UNIONE EUROPEA
Fondo sociale europeo

REGIONE AUTONOMA DELLA SARDEGNA

POR SARDEGNA
obiettivo competitività regionale e occupazione

Università degli Studi di Cagliari

Dipartimento di Scienze Mediche e Sanità Pubblica

DOTTORATO DI RICERCA

Medicina Molecolare e Traslazionale

Ciclo XXX

TITOLO TESI

DEVELOPMENT OF A BIOINFORMATIC PIPELINE

FOR NEXT-GENERATION SEQUENCING DATA ANALYSIS:

Application to circulating cell-free fetal DNA

Settori scientifico disciplinari di afferenza

BIO/11

Presentata da:	Matteo Massidda
Coordinatore Dottorato	Prof. Amedeo Columbano
Tutor	Prof.ssa Maria Cristina Rosatelli

Esame finale anno accademico 2016 – 2017
Tesi discussa Febbraio 2018

Matteo Massidda gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

Señor Jesús,

Me amas tanto que me da vergüenza.

[Kiko Argüello. Anotaciones. 2016]

ABSTRACT

Prenatal diagnosis (PD) of single gene disorders and aneuploidies is currently carried out through chorionic villus sampling (CVS) and amniocentesis, two invasive procedures which have an associated risk of fetal loss of 0.5-2%.

The discovery, in 1997, that cell-free fetal DNA (cffDNA) circulates in maternal plasma during gestation, has offered an alternative source of fetal genetic material which can be collected by non invasive approach.

In a study aimed at developing a protocol for Non Invasive Prenatal Diagnosis of β-thalassemia, we have developed a bioinformatic pipeline to infer the fetal genotype by next generation sequencing of cffDNA.

The approach consists in target sequencing of a region extending 63Kb in the β-globin gene cluster, including the causative mutation (β039) and SNPs with high heterozygosity, which are used for parental haplotypes construction.

In the first part of our protocol the DNAs from parents and fetus (obtained from CVS) are sequenced in order to identify informative SNPs which are then sequenced in the corresponding cffDNA sample.

The pipeline firstly analyzes sequencing data from parents and constructs the parental haplotypes based on a reference panel composed by the 1000 Genomes Reference Panel merged with sequencing data of 39 Sardinian TRIOs (father, mother, child).

Secondly, the fetal genotype prediction is carried out by an algorithm that infers the most likely inherited paternal and maternal alleles and, through Viterbi algorithm, reconstructs the possible haplotype blocks inherited by the fetus.

Finally, the pipeline predicts the fetal HBB genotype and validates the results comparing it with the corresponding CVS sample genotype.

We processed 39 cffDNA samples, however in 7 of them the pipeline could not complete the prediction because of the lack of informative sites or low fetal fraction. In 24 out of 30 samples analyzed by the pipeline (80%) the fetal genotype was correctly determinated.

In the remaining 6 samples (20%) the analysis returned only the correct paternal inherited haplotype determination. In general, obtained results are encouraging and confirm that NIPD is also feasible in couples who are at risk for a monogenic disorder and share the same variant.

CONTENTS

1. BACKGROUND

1.1 NEXT-GENERATION SEQUENCING

The publication of the first drafts of the human genome in 2001 as a result of two successful projects by Collins (International Human Genome Sequencing Consortium, 2001) and Venter (Venter JC, 2001) represents a milestone in the field of DNA sequencing.

A large number of research groups from many countries worked for 13 years to reconstruct the whole human genome sequence.

The project started in 1990 and was completed in 2003 with the filling of gaps in the sequence (International Human Genome Sequencing Consortium, 2004).

Despite the scientific relevance of the obtained results, the high cost and the large amount of time spent on this project highlighted the need for new sequencing technologies which should be cost effective and time saving, also allowing for the parallelization of a great number of samples.

In this way, the workflow of Sanger sequencing to generate read lengths up to 1000 bp with per-base accuracy of 99.999%, represented a limit for sequencing large numbers of human genomes (Shendure J J. H., 2008).

Nevertheless, after dominating the DNA sequencing for approximately 40 years, Sanger sequencing still represents a valid method and is included in many diagnostic workflows.

The first signs of what might revolutionize the sequencing market appeared in 2005 with the landmark publication of the sequencing-by-synthesis technology developed by 454 Life Sciences Corporation (Margulies M, 2005).

The newly developed sequencing system was based on the implementations of cyclic-array strategies described by Shendure and Margulies groups, which were practical and cost competitive with conventional sequencing (Shendure J e. a., Accurate multiplex polony sequencing of an evolved bacterial genome, 2005) (Margulies M, 2005).

Cyclic-array sequencing allows the simultaneous decoding of millions to billions of distinct features located in a two-dimensional array (Shendure J e. a., Overview of DNA Sequencing Strategies, 2011).

In each cycle an enzymatic process is applied to interrogate the identity of the base at a particular position in each feature for all features in parallel.

The enzymatic process is coupled to a detection system of the incorporated nucleotide, which can be the production of light, the measurement of H+ ions release or the incorporation of a fluorescent group, depending on the sequencing platform (Quail MA, 2012).

After multiple cycles, a contiguous sequence for each feature is constructed after incorporation signal decoding and sequences are stored in output files.

This principle is the basic paradigm for several different Next Generation Sequencing platforms, which presents several advantages over Sanger sequencing.

Firstly, the parallelization of multiple features in a single sequencing run provided by NGS resulted in a large number of generated sequences (several gigabases) in comparison with Sanger's capillary sequencing limited output, reducing the sequencing time.

Secondly, NGS instruments do not require a cloning step. Sanger sequencing library construction requires DNA fragments cloning into plasmids, a process that can be completed in approximately a week. By contrast, NGS library preparation can be performed in a few days. DNA to be sequenced is fragmented and synthetic adapters are ligated at each fragment end.

These adapters are universal sequences, specific to each platform, that can be used to clonally amplify the library fragments.

Amplification is required to provide sufficient signal strength to enable the detection of incorporated nucleotides.

The two most commonly used methods are emulsion PCR and solid-phase bridge PCR.

In emulsion PCR (Dressman D, 2003), single strand library fragments are diluted and compartmentalized in water droplets in a water-in-oil emulsion.

Ideally, the dilution is to a degree where each droplet contains a single template molecule and functions as a micro-PCR reactor. The reactors contain a bead covered with complementary adapters, binding the fragment to be amplified, and all amplification required reagents such as deoxynucleotides (dNTPs), primers and DNA polymerase.

After PCR, each bead is covered by thousands of copies of the bound DNA fragment.

Conversely, bridge PCR (Fedurco M, 2006) is performed on a solid surface such as a flow cell.

Library fragments randomly bind to adapters-complementary primers immobilized on the support.

The free end of each fragment can interact with the nearby primers determining a “bridge” structure and so allowing template strand amplification.

After several PCR cycles, the solid surface will present 100-200 million spatially separated template clusters.

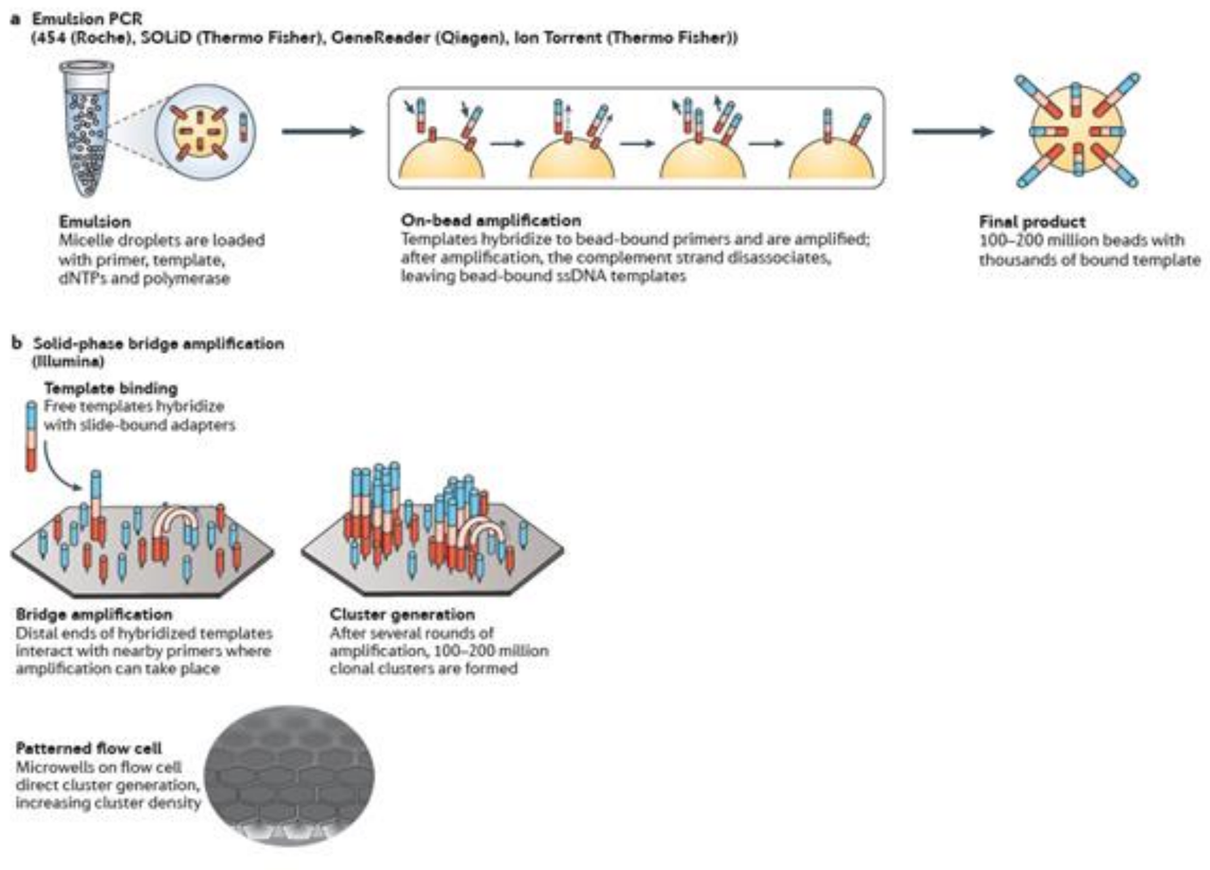

a Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))
Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase
On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates
Final product
100–200 million beads with thousands of bound template
b Solid-phase bridge amplification (Illumina)
Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place
Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed
Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density

Figure 1. **Template amplification strategies. a)** In emulsion PCR, fragmented DNA templates are ligated to adapter sequences and are captured in an aqueous droplet (micelle) along with a bead covered with complementary adapters, deoxynucleotides (dNTPs), primers and DNA polymerase. PCR is carried out within the micelle, covering each bead with thousands of copies of the same DNA sequence. **b)** In solid-phase bridge amplification, fragmented DNA is ligated to adapter sequences and bound to a primer immobilized on a solid support, such as a patterned flow cell. The free end can interact with other nearby primers, forming a bridge structure. PCR is used to create a second strand from the immobilized primers, and unbound DNA is removed. *(Goodwin S, 2016)*

These innovative library amplification strategies are included respectively in Ion Torrent (Rothberg JM, 2011) and Illumina (Bentley DR, 2008) platform workflows.

Once the library preparation step is concluded, sequencing can be performed.

In NGS this process is a cyclic alternance of nucleotide incorporation and base detection steps. Given the shared basic principle, several differences occur depending on the sequencing platform.

Illumina sequencers adopted the cyclic reversible termination strategy.

In these instruments, in each sequencing cycle the polymerase can incorporate only one of the four fluorescent labelled nucleotides, according to base complementarity.

After removing unincorporated nucleotides, a laser system detects the fluorescent dye and determines nucleotide identification. The labelled dye and terminating group are then removed and the next cycle starts with a newly modified nucleotide flow.

Conversely, Ion Torrent strategy is based on pH variation detection. When a nucleotide is incorporated by the polymerase in the growing sequence, H^+ ions are released and determine a change in pH.

The Ion chip is structured in microwells, each one allows the presence of a single bead (after emulsion PCR). A detector is present on the lower surface of the chip that permits the identification of single well local pH variations, which is translated into incorporated nucleotide identity.

In this way, unlike Illumina sequencers in which labelled nucleotides allow contemporary flow, the base incorporation step is subsetted into four separate nucleotide flows, each one followed by the removal of unincorporated nucleotides and detections in pH change at single chip well resolution.

The association of the identity of flowed nucleotides with pH variation revelation in each cycle permits the construction of sequences for all library fragments and their clonally amplified copies associated within a single bead located in the chip.

