. 6, pp. 648–655, 2015.

[8] M. L. Metzker, “Sequencing technologies the next generation,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010.

[9] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, “PEAR: A fast and accurate Illumina Paired-End reAd mergeR,” *Bioinformatics*, vol. 30, no. 5, pp. 614–620, 2014.

[10] N. Whiteford *et al.*, “An analysis of the feasibility of short read sequencing,” *Nucleic Acids Res.*, vol. 33, no. 19, pp. 1–6, 2005.

[11] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, “Sequencing depth and coverage: Key considerations in genomic analyses,” *Nat. Rev. Genet.*, vol. 15, no. 2, pp. 121–132, 2014.

[12] X. Bian *et al.*, “Comparing the performance of selected variant callers using synthetic data and genome segmentation,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–11, 2018.

[13] G. Yang, C. Sau, W. Lai, J. Cichon, and W. Li, “Next-generation sequencing in the clinic: Promises and challenges,” *Cancer Lett.*, vol. 344, no. 6188, pp. 1173–1178, 2015.

[14] F. Mertes *et al.*, “Targeted enrichment of genomic DNA regions for next-generation sequencing,” *Brief. Funct. Genomics*, vol. 10, no. 6, pp. 374–386, 2011.

[15] C. S. Ku, D. N. Cooper, and G. P. Patrinos, “The Rise and Rise of Exome Sequencing,” *Public Health Genomics*, vol. 19, no. 6, pp. 315–324, 2017.

[16] C. Pommerenke *et al.*, “Enhanced whole exome sequencing by higher DNA insert lengths,” *BMC Genomics*, vol. 17, no. 1, pp. 1–8, 2016.

[17] M. Choi *et al.*, “Genetic diagnosis by whole exome capture and massively parallel DNA sequencing,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 45, pp. 19096–19101, 2009.

[18] Q. Wang, C. S. Shashikant, M. Jensen, N. S. Altman, and S. Girirajan, “Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, 2017.

[19] S. B. Ng *et al.*, “Targeted capture and massively parallel sequencing of 12 human exomes,” *Nature*, vol. 461, no. 7261, pp. 272–276, 2009.

[20] Y. Hasin-Brumshtein, M. C. M. Ramirez, L. Arbiza, and R. Zeitoun, “The Importance of Coverage Uniformity Over On-Target Rate for Efficient Targeted NGS,” *Twist Biosci.*, vol. 61, no. 3, pp. 610–612, 1966.

[21] L. Y. Ballester, R. Luthra, R. Kanagal-Shamanna, and R. R. Singh, “Advances in clinical next-generation sequencing: Target enrichment and sequencing technologies,” *Expert Rev. Mol. Diagn.*, vol. 16, no. 3, pp. 357–372, 2016.

[22] M. K. Sakharkar, V. T. K. Chow, and P. Kangueane, “Distributions of exons and introns in the human genome,” *In Silico Biol.*, vol. 4, no. 4, pp. 387–393, 2004.

[23] S. Gudlaugsdottir, D. R. Boswell, G. R. Wood, and J. Ma, “Exon size distribution and the origin of introns,” *Genetica*, vol. 131, no. 3, pp. 299–306, 2007.

[24] S. R. Head *et al.*, “Library construction for next-generation sequencing: Overviews and challenges,” *Biotechniques*, vol. 56, no. 2, pp. 61–77, 2014.

[25] M. T. W. Ebbert *et al.*, “Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches,” *BMC Bioinformatics*, vol. 17, no. Suppl 7, 2016.

[26] “PANINI Website.” [Online]. Available: https://www.birmingham.ac.uk/generic/panini/index.aspx.

[27] G. A. Van der Auwera *et al.*, *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*, no. SUPL.43. 2013.

[28] J. Shang, F. Zhu, W. Vongsangnak, Y. Tang, W. Zhang, and B. Shen,

“Evaluation and comparison of multiple aligners for next-generation sequencing data analysis,” *Biomed Res. Int.*, vol. 2014, 2014.