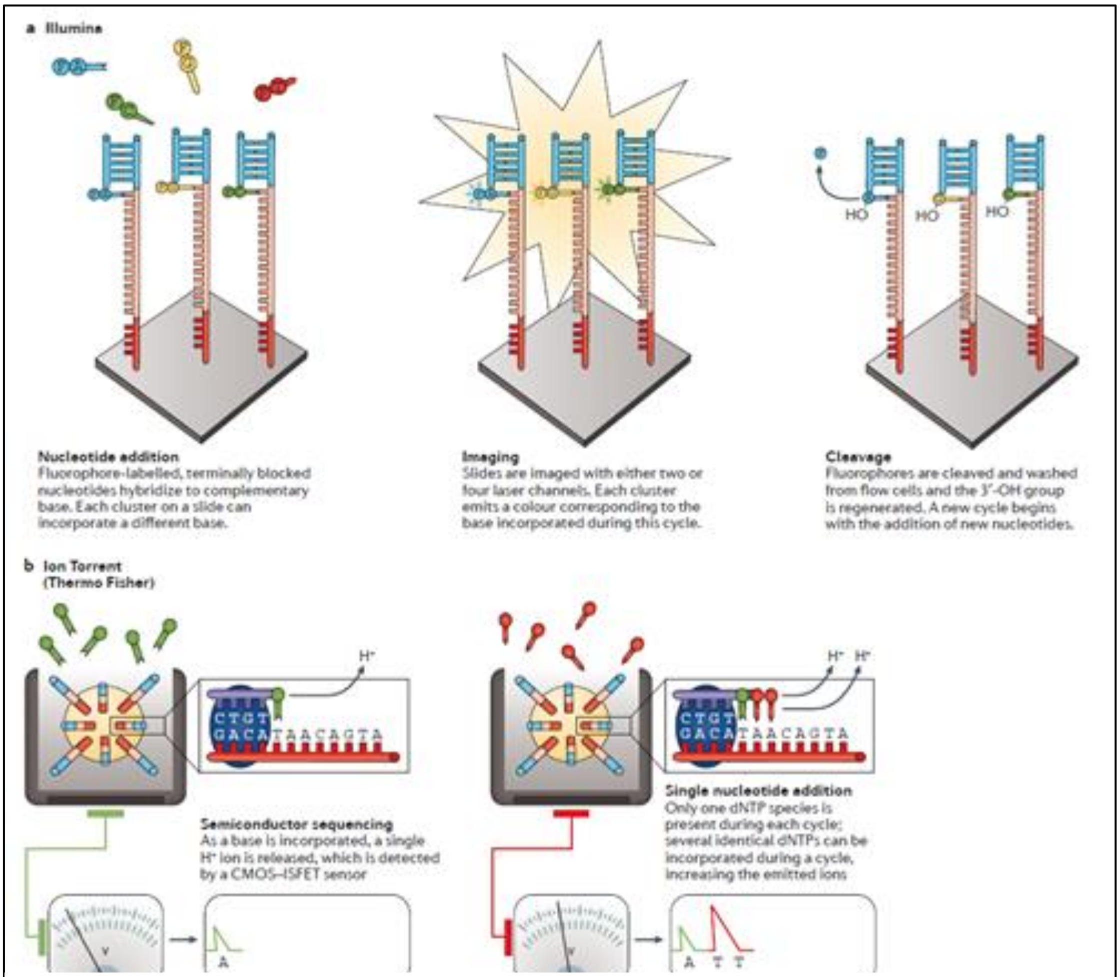

a Illumina
HO
HO
HO
Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.
Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.
b Ion Torrent (Thermo Fisher)
H+
CTGT
GACA
TAACAGTA
Semiconductor sequencing
As a base is incorporated, a single H+ ion is released, which is detected by a CMOS–ISFET sensor
V
A
H+ H+
CTGT
GACA
TAACAGTA
Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions
V
A T T

Figure 2. **Sequencing by synthesis. a) Illumina**. After solid-phase template enrichment, a mixture of primers, DNA polymerase and modified nucleotides are added to the flow cell. Each nucleotide is blocked by a 3′-O-azidomethyl group and is labelled with a base-specific, cleavable fluorophore (F). During each cycle, fragments in each cluster will incorporate just one nucleotide as the blocked 3′ group prevents additional incorporations. After base incorporation, unincorporated bases are washed away and the slide is imaged by total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels; the colour (or the lack or mixing of colours in the two-channel system used by NextSeq) identifies which base was incorporated in each cluster. The dye is then cleaved and the 3′-OH is regenerated with the reducing agent tris(2-carboxyethyl)phosphine (TCEP). The cycle of nucleotide addition, elongation and cleavage can then begin again. **b) Ion Torrent**. After bead-based template enrichment, beads are carefully arrayed into a microtitre plate where one bead occupies a single reaction well. Nucleotide species are added to the wells one at a time and a standard elongation reaction is performed. As each base is incorporated, a single H+ ion is generated as a by-product. The H+ release results in a 0.02 unit change in pH, detected by an integrated complementary metal-oxide semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) device. After the introduction of a single nucleotide species, the unincorporated bases are washed away and the next is added. *(Goodwin S, 2016)*

After sequencing, the output produced by NGS is composed of short sequences that are commonly called “reads”. A genomic region will be represented by such N reads as the N number of its copies in the sequencing library. More details on NGS data are treated in the next section.

The above described methods show the advantage provided by NGS in sequencing data production, which is performed in a stepwise manner due to coupled sequencing-detection processes. However, Sanger sequencing data production is performed in two separated steps, which are the production of fluorescently labelled fragments and their consecutive electrophoresis separation, associated with their fluorescence-based detection.

Another key difference between NGS and Sanger sequencing is represented by read length.

Sanger sequencing allows the generation of reads ranging between 600 and 800 bp and up to 1000 bp. In contrast, NGS platforms produce sequencing reads of about 100-400 bp. These short reads need to be aligned to a reference genome before performing downstream data analysis and interpretation, determining the impossibility of genome assembly by using shared sequences overlaps as done for Sanger sequencing reads (Mardis, 2017).

An extraordinary improvement in DNA sequencing determined by NGS technology is represented by James Watson's genome sequencing (Wheeler DA, 2008).

In 2008 Watson's genome was sequenced in four months and with an associated cost of less than $ 1 million on a 454 NGS platform. Compared to the Human Genome Project and the first personal sequenced genome published by Venter (Levy S, 2007), NGS-based strategy was seen to be both cost effective and time-competitive, with a 1000 and 100-fold drop in sequencing cost respectively.

NGS technologies has continued to evolve, reaching high throughput with the possibility of sequencing a human genome in 1-3 days and bringing the cost of sequencing down to around $ 1000 (Veritas Genetics, 2016) (Wetterstrand, 2016).

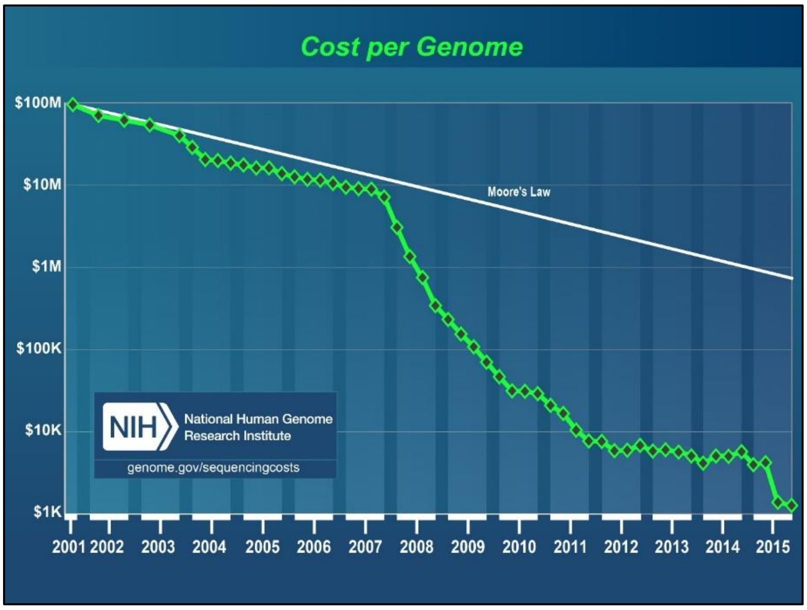


Figure 3. **Graph of "Cost per Genome".** This graph illustrates the nature of the reductions in sequencing costs and also hypothetical data reflecting Moore's Law. Adapted from *(Wetterstrand, 2016)*

These important features have allowed the feasibility of large projects such as the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015) (Sudmant PH, 2015), the UK10K (Wetterstrand, 2016) project (The UK10K Consortium, 2015) and the Haplotype Reference Consortium (The Haplotype Reference Consortium, 2016). Aside from each study aim and their associated results, the importance of large projects lies in their data availability as public open-source data sets, providing great statistical power even to small research groups.

Recently, NGS started to be adopted in several clinical applications, from target panels of genes to whole genome sequencing, in different areas of interest like oncology, reproductive health and other complex diseases.

Although successful, the NGS approach has not replaced Sanger sequencing, which is still involved in large numbers of diagnostic tests due to its high sequence accuracy and its low throughput, compatible with most diagnostic centers afference.

The above mentioned progress in DNA sequencing provided by NGS is coupled with a parallel renovation of the computational biology field, with an increased importance of data analysts and bioinformaticians.

1.2 BIOINFORMATIC NGS DATA ANALYSIS

The high throughput yielded by NGS has determined a new challenge: the big data management.

While sequencing output can vary depending on the NGS deep level application and the extension of the region to be sequenced, data analysis requires sophisticated computational techniques and the usage of high performance computers or infrastructures such as clouds or CPU clusters (Schmidt B, 2017).

In a context where data analysis represents the main part of projects including NGS, bioinformaticians started to have a prominent role in research groups, because of their capability of integrating biology with computer science and statistics.

A large number of tools have been developed recently for a high number of applications, covering the whole data analysis process from raw data produced by the instruments to the target level of information desired.

This can be assessed by the increased number of scientific publications in concomitance with NGS technology development.

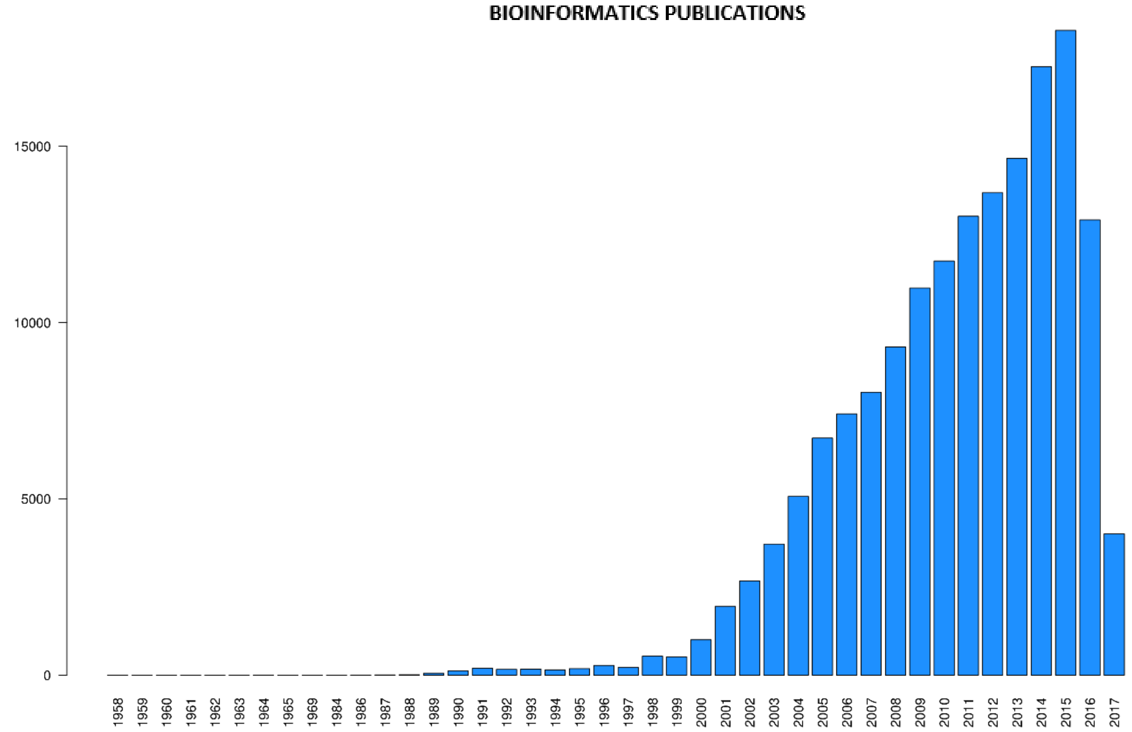
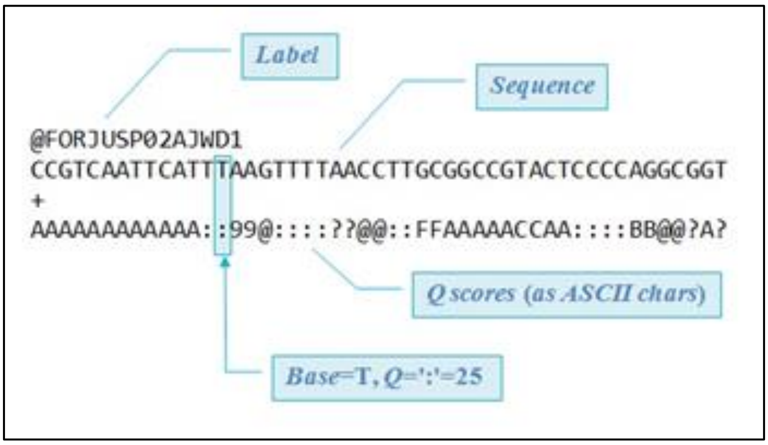


Figure 4. **Bioinformatics Publications.** Number of publications per year for the topics “bioinformatics” and “computational biology” from NCBI PubMed Search.

A Next-Generation Sequencing data analysis basic workflow can be summarized into four main steps: base calling, alignment, variant calling and downstream analysis (Nielsen R, 2011).

- **BASE CALLING**

Raw signals generated by the sequencing instruments are converted into nucleotide bases with associated quality scores. Base calling is performed by the software integrated within the sequencing instrument, which is specific for the sequencing platform used according to instrument sequencing chemistry, and all sequenced fragments, called reads, are stored in FASTQ format (Cock PJA, 2010).



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Figure 5. **FASTQ file format**. There are four line types in the FASTQ format. First a '@' title line which often holds just a record identifier. Second comes the sequence line(s), which as in the FASTA format can be line wrapped. Third, to signal the end of the sequence lines and the start of the quality string, comes the '+' line. Finally, comes quality line(s) which again can be wrapped. These use a subset of the ASCII printable characters (at most ASCII 33–126 inclusive) with a simple offset mapping. Crucially, after concatenation (removing line breaks), the quality string must be equal in length to the sequence string. *(Cock PJA, 2010) (Drive5, s.d.)*

FASTQ format is composed of four line types. The first line, which starts with "@", is the title line with run and sequence read identification IDs. The second line is the single letter code string of the read sequence. Then, a "+" line signals the end of the sequence lines and the start of the quality string. Finally, comes quality line coded with a subset of the ASCII printable characters (at most ASCII 64–126 inclusive), accounting for per-base PHRED quality score, defined in terms of estimated probability of error and computed according to the formula Q PHRED = -10 * log 10 (P error). A PHRED score of 20 corresponds to a 1% error rate in base calling.

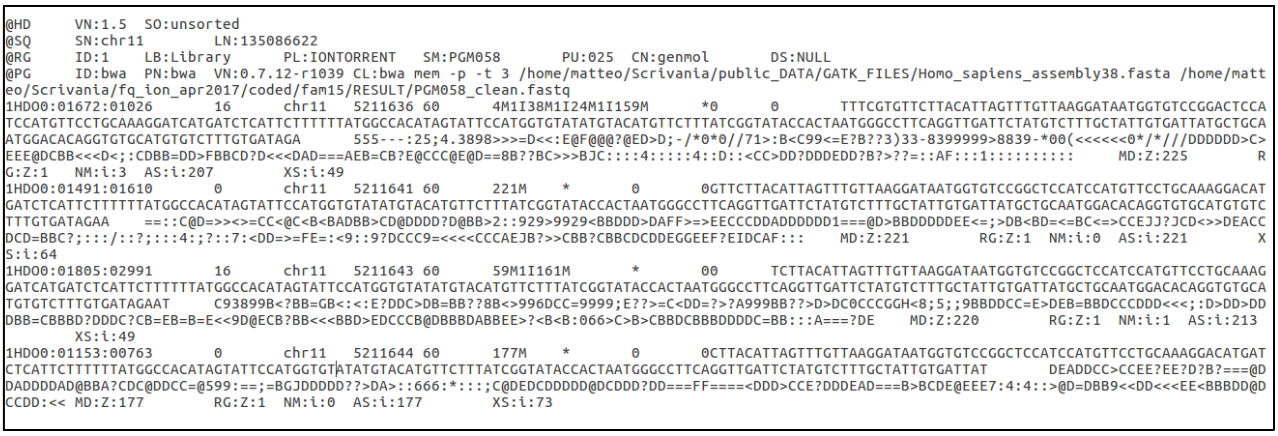
- **ALIGNMENT**

After base calling, reads coming from a single sample are stored in one or more FASTQ files, depending on sequence extension length. While NGS produce reads with a length range between 100-400 bp, it is possible to obtain a continuous DNA sequence (Mardis, 2017). This is achieved with alignment step, also called mapping, in which a mapping algorithm tries to locate each read present in the FASTQ file on a reference DNA sequence. A large amount of software has been developed in this way with different strategies and algorithms. The application of the Burrows-Wheeler Transform (BWT) has particular relevance to string matching, which led to the development of the widely used software BWA, SOAP2 and Bowtie. These tools, in respect to previously used MAQ, reach similar high mapping accuracy in a 10-20 fold time reduction. In fact, while MAQ works by hashing reads and scanning through the reference genome to find matches, BWT-based aligners compress data and index transformed strings for faster read query.

Aligners generally output mapped reads in SAM format (Sequence Alignment/Map), a file format ideated by Li et al. to standardize alignment output from different tools and create an interface between alignment and downstream analysis (Li H, 2009).

The SAM format consists of one header section and one alignment section.

The header section starts with the "@" character and contains information about file format, sorting order, reference sequence used for alignment, sample name, read group, the software and parameters used for alignment and other optional fields. The alignment section is composed of 11 mandatory TAB-delimited fields which carry information of read name, reference sequence name, read mapping position on the reference, PHRED-scaled mapping quality score, a CIGAR string showing the location of points of variations from the reference sequence, the read sequence and its per-base quality score.



```
@HD     VN:1.5  SO:unsorted
@SQ     SN:chr11        LN:135086622
@RG     ID:1    LB:Library      PL:IONTORRENT   SM:PGM058       PU:025  CN:genmol       DS:NULL
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -p -t 3 /home/matteo/Scrivania/public_DATA/GATK_FILES/Homo_sapiens_assembly38.fasta /home/matt
eo/Scrivania/fq_ion_apr2017/coded/fam15/RESULT/PGM058_clean.fastq
1HDO0:01672:01026       16      chr11   5211636 60      4M1I38M1I24M1I159M      *0      0       TTTCGTGTTCTTACATTAGTTTGTTAAGGATAATGGTGTCCGGACTCCA
TCCATGTTCCTGCAAAGGATCATGATCTCATTCTTTTTTATGGCCACATAGTATTCCATGGTGTATATGTACATGTTCTTTATCGGTATACCACTAATGGGCCTTCAGGTTGATTCTATGTCTTTGCTATTGTGATTATGCTGCA
ATGGACACAGGTGTGCATGTGTCTTTGTGATAGA      555---:25;4.3898>>>=D<<:E@F@@@?@ED>D;-/*0*0//71>:B<C99<=E?B??3)33-8399999>8839-*00(<<<<<<0*/*///DDDDDD>C>
EEE@DCBB<<<D<;:CDBB=DD>FBBCD?D<<<DAD===AEB=CB?E@CCC@E@D==8B??BC>>>BJC::::4:::::4::D::<CC>DD?DDDEDD?B?>??=::AF:::1::::::::::     MD:Z:225        R
G:Z:1   NM:i:3  AS:i:207        XS:i:49
1HDO0:01491:01610       0       chr11   5211641 60      221M    *       0       0GTTCTTACATTAGTTTGTTAAGGATAATGGTGTCCGGCTCCATCCATGTTCCTGCAAAGGACAT
GATCTCATTCTTTTTTATGGCCACATAGTATTCCATGGTGTATATGTACATGTTCTTTATCGGTATACCACTAATGGGCCTTCAGGTTGATTCTATGTCTTTGCTATTGTGATTATGCTGCAATGGACACAGGTGTGCATGTGTC
TTTGTGATAGAA    ==::C@D=>><>=CC<@C<B<BADBB>CD@DDDD?D@BB>2::929>9929<BBDDD>DAFF>=>EECCCDDADDDDDD1===@D>BBDDDDDEE<=;>DB<BD=<=BC<=>CCEJJ?JCD<>>DEACC
DCD=BBC?;:::/::?;:::4:;?::7:<DD=>=FE=:<9::9?DCCC9=<<<<CCCAEJB?>>CBB?CBBCDCDDEGGEEF?EIDCAF:::    MD:Z:221        RG:Z:1  NM:i:0  AS:i:221        X
S:i:64
1HDO0:01805:02991       16      chr11   5211643 60      59M1I161M       *       00      TCTTACATTAGTTTGTTAAGGATAATGGTGTCCGGCTCCATCCATGTTCCTGCAAAG
GATCATGATCTCATTCTTTTTTATGGCCACATAGTATTCCATGGTGTATATGTACATGTTCTTTATCGGTATACCACTAATGGGCCTTCAGGTTGATTCTATGTCTTTGCTATTGTGATTATGCTGCAATGGACACAGGTGTGCA
TGTGTCTTTGTGATAGAAT     C93899B<?BB=GB<:<:E?DDC>DB=BB??8B<>996DCC=9999;E??>=C<DD=?>?A999BB??>D>DC0CCCGGH<8;5;;9BBDDCC=E>DEB=BBDCCCDDD<<<;:D>DD>DD
DBB=CBBBD?DDDC?CB=EB=B=E<<9D@ECB?BB<<<BBD>EDCCCB@DBBBDABBEE>?<B<B:066>C>B>CBBDCBBBDDDDC=BB:::A===?DE    MD:Z:220        RG:Z:1  NM:i:1  AS:i:213
        XS:i:49
1HDO0:01153:00763       0       chr11   5211644 60      177M    *       0       0CTTACATTAGTTTGTTAAGGATAATGGTGTCCGGCTCCATCCATGTTCCTGCAAAGGACATGAT
CTCATTCTTTTTTATGGCCACATAGTATTCCATGGTGTATATGTACATGTTCTTTATCGGTATACCACTAATGGGCCTTCAGGTTGATTCTATGTCTTTGCTATTGTGATTAT       DEADDCC>CCEE?EE?D?B?===@D
DADDDDAD@BBA?CDC@DDCC=@599:==;=BGJDDDDD??>DA>::666:*:::;C@DEDCDDDDD@DCDDD?DD===FF====<DDD>CCE?DDDEAD===B>BCDE@EEE7:4:4::>@D=DBB9<<DD<<<EE<BBBDD@D
CCDD:<< MD:Z:177        RG:Z:1  NM:i:0  AS:i:177        XS:i:73
```

Figure 6. **SAM file format**. The '@SQ' line in the header section gives information of reference sequences. In the '@PG' lines are shown the program and commands used for alignment. Then the alignment section contains mapped reads one record per row.

All this information can be used to select high quality reads and obtain good quality data for downstream analysis. Read mapping represents a very important step for variant detection and needs to be accurate to avoid errors in SNP and genotype calling due to incorrectly aligned reads.

For this reason the alignment step is generally followed by some refinement steps such as quality filtering, PCR duplicates remove and re-alignment around putative insertion/deletion sites.

The text-based SAM file can furthermore be converted into its equivalent binary representation, the Binary Alignment/Map (BAM) format (Li H, 2009), which is compact in size and supports fast retrieval of alignments in specified regions.

BAM files can also be visualized with a tool developed by the Broad Institute, the Integrative Genomics Viewer (Robinson JT, 2011). IGV allows sequence reads visualization with a single-base resolution, showing the number of sequenced reads (DEPTH) and the allele they carry. It is integratable with SNP IDs and genomic annotations from public sources.

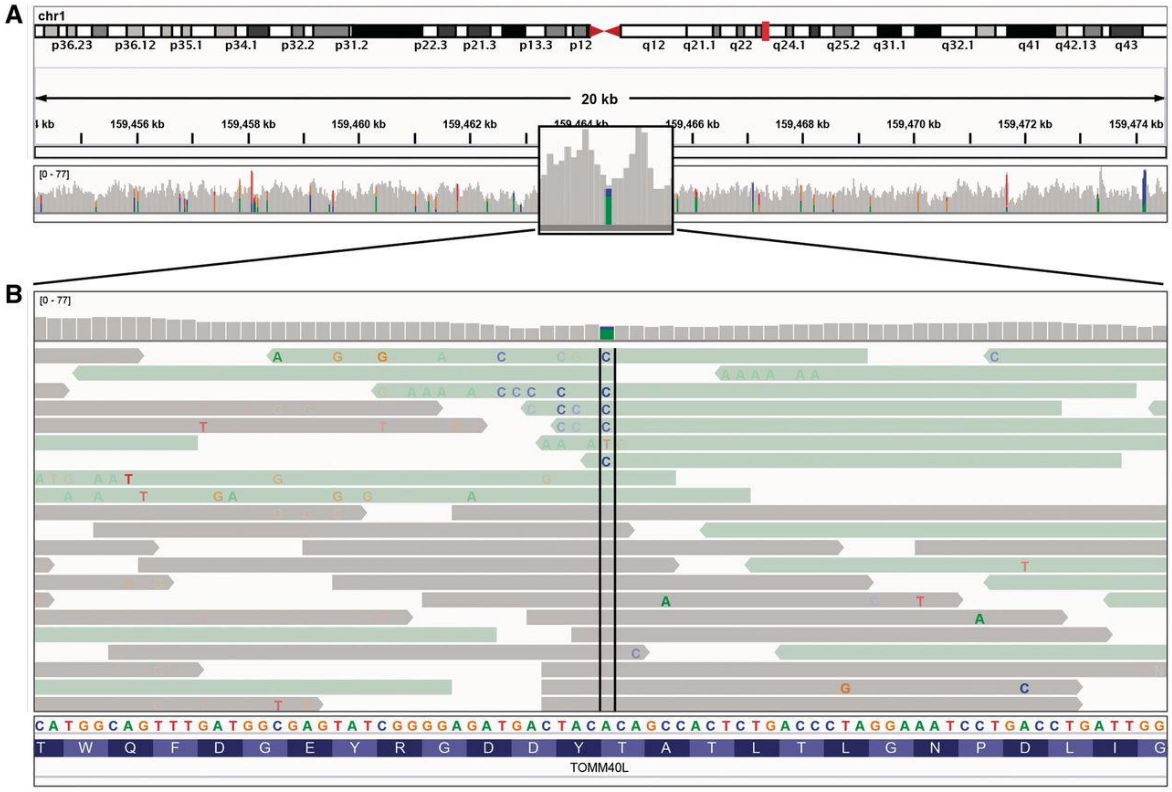

A
chr1
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q41 q42.13 q43
20 kb
159,456 kb 159,458 kb 159,460 kb 159,462 kb 159,464 kb 159,466 kb 159,468 kb 159,470 kb 159,472 kb 159,474 kb
[0 - 77]
B
[0 - 77]
CATGGCAGTTTGATGGCGAGTATCGGGGAGATGACTACACAGCCACTCTGACCCTAGGAAATCCTGACCTGATTGG
T W Q F D G E Y R G D D Y T A T L T L G N P D L I G
TOMM40L

Figure 7. **Read alignment views at 20 kb and base pair resolution**. IGV displays varying level of data detail depending on the zoom level, and uses color and transparency to highlight interesting events in the data. (**A**) Reads are summarized as a coverage plot. Positions with a significant number of mismatches with respect to the reference are highlighted with color bars indicative of both the presence of mismatches and the allele frequency. (**B**) Individual base mismatches are displayed with alpha transparency proportional to quality. In this example, the reads have been sorted and colored by strand. *(Thorvaldsdóttir H, 2012)*

- **VARIANT CALLING**

Sequencing product stored in BAM format has to be converted in an analyzable file which undergoes downstream analysis.