[29] S. Laurie *et al.*, “From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing,” *Hum. Mutat.*, vol. 37, no. 12, pp. 1263–1271, 2016.

[30] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, “Systematic comparison of variant calling pipelines using gold standard personal exome variants,” *Sci. Rep.*, vol. 5, no. December, pp. 1–8, 2015.

[31] S. W. Kong, I. H. Lee, X. Liu, J. N. Hirschhorn, and K. D. Mandl, “Measuring coverage and accuracy of whole-exome sequencing in clinical context,” *Genet. Med.*, vol. 20, no. 12, pp. 1617–1626, 2018.

[32] E. V. Koonin and Y. I. Wolf, “Constraints and plasticity in genome and molecular-phenome evolution,” *Nat. Rev. Genet.*, vol. 11, no. 7, pp. 487–498, 2010.

ACKNOWLEDGEMENTS

Per questa tesi, vorrei ringraziare innanzitutto il Professor Delledonne, che in questi anni mi ha guidato nell'ambito della ricerca insegnandomi a sviluppare un'animata curiosità, dedita a non dare mai nessuna informazione per scontata. La passione che mette ogni giorno nel suo lavoro sarà sempre per noi del gruppo un esempio da seguire.

Ringrazio tutti i membri del laboratorio di Functional Genomics: lavoratori instancabili e uniti. In particolare ringrazio: Marzia Rossato, per essere per tutti noi una seconda guida, per avere avuto sempre una parola di conforto anche nei momenti più difficili e per avermi spesso spronato ad essere la versione migliore di me stessa; il Dott. Salviati, per averci insegnato ogni giorno qualcosa col sorriso; Emanuela Cosentino, una delle persone più esuberanti e di cuore che io abbia mai conosciuto, per aver reso specialissimi questi anni di dottorato; Luca Marcolungo, per tutti gli attenti suggerimenti che mi hanno aiutato a crescere, sia lavorativamente che umanamente; Simone Maestri, per tutti i confronti costruttivi ma anche per tutte le risate nei momenti in cui non ci restava che piangere; e Denise Lavezzari, per il costante supporto e la pazienza in tutti i momenti di panico per le deadline dell'ultimo minuto (su BAM e varianti notturne!). Ringrazio inoltre tutti i bioinformatici che hanno lavorato con noi in passato, in particolare Luciano Xumerle e Marianna Garonzi, per avermi insegnato tanto nei primi anni di dottorato e con cui si è formato un legame speciale. In più, ringrazio tutti i membri del laboratorio di Personal Genomics, passati e presenti, per tutto il lavoro svolto insieme e per le persone belle che mi sono state a fianco (in particolare Maria Giovanna, Carla, Anna Scandalo, AE ed Elisa).

Thank you to the PANINI consortium, which is an amazing group of people with so much passion for their research and that taught me a lot. Also, thank you to the Professor van Riel and Andrea Cabbia, for the warm hospitality at the University of Eindhoven during my secondment and the very interesting collaboration that we had created together.

Ringrazio la mia Famiglia: mia madre, mio padre, le mie sorelle e le mie nonne, che mi son sempre stati vicino e di supporto in tutto quello che ho fatto nella mia vita. In particolare, dedico a mia madre questo lavoro, che mi ha sempre dato un'immensa forza per non mollare mai.

Ai miei amici di sempre, un ringraziamento dal cuore per tutto il sostegno. Tutta la mia Modern Family: Daniele, Armando, Claudio, Beppe e Vanzi, per avermi in qualche modo guidato fino a qui; Vì, per il sostegno che ormai da 13(!) anni non mi fa mai mancare; e la Simo, la voce "per" la mia coscienza e la mia roccia di sempre e per sempre.

A tutti voi un grazie enorme, così come a tutte le persone che ho conosciuto e che mi hanno donato un po' di loro stesse e a cui io ho donato un po' di me stessa.