This process is called "Variant Calling" and consists of the comparison of sequenced reads from their point of alignment on the reference genome sequence to identify positions in which bases are different.

Variant identification’s major challenge is the selection of true variants from sequencing errors among all reads.

Early variant calling methods were based on simple allele counting after a filtering step to keep only high-confidence bases. Generally, sequenced bases with a PHRED quality score were below 20 (accounting for a sequencing error rate greater than 1%) were filtered out (Nielsen R, 2011).

Counting alleles present in the remaining reads allowed for the identification of sequence variants and the determination of individual genotypes for that genomic position based on the alleles ratio.

More recent variant caller tools integrated this strategy with several probabilistic methods which combine base quality scores and public database information, for example dbSNP allele frequencies, to compute a genotype probability. This additional information can then be used to keep only variation points with high confidence and proceed with downstream analysis with good quality data.

Information decoded by variant calling is stored in a specific file format, the Variant Call Format (VCF) proposed by Danecek et al as a standard format for sequence variation data (Danecek P, 2011).

A VCF file consists of three parts: a meta-information section, a header line and a data section.

Meta-information lines start with “##” characters and provide information on file creation like date, used software and parameters and, most importantly, a description of all data section fields.

The VCF header line starts with the “#” character and names data columns. It is composed of eight mandatory fields which, if genotype data is present in the file, are followed by a FORMAT and sample ID column headers, according to the number of samples included in the VCF.

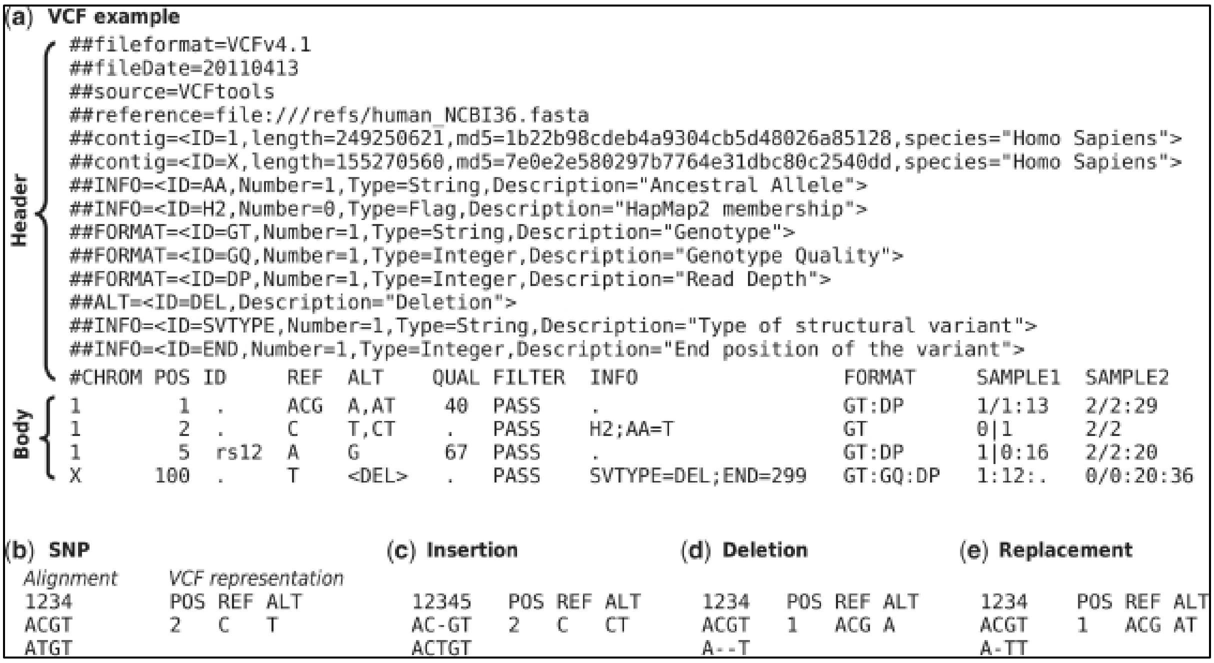
The header line and data section are tab-delimited.

The mandatory fields of a VCF are:

- CHROM: chromosome in which the variant has been mapped;
- POS: 1-based position of the variant on the chromosome;
- ID: variant ID, for example rsIDs from dbSNP;
- REF: reference allele;
- ALT: comma separated list of alternate non-reference alleles;
- QUAL: phred-scaled quality score;
- FILTER: site filtering information;
- INFO: additional, user extensible information.

The 9th column is generally the FORMAT field, which specifies what kind of information is present in the subsequent sample column, the last labelled with individual ID. An example could be a FORMAT field “GT:GQ:DP” which indicates that in the sample column in order the called genotype (GT), the computed genotype quality (GQ) and the read depth (DP) separated by the character “:” can be found. During variant calling it is possible to define what kind of information will be kept for analyzed samples and so in FORMAT field.

According to the fields described in the header line, data is TAB-separated and organized into one record per-row.


(a) VCF example
Header
Body
(b) SNP
(c) Insertion
(d) Deletion
(e) Replacement

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID      REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1  .      ACG  A,AT    40  PASS    .                    GT:DP      1/1:13   2/2:29
1        2  .      C    T,CT    .   PASS    H2;AA=T              GT         0|1      2/2
1        5  rs12   A    G       67  PASS    .                    GT:DP      1|0:16   2/2:20
X      100  .      T    <DEL>   .   PASS    SVTYPE=DEL;END=299   GT:GQ:DP   1:12:.   0/0:20:36
```

```
Alignment   VCF representation
1234        POS REF ALT          12345   POS REF ALT     1234   POS REF ALT     1234   POS REF ALT
ACGT        2   C   T            AC-GT   2   C   CT      ACGT   1   ACG A       ACGT   1   ACG AT
ATGT                             ACTGT                   A--T                   A-TT
```

Figure 8. (**a**) **Example of valid VCF**. The header lines ##fileformat and #CHROM are mandatory, the rest is optional but strongly recommended. Each line of the body describes variants present in the sampled population at one genomic position or region. All alternate alleles are listed in the ALT column and referenced from the genotype fields as 1-based indexes to this list; the reference haplotype is designated as 0. For multiploid data, the separator indicates whet her the data are phased (|) or unphased (/). Thus, the two alleles C and G at the positions 2 and 5 in this figure occur on the same chromosome in SAMPLE1. The first data line shows an example of a deletion (present in SAMPLE1) and a replacement of two bases by another base (SAMPLE2); the second line shows SNP and an insertion; the third a SNP; the fourth a large structural variant described by the annotation in the INFO column, the coordinate is that of the base before the variant. (**b–e**) **Alignments and VCF representations of different sequence variants**: SNP, insertion, deletion, and replacement. The REF columns shows the reference bases replaced by the haplotype in the ALT column. The coordinate refers to the first reference base. *(Danecek P, 2011)*

All missing values are specified with the "." character. This is important to keep the file structure integrity, allowing the construction of indexes which are responsible for VCFs fast data access.

- **DOWNSTREAM ANALYSIS**:

The raw VCF produced after variant calling represents the downstream analysis starting point.

Primarily, detected variants have to be annotated to determine their biological significance.

A large number of public sources are available and provide various classes of information.

Among them, frequency-based annotations allow the restriction of the candidate polymorphisms with biological significance based on single variant frequency in a given population.

Structural-based annotations permit the variant prioritization on the resulting changes in the protein structure determined by amino acid substitution.

The main objective of downstream analysis is to focus on the relevant information surrounded by the large number of sequences produced by NGS and support scientists in the identification of the biological cause they are investigating.

Naturally, the wide range of biological questions has led to the development of a great amount of bioinformatic software, specific for each field of application.

In a project aimed at developing a Non-Invasive Prenatal Diagnosis protocol for β-thalassemia, target sequencing of cell-free fetal DNA present in maternal plasma was performed with the benchtop NGS instrument Ion PGM (Personal Genome Machine - Thermo Fisher). The subsequent data analysis was conducted through an in-house developed bioinformatic pipeline.

The next section provides a biological background on circulating fetal derived DNA and its application to Non Invasive Prenatal Diagnosis.

1.3 CIRCULATING NUCLEIC ACIDS: cell-free fetal DNA

After the demonstration in 1987 that senescent tumoral cells are able to release DNA molecules in peripheral blood (Stroun M, 1987), D.Lo hypothesized that also fetal cells should release DNA in maternal circulation.

This hypothesis was confirmed in 1997 by Real Time PCR identification of Y chromosome sequences (for example DYS14 sequence) in plasma of women carrying male fetuses (Lo YMD C. N., 1997). This discovery led several research groups to focus their studies on cell-free fetal DNA (cffDNA).

The same group responsible for cffDNA discovery reported the quantification through Real Time PCR of fetal DNA fraction in maternal plasma, observing that cffDNA concentration ranges from 3% in the first trimester to 6% at the end of pregnancy (Lo YMD T. M., 1998).

The availability of new and more sensitive technologies, like Digital PCR, allowed the demonstration that cffDNA concentration in maternal plasma is on average around 10%, growing proportionally with gestational age (Lun FM, 2008), reaching 26% in the third trimester of pregnancy.

Furthermore, it has been observed that an increase in cffDNA concentration is related to several pathogenic factors such as preeclampsia (Lo YMD L. T., 1999) and trisomy 21.

The origin of cffDNA was demonstrated by Flori and Alberry to be the syncytiotrophoblast (Flori E, 2004) (Alberry M, 2007). These cells, after apoptosis, release fragments of DNA which are able to cross the placenta and enter maternal blood circulation (Ariga H, 2001).

Apoptotic origin is also supported by cffDNA fragments length, ranging from 50 bp to 143 bp with a periodicity of 10 bp (Lo YMD C. K., 2010).

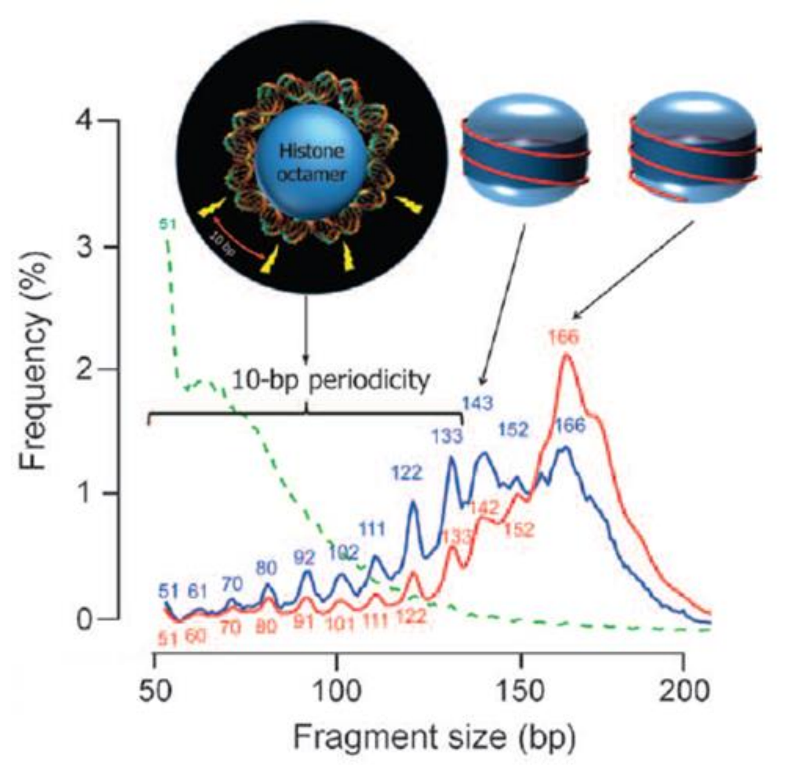
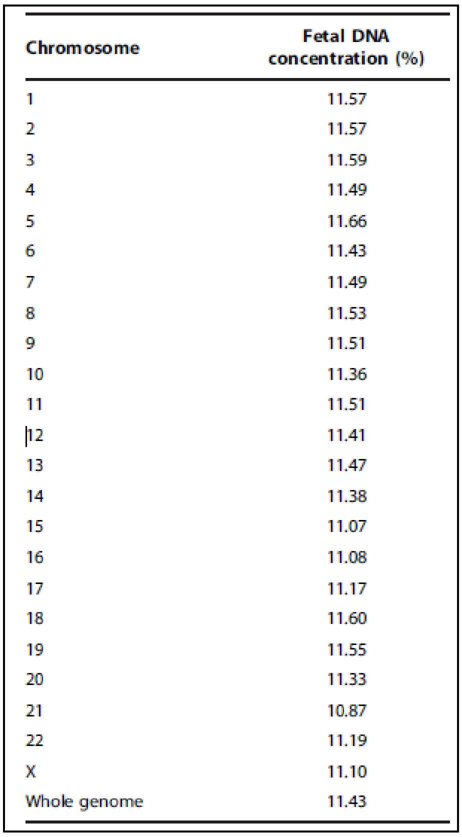


Figure 9. **Sequencing of fetal and total DNA in maternal plasma**.Size distribution of fetal DNA (blue curve), total DNA (red curve), and mitochondrial DNA (green broken curve). Numbers denote the DNA size at the peaks. Schematic illustrations of the structural organization of a nucleosome are shown above the graph. From left to right, DNA double helix wound around a nucleosomal core unit with the sites for nuclease cleavage shown; a nucleosome core unit with ~146 bp of DNA (red tape) wound around it;and a nucleosomal core unit with an intact ~20-bp linker sequence. *(Lo YMD C. K., 2010)*

Trophoblasts are not the only cells that release DNA in maternal plasma, which is composed of 90% of free maternally-derived DNA fragments of around 166 bp and only 10% of fetal DNA.

Although cffDNA has a low percentage concentration in maternal plasma and has a fragmented status, the advent of NGS has shown that it contains the whole fetal genome, with an equal representation of all fetal chromosomes (Lo YMD C. K., 2010).



Chromosome	Fetal DNA concentration (%)
1	11.57
2	11.57
3	11.59
4	11.49
5	11.66
6	11.43
7	11.49
8	11.53
9	11.51
10	11.36
11	11.51
12	11.41
13	11.47
14	11.38
15	11.07
16	11.08
17	11.17
18	11.60
19	11.55
20	11.33
21	10.87
22	11.19
X	11.10
Whole genome	11.43

Figure 10. Fractional concentrations of fetal DNA for different chromosomes. (Lo YMD C. K., 2010)

Another important cffDNA feature is its short half-life of around 16 minutes (Benachi A, 2003).

In fact, it is rapidly cleared from maternal circulation a few hours after delivery through kidney excretion, thus excluding genetic contamination between consecutive pregnancies (Lo YMD Z. J., 1999).

Described features shows that cffDNA represents an optimal fetal genetic source for non invasive prenatal diagnosis. Nevertheless, due to its low concentration in respect to maternal free DNA (1:10), cffDNA analysis requires highly sensitive techniques which allow circulating fetal DNA enrichment.

1.4 NON INVASIVE PRENATAL DIAGNOSIS

One of the first applications of cffDNA analysis to Non Invasive Prenatal Diagnosis (NIPD) was fetal gender determination.

This simple method is based on the identification of Y chromosome mapping sequences that the fetus inherits from the father and which are absent in the maternal genome.

Non invasive fetal gender determination in early gestational age has an important diagnostic significance for prenatal diagnosis of X-linked diseases (Hemophilia, DMD) and Congenital Adrenal Hyperplasia (CAH).

In the field of X-linked diseases, fetal gender determination could avoid invasive diagnostic procedures for all women carrying female fetuses (Costa JM, 2002).

CAH at risk pregnancies are routinely subjected to dexamethasone prenatal treatment from the first weeks of gestation, to reduce the virilization risk of female fetuses.

Non Invasive fetal gender identification should allow the interruption of this pharmacological treatment in male fetuses several weeks before echographic evidence (Rijnders RJ, 2001).

Mackie et al analyzed 60 studies published between 1997 and 2015 (11179 tests) to estimate diagnostic performance of non invasive fetal gender determination through cffDNA analysis. Quantitative Real Time PCR of Y chromosome mapping sequences has allowed fetal gender identification starting from the 5th week of gestation with high sensitivity and specificity, respectively 98.9% and 99.9%, in a large cohort of samples (Mackie FL, 2017).

The successful detection power of this method has been translated into clinical practice for several years in some European countries and the United States, with a consequent reduction in the number of performed invasive tests (Finning KM, 2008).

Another important application of cffDNA analysis is fetal Rhesus D genotype determination (Lo YMD H. N., 1998).

RhD gene maps on the short arm of chromosome 1 and codes for an antigenic protein on erythrocyte cell membrane. Due to RhD factor dominant autosomic inheritance, its status determination assumes great importance in caucasian-derived populations, in which individuals with negative Rhesus phenotype show deletion of this gene.

During pregnancy evolution, fetal-maternal RhD factor incompatibility determine the Hemolytic Disease of the Newborn (HDN), also known as Erythroblastosis Fetalis.

In cases of Rh-positive fetuses, Rh-negative mothers produce G class immunoglobulins (IgG) against fetal erythrocyte D antigens.

Ig production generally starts at childbirth and is reactivated during successive pregnancies if the fetus presents RhD factor. Maternal Igs can cross the placenta and binds to fetal erythroblasts causing hemolysis.

To avoid this situation, during the first trimester of pregnancy, Rh-negative women are routinely subjected to anti-D immunoglobulin prophylaxis (Dovc-Drnovšek T, 2013).

Non Invasive Prenatal genotypization of fetal RhD factor permits early fetal genotype determination and the interruption of maternal immunoprophylaxis if the fetus is Rh-negative.

This simple test has been translated into clinical practice since 2001 (Finning K, 2004) and is routinely performed in several centers in Europe and the United States, with a sensitivity of 99.7% and specificity of 98.4%.

cffDNA analysis also represents a great target for NIPD of monogenic diseases such as β-thalassemia, cystic fibrosis and achondroplasia.

Autosomal dominant single gene disorders were the first class of monogenic diseases studied with this approach.

NIPD of these pathologies is performed through the identification of paternally-derived mutations inherited by the fetus which are not present in the maternal genome.

With this approach, responsible mutations of achondroplasia (G→A 1138 of FGFR3 gene) (Saito H, 2000), Huntington's Disease (Bustamante-Aragones A d. A.-T.-R.-L., 2012) and Myotonic Dystrophy (Gahan, 2013) have been identified non invasively in several couples at risk.

Also for autosomal recessive diseases diagnosis, the most simple approach involves the paternally inherited mutation which is not shared with the maternal genome.

In this way it is possible to assess that the fetus has a 50% probability of being affected if paternal mutation is detected, otherwise it is possible to exclude the inherited disease.

Unfortunately, this approach does not permit the determination of maternal mutation inheritance and, for this reason, is not a suitable procedure for families with parent carriers of the same mutation.

Due to circulating DNA composition of fetal-maternal DNA mixture in a 1:10 ratio, the discrimination of fetal fraction from maternal background requires highly sensitive technologies.

A suitable technology that can be used for this purpose is Digital PCR.

Firstly described in 1999 by Vogelstein and Kinzler (Vogelstein B, 1999), Digital PCR requires template dilution, to statistically determine the presence of a single DNA strand in each plate well, allowing in this manner the spatially separated amplification of a single allele. Allele-specific fluorescent probes then permit, after image capture and signal conversion, the revelation of the identity of sequences located in each well and obtain an accurate estimation of allele counts.

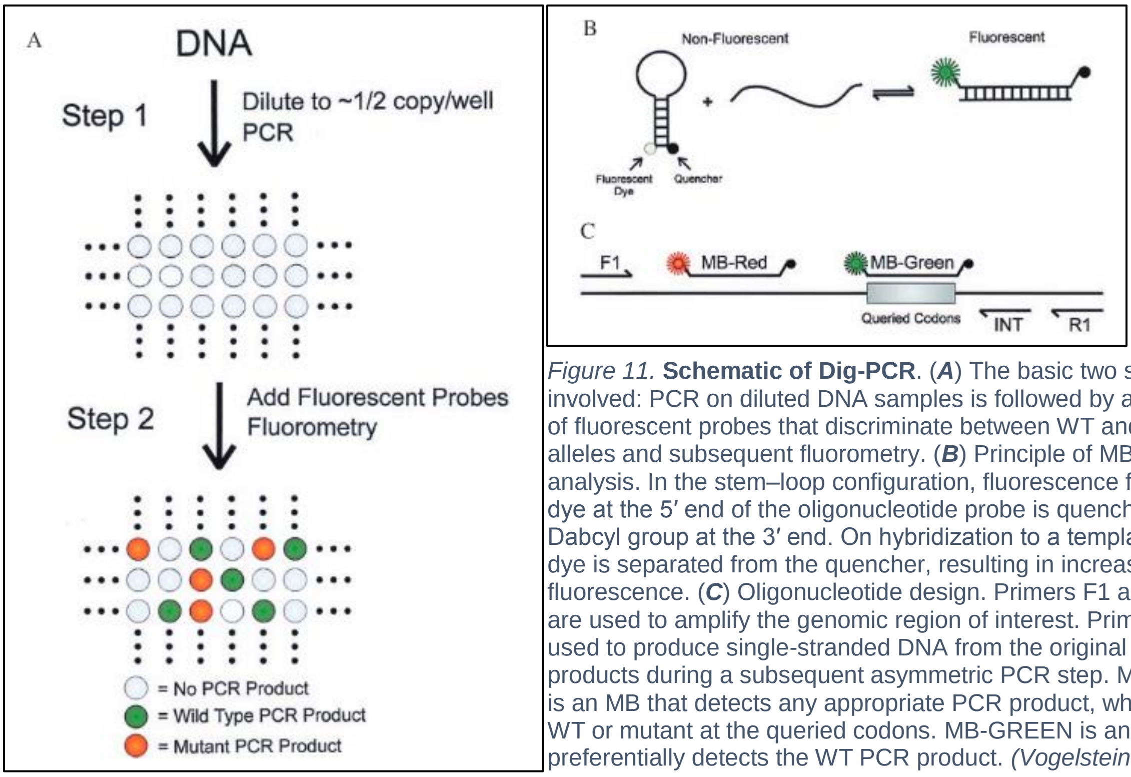


Figure 11. **Schematic of Dig-PCR**. (***A***) The basic two steps involved: PCR on diluted DNA samples is followed by addition of fluorescent probes that discriminate between WT and mutant alleles and subsequent fluorometry. (***B***) Principle of MB analysis. In the stem–loop configuration, fluorescence from a dye at the 5′ end of the oligonucleotide probe is quenched by a Dabcyl group at the 3′ end. On hybridization to a template, the dye is separated from the quencher, resulting in increased fluorescence. (***C***) Oligonucleotide design. Primers F1 and R1 are used to amplify the genomic region of interest. Primer INT is used to produce single-stranded DNA from the original PCR products during a subsequent asymmetric PCR step. MB-RED is an MB that detects any appropriate PCR product, whether it is WT or mutant at the queried codons. MB-GREEN is an MB that preferentially detects the WT PCR product. *(Vogelstein B, 1999)*

The application of Digital PCR to plasmatic DNA allows the determination of allele ratios, which can be analyzed with the Relative Mutation Dosage approach (RMD). This approach permits the definition of fetal genotypes by quantifying observed allelic imbalances in cell-free DNA data coming from Digital PCR, as a consequence of fetal-maternal DNA mixture presence.

In a situation in which parents are both carriers for the same mutation, it is possible to hypothesize that:

- if the fetus is homozygous for the variant allele, an allelic imbalance is expected in plasmatic DNA to be M>N (overexpression of mutant allele despite wild-type);
- If the fetus presents wild-type homozygous genotype, wild type allele is expected to be overrepresented ,   M<N;
- if the fetus has heterozygous genotype no allelic imbalance is expected, M=N.

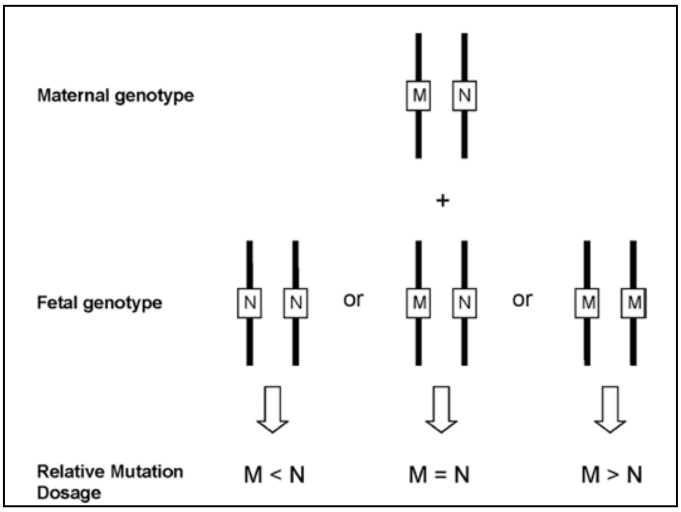


Figure 12. **Schematic illustration of the principle of digital RMD**. When a pregnant woman and her fetus are both heterozygous for a gene mutation, the amounts of the mutant allele (M) and wild-type allele (N) would be in allelic balance in maternal plasma. When the fetus is homozygous for the wild-type or mutant allele, there would be an underrepresentation or over-representation of the mutant allele, respectively. Digital RMD determines if the mutant and wild-type alleles in maternal plasma are in allelic balance or imbalance. *(Lun FMF, 2008)*

This approach has been proposed for NIPD of haemoglobinopathies (Lun FMF, 2008) and hemophilia (Lo YMD C. R., 2011).

In 2008, the Digital RMD approach was tested for non invasive fetal genotype determination in a cohort of 10 women with a gestational age ranging from 18 to 20 weeks (second trimester). All women involved were carriers of male fetuses, to ascertain cffDNA presence and estimate its fraction (Lun FMF, 2008). In 5 samples out of 10 the fetal genotype was correctly determined while in only 1 sample the genotype was incorrect. The remaining 4 samples resulted in indeterminate genotype due to low cffDNA fraction (<10%).

In 2011, Digital RMD allowed the correct fetal genotype determination in all analyzed cffDNA samples of a cohort of 7 pregnant women at risk for hemophilia (Tsui NBY, 2011).

More recently, Barrett et al applied Digital RMD to the NIPD of Sickle Cell Disease in a group of 65 women carriers of the HBB codon 6 mutation, reaching an 80% success rate which increased to 100% when considering samples with estimated fetal fraction greater than 7% (Barrett AN, 2012).

Unfortunately, Digital PCR is suitable only for sequences with a complementary dye-labelled probe, a feature that limits the cffDNA analysis to a restricted number of fragments.

The analysis of all cffDNA fragments present in maternal plasma became possible with the advent of NGS. The sensitivity of this class of instrument also permits the detection and the subsequent sequencing of less represented fragments.

The evidence that cffDNA fragments represent the whole fetal genome and that all chromosomes are equally represented with the same average concentration was reported in 2010 by Dennis Lo (Lo YMD C. K., 2010).

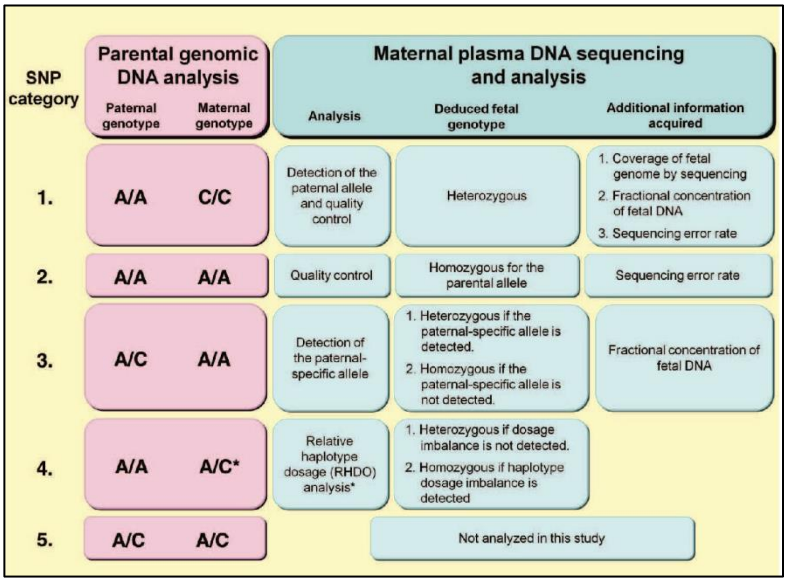
The study represents the first reported application of NGS in NIPD of β-thalassemia, in a family where parents were carriers for different mutations in the HBB gene.

The adopted analysis strategy was an extension of the RMD approach, the Relative Haplotype Dosage (RHDO).

Briefly, RHDO extends the allelic imbalance observation at the mutation site to a set of associated SNPs which constitute an haplotype block. This method requires an *a priori* knowledge of parental haplotypes.

SNPs are grouped in 5 different classes depending on the information they provide, as shown in Figure 13**.**

The determination of the paternally transmitted haplotype is possible by analyzing SNPs for which the father is heterozygous and the mother homozygous (SNP category 3).



SNP category	Parental genomic DNA analysis		Maternal plasma DNA sequencing and analysis		
	Paternal genotype	Maternal genotype	Analysis	Deduced fetal genotype	Additional information acquired
1.	A/A	C/C	Detection of the paternal allele and quality control	Heterozygous	1. Coverage of fetal genome by sequencing 2. Fractional concentration of fetal DNA 3. Sequencing error rate
2.	A/A	A/A	Quality control	Homozygous for the parental allele	Sequencing error rate
3.	A/C	A/A	Detection of the paternal-specific allele	1. Heterozygous if the paternal-specific allele is detected. 2. Homozygous if the paternal-specific allele is not detected.	Fractional concentration of fetal DNA
4.	A/A	A/C*	Relative haplotype dosage (RHDO) analysis*	1. Heterozygous if dosage imbalance is not detected. 2. Homozygous if haplotype dosage imbalance is detected	
5.	A/C	A/C	Not analyzed in this study		

Figure 13. **Noninvasive fetal genomic analysis from maternal plasma DNA**. Parental SNP combinations can be grouped into five categories. Categories 1, 2, and 3 allow the basic parameters for maternal plasma DNA sequencing to be established, including the percentage coverage of the fetal genome, fractional concentration of fetal DNA, and sequencing error rate. Category 3 also allows the fetal inheritance status of SNP alleles unique to the father to be studied. Mutations uniquely carried by the father can be regarded as category 3. Category 4 allows the inheritance status of the maternal haplotype to be studied. One application is the tracking of fetal inheritance of a haplotype block close to a mutation carried by the mother. Here, noninvasive fetal genomic analysis was carried out for a family undergoing prenatal diagnosis for b-thalassemia. Asterisk denotes that information on the maternal haplotype is required for the RHDO analysis. Category 5 SNPs were not analyzed in this study, but might be useful for the prenatal diagnosis of autosomal recessive disorders with consanguineous parents or genetic diseases with a strong founder effect. *(Lo YMD C. K., 2010)*

For maternally inherited haplotype determination the RHDO approach was applied. A cumulative count of category 4 SNPs (father homozygous and mother heterozygous) reads was performed for the two possible maternal haplotypes and evaluated through a Sequential Probability Ratio Test (SPRT), which determines the statistical significance of any allelic imbalance seen (Wald, 1947) (Royall, 1997).

When the classification threshold for SPRT is reached, the fetal inheritance of a maternal haplotype is established.

With the described approach it was possible to correctly determine the inheritance of the paternal mutation and the maternal wild type allele.

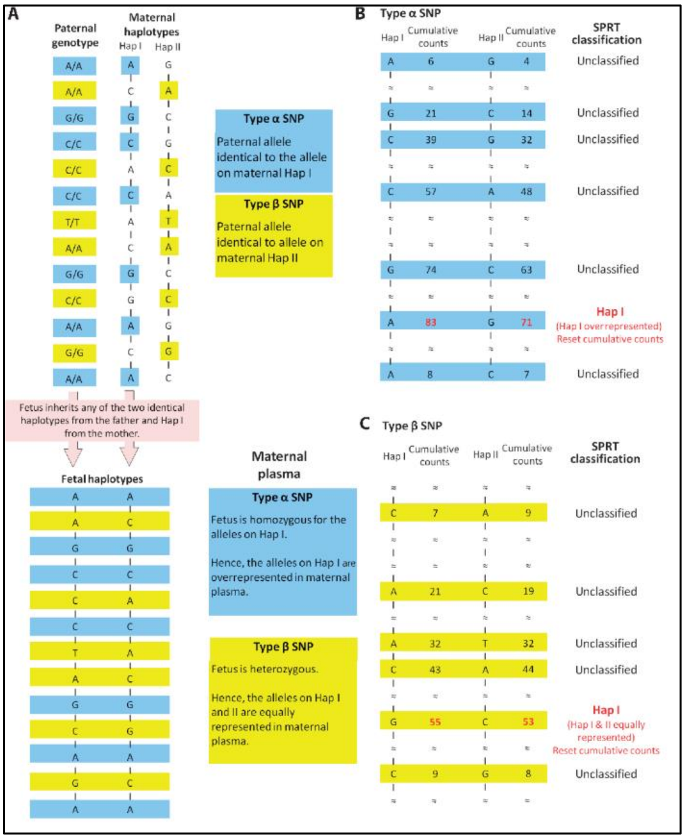


Figure 14. **Relative haplotype dosage (RHDO) analysis**. (**A**) In type α SNPs, paternal alleles are identical to the maternal alleles on Hap I. In type β SNPs, paternal alleles are identical to the maternal alleles on Hap II. If the fetus inherits Hap I from the mother, it is homozygous for type a and heterozygous for type β SNPs. (**B**) For type α SNPs, Hap I is overrepresented in maternal plasma. (**C**) For type β SNPs, there is no significant difference between the cumulative counts for Hap I and Hap II SNPs. Given that the fetus in this case inherits Hap II from the father, the sequential probability ratio test (SPRT) deduces the inheritance of Hap I from the mother.

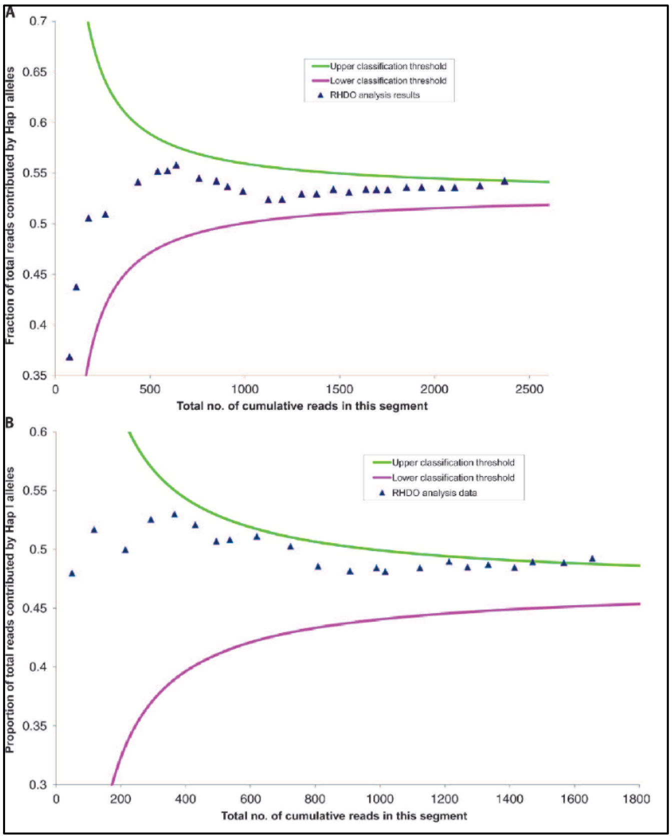


Figure 15. **SPRT classification**. (A and B) SPRT classification process for RHDO analysis of (**A**) type α and (**B**) type β SNPs in a region close to the pter of chromosome 1. The classification process runs in the direction from the telomeric end to the centromere.

The Haplotype-based strategy reported by Lo represents a valid method to determine fetal genotype with higher accuracy in respect to single variant site allelic imbalance observation. A limit of this procedure is represented by SNPs for which both parents are carriers (SNP category 5). These sites were excluded because of the lack of paternal haplotype information.

In 2012 Fan et al reported fetal genome non invasive determination by parental haplotype representation dosage in plasmatic DNA from 3 pregnant women reaching high accuracy values (Fan HC, 2012).

In the same year, Kitzman et al improved on haplotype counting methods with the development of a Hidden Markov Model (HMM) to infer the inherited maternal haplotype (Kitzman JO, 2012).

For each maternal-specific heterozygous site the HMM emits probabilities associated to the transmission of the haplotype with the allele shared with the father's genome and the haplotype carrying the different allele.

These probabilities are given by a binomial distribution that takes into account the fetal DNA fraction and the number of reads.

If the fetus is supposed to be homozygous for a given maternal heterozygous site, the probability of observing *k* such alleles among *N* total reads with fetal percentage *F* is:

$$\Pr(K = k|N, F) = Bin\left(N, \frac{1-F}{2} + \frac{F}{2} + \frac{F}{2}\right)$$

where $\frac{1-F}{2}$ represents maternal DNA, $\frac{F}{2}$ respectively paternal and maternal contributions.

If the inherited maternal allele and the paternal allele differ at a given site (fetus heterozygous), the probability of observing k inherited alleles simplifies to

$$\Pr(K = k|N, F) = Bin\left(N, \frac{1 - F}{2} + \frac{F}{2}\right) = Bin(N, 0.5)$$

The Viterbi algorithm (Viterbi, 1967) was then used to determine the most probable path among observed data, determining a prediction of the maternally transmitted haplotype.

For the paternal haplotype inheritance determination, paternal informative SNPs were used to assess the presence or absence of the paternal specific allele. The predicted haplotype was the one with a higher likelihood.

This method was successfully applied to NIPD of congenital adrenal hyperplasia (CAH) by Ma et al, who were able to infer maternal and paternal alleles with an accuracy of around 96% and 97% respectively (Ma D, 2014). In this study the transmission of paternal mutation and the inheritance of the maternal haplotype associated with the wild-type allele were correctly determined .

These encouraging results led to the development of a bioinformatic pipeline by adapting several described methods for cffDNA analysis to a particular familiar condition where parents are carriers of the same causative mutation.

1.5 NON INVASIVE PRENATAL TESTING

Although cffDNA analysis for monogenic diseases has shown promising prospectives, several improvements are needed for its translation into clinical practice. However, there has been considerable success in non invasive prenatal screening of common fetal aneuploidies.

After the demonstration that the fetal genome is entirely represented in maternal plasma and with a very similar chromosomal average fraction (Lo YMD C. K., 2010) , several research groups focused on the identification of differences in chromosome number (Lo, 2013).

The rapid development of specific sequencing protocols and dedicated bioinformatic algorithms (Boon EM, 2013), has led to the production and commercialization of Non Invasive Prenatal Tests from several private companies.

The most representative are the Materni T21 PLUS (Sequenom Laboratories), Verifi prenatal (Verinata Health, Illumina), NIFTY Test (BGI), IONA Test (Premaitha Health), Harmony prenatal (Ariosa Diagnostics, Roche) and Panorama (Natera inc.).

These tests are able to screen plasmatic DNA for the most common aneuploidies, such as 21, 13 and 18-trisomy with a high rate of success and large scale validation (Gil MM, 2017).

These tests, however, do not replace invasive clinically performed procedures. They can be included into routine screening tests with the aim of addressing to invasive procedures, with their related risk of fetal loss, only patients with positive results from the non invasive test.

2. AIM OF THE STUDY

The aim of this study was the development of a bioinformatic pipeline for NGS data analysis and fetal genotype prediction for an in-house NIPD protocol for β-thalassemia, through next-generation sequencing of cell-free fetal DNA present in maternal plasma.

β-thalassemia is the most common autosomal recessive single-gene disorder in Sardinia, where approximately 10.3% of the population is a carrier.

The chr11:g.5226774G>A variant in the HBB gene (rs11549407, dbSNP), better known as the nonsense β°39 variant, is the most common variant, accounting for 95.7% of the β-thalassemia variants in Sardinians with an allele frequency of 4.8% (Saba L, 2017).

Due to the composition of the Sardinian population, a non invasive valid method is required to safely determine the status of the fetus in all couples at risk for β-thalassemia .

In this way, a workflow from sample collection and processing to data analysis results is proposed in this thesis.

Described method and reported results were recently published in the European Journal of Human Genetics (Saba L, 2017).

3. METHODS

A cohort of 39 couples at risk for β-thalassemia, who underwent prenatal diagnosis because the parents were carriers of the chr11:g.5226774G>A variant in the HBB gene (rs11549407, GRCh38 chr11:5226774), was recruited by the Screening and Genetic Counselling Service of the "Ospedale Pediatrico Microcitemico - A.Cao".

After counselling and a signature of informed consent, 20 ml of maternal peripheral blood were collected in EDTA-containing tubes prior to villocentesis, at a gestational age of 7 weeks to 14 weeks+3 days (mean 9 weeks +6 days).

Plasma samples were separated after whole blood centrifugation at 1,600 g for 10 min and at 16,000 g for 10 min, aliquoted into 1.5 ml tubes and finally frozen at −80 °C until cffDNA extraction.

Samples were processed according to the experimental workflow described in the next section and data analysis was performed through a dedicated bioinformatic pipeline.

3.1 MOLECULAR BIOLOGY

cffDNA samples were isolated from 8 ml of thawed plasma using the QIAamp Circulating Nucleic Acid Kit from Qiagen with a Qiagen vacuum manifold, following the manufacturer's protocol (Qiagen GmbH, Hilden, Germany). Final DNA was eluted into 150 µl of AVE buffer.

Parental genomic DNA was extracted from 500 µl of whole blood with a DiaSorin Blood DNA 500 extraction kit (DIASORIN S.P.A., SALUGGIA (VC), Italy) and NorDiag Arrow System (ISOGEN Life Science, Utrecht, The Netherlands).

The corresponding trophoblast DNA samples were obtained by villocentesis at 11–14 weeks of gestation. After maternal decidual tissue dissection, DNA was extracted from the trophoblast tissue samples with the salting-out method.

The chr11:g.5226774G>A variant was detected in both trophoblast DNA and in parental DNA using the Nuclear Laser Medicine Beta Globin Test kit (Nuclear Laser Medicine s.r.l, Italy)

The principle of the described approach is to use the cffDNA samples to infer the parental haplotypes most likely inherited by the foetus and establish the fetal HBB chr11:g.5226774G>A genotype accordingly.

The analysis of each cffDNA sample is preceded by the semiconductor sequencing of the corresponding trio of familial DNA samples (maternal, paternal and trophoblast) in a 62.7 kb target region of the β-globin gene cluster

(NC_000011.10 chr11: 5209079-5271945, GRCh38, UCSC Genome Browser) that contains both the HBB chr11:g.5226774G>A variant and SNPs that are potentially useful in determining the parental haplotype structure.

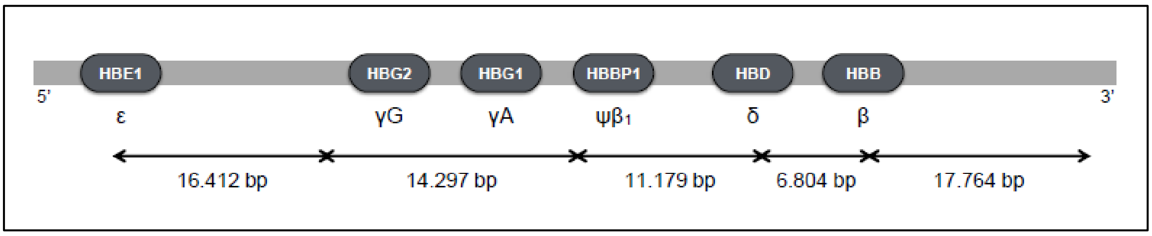


Figure 16. **β-globin gene cluster scheme**. Representation of β-globin gene cluster target region; the 5 arrows below represent the long amplicons positions. *(Saba L, 2017)*

SNPs are considered informative when both the mother and the father are homozygous for different alleles or when at least one parent is heterozygous.

For each cffDNA sample, short regions (80–120 bp) containing the informative SNPs and the HBB c.118C>T variant were individually amplified.

After pooling, the short amplicons were purified using Agencourt AMPure XP Reagent (Beckman Coulter, Brea, CA, USA), quantified with a Qubit dsDNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA) and subjected to library preparation as described in the Ion Xpress Plus Library Kit (Life Technologies) protocol. The adapter-ligated library was quantified using a High Sensitivity DNA kit and a 2100 Bioanalyzer (Agilent).

One hundred picomoles of the ligated pooled libraries was subjected to template preparation with the Ion OneTouch Template Kit and Ion OneTouch

System v2 (Life Technologies, Carlsbad, CA, USA).

Semiconductor sequencing was performed in 314/316 chips using the Ion PGM Sequencing 200 Kit v2 (Life Technologies) in a PGM system at 500 flows, in accordance with the manufacturer’s protocol.

3.2 PIPELINE

The bioinformatic pipeline is written in Python 2.7 programming language (Python Software Foundation., 2010). While a graphical interface has not been developed, pipeline execution can be performed through a linux teminal.

During the analysis, several open-source bioinformatic types of software are used, like BWA (Burrows-Wheeler Aligner) (Li H., 2009), SAMtools (Li H, 2009) (Li, 2011), SHAPEIT (Delaneau O Z. J., 2013), VCFtools (Danecek P, 2011), and some Python modules as NumPy (Van der Walt S, 2011) and Pandas (McKinney, 2010), specific for data frames and array managment. To correctly execute the pipeline, it is necessary to install and include them into the environment path.

Due to the restricted extension of the genomic region to be analyzed, the analysis can be performed on common personal computers at the expense of analysis time which is directly affected by computational performance.

The pipeline is composed of 8 scripts, involved in different steps of the analysis, which are subsequently called in a precise order by a master script.

The process is fully automatized and can be started by the execution of the main python script. This is made possible by a specific system of folders in which required input files must be placed and are expected to be located by the pipeline.

There are three main folders where are respectively located the samples, the pipeline with its scripts and analysis accessory database files.

In the samples folder, single family sequencing data is grouped in a subfolder labelled with the family ID. This practical subset allows the analysis to be performed for a single family by typing the desired family ID when launching the main script or on all samples present in the folder, the default parameter.

Analysis results files are located in the appropriate family folder.

Analysis accessory files are placed in the “reference” folder, which include the reference human genome sequence (Genome Reference Consortium Human Build 38), against which sequencing reads will be aligned, and a reference panel of haplotypes, described later.

Once the main script is executed and the target of the analysis has been specified , the system checks for sequencing data files existence in each family folder and then starts processing families based on IDs alphabetical order.

3.2.1 Script #1: PARENTS ANALYSIS AND PHASING

The first script processes parental NGS raw data to obtain a file containing parental haplotypes.

The first time the pipeline is executed, an additional step constructs, through BWA software, all dependencies and indexes for reference sequence FASTA file, required for subsequent sample reads alignment.

Once these files are made up and stored into the “reference” folder, it is possible to avoid repeating this step in the next analysis by “commenting out” the line of the script code where the function is called with the “#” character.

Analysis of samples starts with parental FASTQ files that are aligned to the reference sequence of chromosome 11 (chr11.fasta) with the MEM function of BWA software. Aligned reads are stored in the conventional SAM file format.

SAMtools software then converts SAM files into their binaries BAM files, which are sorted and indexed to allow rapid sequence queries.

This step is also useful because produced files allow sequenced regions visualization with software like IGV.

Through SAMtools and BCFTOOLS the pipeline performs the calling of variants represented in the reads producing separated variant call files (VCF) for the father and the mother.

Raw VCFs then undergo several filtering steps which are aimed at removing low quality data. Variants with quality scores lower than 10 are filtered out as ambiguous duplicated sites. Furthermore, to speed up the analysis, VCFs are cleaned from additional sequencing information that is not required in downstream analysis.

Single parental VCFs are then merged into a single file which is filtered for genomic positions present in the 1000 Genomes Project phase 3 panel (Sudmant PH, 2015). Simultaneously, the annotation of single nucleotide polymorphisms (SNPs) rsIDs is performed by dbSNP database interrogation (Sherry ST, 2001).

SHAPEIT software and a population-specific constructed reference panel of haplotypes, composed of 1000 Genomes Project samples and 39 Sardinian TRIOs, are used to phase parental genotypes.

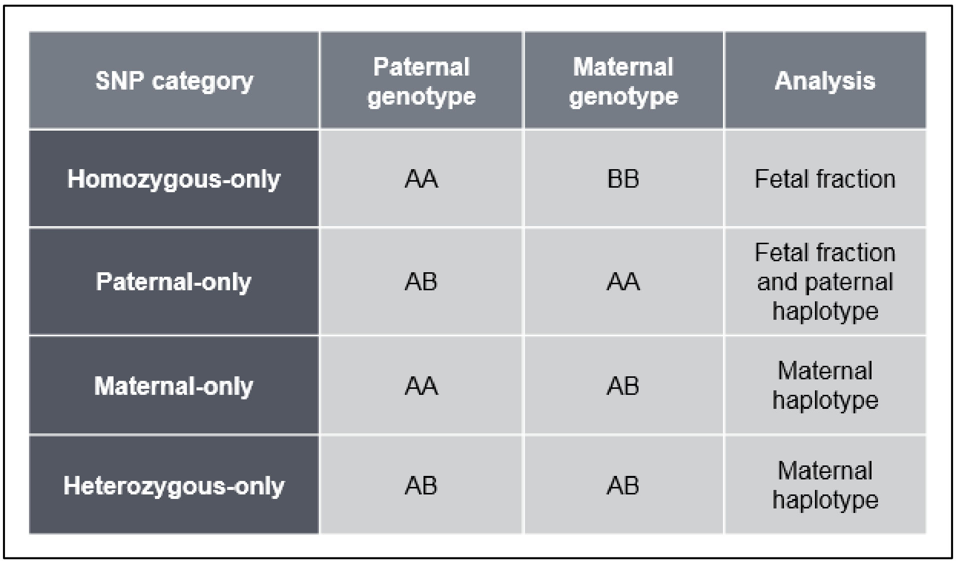
Haplotype construction accuracy is verified at the end of the analysis by comparing them with haplotypes determined by TRIO analysis.

Parental phased genotypes are stored in a VCF file which represents the scaffold for the determination of the paternal and maternal inherited haplotype by the fetus.

3.2.2 Script #2: SNPs CLASSIFICATION

Parental genotype combinations are determined by the second script of the pipeline which is aimed at finding and classify informative SNPs in 4 categories(Figure 17):

- homozygous-only: SNPs for which both parents are homozygous but for different alleles
- paternal-only: SNPs for which the father is heterozygous and the mother homozygous
- maternal-only: SNPs for which the mother is heterozygous and the father homozygous
- heterozygous-only: SNPs for which both parents are heterozygous



SNP category	Paternal genotype	Maternal genotype	Analysis
Homozygous-only	AA	BB	Fetal fraction
Paternal-only	AB	AA	Fetal fraction and paternal haplotype
Maternal-only	AA	AB	Maternal haplotype
Heterozygous-only	AB	AB	Maternal haplotype

Figure 17. **Informative SNPs categories**. Classification of informative SNPs involved in the analysis strategy and their relative usage.

For each category of SNPs a file containing the genomic coordinates and parental observed phased genotypes is generated.

These files will be used for fetal fraction estimation and paternal and maternal haplotype inheritance determination. Their absence results in the interruption of the analysis.

3.2.3 Script #3-4: Cell-free DNA sequence analysis

This script is responsible for cffDNA sample sequences analysis.

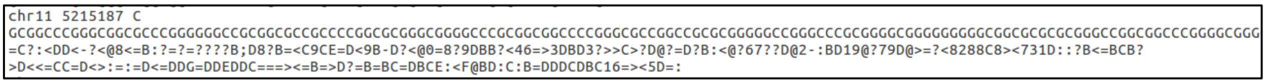
The steps and software involved for processing the FASTQ file until the generation of a sorted and indexed BAM file are the same as that used for parental analysis.

Due to the presence of fetal-maternal DNA mixture, performing a common calling of variants with the determination of genotypes, does not provide correct information. In fact, in cffDNA are represented the two distinct genotypes of the mother and the fetus that are not distinguishable by variant caller software.

For this reason the pipeline collects per-nucleotide resolution sequencing information for each genomic position sequenced by piling-up reads through SAMtools software “mpileup” command.

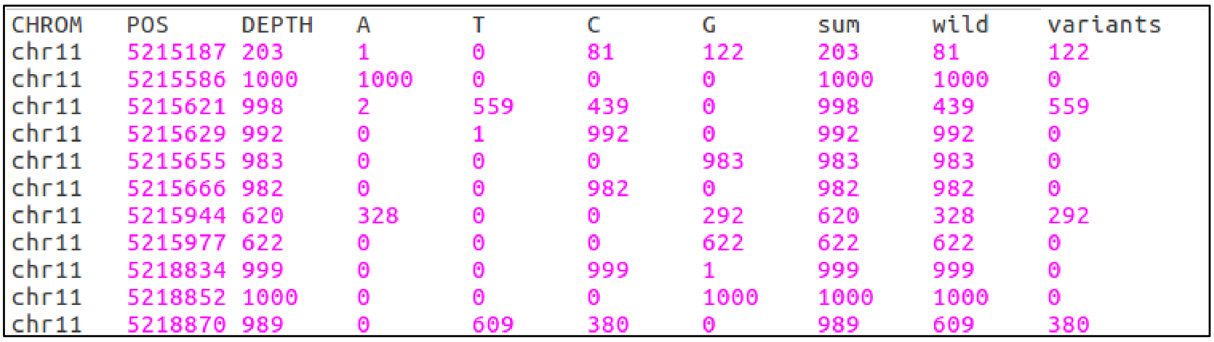
Generated output is in the pileup format (Figure 18), a tab separated text file with the complete list of observed nucleotides and the corresponding quality value for each sequenced genomic position.

To facilitate the analysis of information carried by this file, the pipeline converts it into a more manageable CSV format (Figure 19).



```
chr11 5215187 C
GCGGCCCGGGCGGCGCCCGGGGGGCCGCGGCGCCGCCCCGGCGCGGGCGGGGCCCGCGGCGGCCCCGGGCGCCGGCCGCGCGGGGGCCGGGCCCGCGGGGCGGGGGGGGCGGCGCGCGGGCCGGCGGCCCGGGGCGGG
=C?:<DD<-?<@8<=B:?=?=????B;D8?B=<C9CE=D<9B-D?<@0=8?9DBB?<46=>3DBD3?>>C>?D@?=D?B:<@?67??D@2-:BD19@?79D@>=?<8288C8><731D::?B<=BCB?
>D<<=CC=D<>:=:=D<=DDG=DDEDDC===><=B=>D?=B=BC=DBCE:<F@BD:C:B=DDDCDBC16=><5D=:
```

Figure 18. **Pileup file**. Example of pileup output file for a cffDNA sample for a single genomic position (chr11:5215187); in the first line are shown the reference chromosome and the genomic position with the reference allele. In the second line are listed the observed nucleotides for the genomic position. The single base quality is coded and recorded as string in the third line.



CHROM	POS	DEPTH	A	T	C	G	sum	wild	variants
chr11	5215187	203	1	0	81	122	203	81	122
chr11	5215586	1000	1000	0	0	0	1000	1000	0
chr11	5215621	998	2	559	439	0	998	439	559
chr11	5215629	992	0	1	992	0	992	992	0
chr11	5215655	983	0	0	0	983	983	983	0
chr11	5215666	982	0	0	982	0	982	982	0
chr11	5215944	620	328	0	0	292	620	328	292
chr11	5215977	622	0	0	0	622	622	622	0
chr11	5218834	999	0	0	999	1	999	999	0
chr11	5218852	1000	0	0	0	1000	1000	1000	0
chr11	5218870	989	0	609	380	0	989	609	380

Figure 19. **cffDNA CSV file**. Example of CSV file for a cffDNA sample obtained from pileup file conversion. For each record in the pileup file nucleotides are summarized and listed in the A-T-C-G columns, then according to the reference sequence are determined the numbers of REF and ALT observed alleles (columns “wild” and “variants”)

3.2.4 Script #5: cffDNA FRACTION DETERMINATION

The determination of fetal DNA concentration, by now called fetal fraction, is based on searching in cffDNA something that is not present in the maternal genome, such as Y chromosome sequences.

While the described method is based on the study of parental SNPs, the pipeline estimates the fetal fraction by analyzing two groups of SNPs:

- Paternal-only SNPs, in which the father is heterozygous and the mother homozygous
- Homozygous-only SNPs, in which both parents are homozygous but for a different allele.

A temporary data frame is generated by merging the files containing the above mentioned categories of SNPs with the previously obtained CSV file.

In this way it is possible to determine, for each informative site included in the data frame, the paternal allele fraction as the number of reads for the allele not shared with the maternal genome on the total number of reads for that site.

The fetal fraction is then calculated as the double of the mean of sites which show paternal allele fraction between 0.014 and 0.11 (accounting for a hypothetical fetal fraction ranging between 2.8% and 22%), according to the formula:

$$f = \frac{2y}{x+y}$$

where y represents the paternal allele counts and x the maternal ones.

The presence of less than 2 SNPs from the above mentioned categories induces a stop of the analysis because of the importance of fetal fraction on which the determination of the paternal and maternal transmitted haplotypes are based.

3.2.5 Script #6: PATERNALLY TRANSMITTED HAPLOTYPE DETERMINATION

Paternally transmitted haplotype determination is performed through the analysis of paternal-only SNPs present in cffDNA sequences with a sequencing DEPTH greater than 1000.

Ideally, if the paternal allele not shared with the mother is transmitted to the fetus, $N\, x \frac{f}{2}$ reads are expected for that allele, otherwise the absence of the paternal only allele is expected.

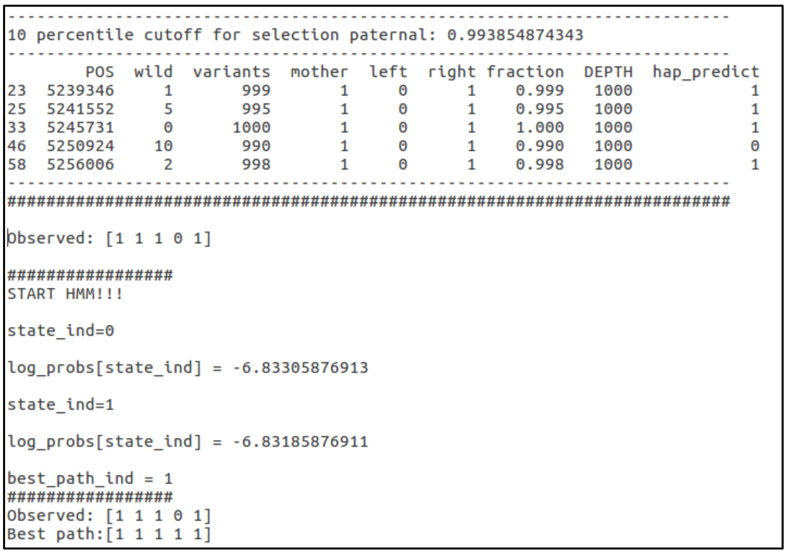
For the discrimination of paternal-only allele presence/absence, a cutoff value is set as the 10th percentile of 100 equally distributed parts of a numeric series with start point 0 and endpoint fetal fraction.

For each paternal-only SNP, if the paternal allele fraction is below this cutoff, the presence of the paternal-only allele is inferred, otherwise the pipeline infers its absence. In this way a site-by-site path of left/right paternally transmitted haplotype is constructed.

Although the generated path generally permits the determination of the transmitted haplotype by simple comparison of the number of sites accounting for the left or right haplotype, the Viterbi Algorithm is applied to obtain a more robust result, based on a probabilistic model, and overstep situations in which an observation-derived path perfectly splits between the two haplotypes.

The algorithm performs a computation of the most probable path by emitting a probability associated within each state, represented by the left or the right haplotype.

Starting from the same probability for both states, the algorithm analyzes each observed value of the path considering all possible combinations and assigning them a probability value. Finally the most probable path is determined as the one that has the higher cumulative probability, indicating the predicted transmitted haplotype.



```
--------------------------------------------------------------------------
10 percentile cutoff for selection paternal: 0.993854874343
--------------------------------------------------------------------------
        POS  wild  variants  mother  left  right fraction  DEPTH  hap_predict
23  5239346     1       999       1     0      1    0.999   1000            1
25  5241552     5       995       1     0      1    0.995   1000            1
33  5245731     0      1000       1     0      1    1.000   1000            1
46  5250924    10       990       1     0      1    0.990   1000            0
58  5256006     2       998       1     0      1    0.998   1000            1
--------------------------------------------------------------------------
##########################################################################

Observed: [1 1 1 0 1]

#################
START HMM!!!

state_ind=0

log_probs[state_ind] = -6.83305876913

state_ind=1

log_probs[state_ind] = -6.83185876911

best_path_ind = 1
#################
Observed: [1 1 1 0 1]
Best path:[1 1 1 1 1]
```

Figure 20. **Paternal Haplotype Prediction**. A section of the report file including the paternal haplotype prediction. In the first row is represented the 10thpercentile cutoff. For each record in the second section is determined the inference of the “left” or “right” paternal haplotype (respectively coded as “0” and “1” in the “hap_predict” column) according to the cutoff. The “Observed” path is then corrected by Viterbi Algorithm which indicates the haplotype with the higher cumulative probability (in this case “1”, the “right” paternal haplotype).

3.2.6 Script #7: MATERNALLY TRANSMITTED HAPLOTYPE DETERMINATION

For the prediction of the maternally inherited haplotype the pipeline takes into account SNPs for which the mother is heterozygous and the father homozygous (maternal-only SNPs).

Due to the restricted number of these sites, the SNPs for which both parents are heterozygous (heterozygous-only SNPs) are included by constraining maternally transmitted haplotype determination to the paternal one.

The knowledge of the paternally transmitted haplotype permits the consideration of heterozygous-only SNPs as they were maternal-only, because only the fetal inheritance from the mother needs to be determined.

With this strategy it is possible to increase the number of informative sites available for this part of the analysis, which is the most challenging.

The high variability introduced by cffDNA short fragments amplification, which results in different sequencing depths, represent the principal cause of the introduction of a depth cutoff of 1500.

This cutoff is higher than that used for paternal-only SNPs filtering (set to 1000) because of the different strategy adopted for the haplotype determination, which is based on allelic imbalance evaluation rather than different allele presence/absence discrimination.

The analysis proceeds with the allelic fraction percentage calculation for each site passing the filter step as the number of reads accounting for each allele on the total number of reads.

Ideally, a heterozygous genotype for a given allele is represented by a percentage of 50% for the wild-type allele and 50% for the variant. The fetal-maternal DNA mixture determines fetal fraction-related changes in this balance depending on the fetal genotype.

As reported by Kitzman et al, the pipeline calculates two probabilities for the homozygous or heterozygous status of the fetus depending on the inherited maternal allele (Kitzman JO, 2012).

The probabilities p1 and p2 indicate the single success probability and are calculated with the formulas:

$$p1 = \left(\frac{1-F}{2} + \frac{F}{2} + \frac{F}{2}\right)$$

where $\frac{1-F}{2}$ represents maternal DNA fraction, $\frac{F}{2}$ respectively paternal and maternal contributions. Probability p1 is related to the inheritance of the maternal allele shared with paternal genome and then to a fetal homozygous genotype.

Otherwise, p2 is the probability that the maternally inherited allele is different from the paternally inherited one (fetus heterozygous) and is calculated with the formula:

$$p2 = \left(\frac{1-F}{2} + \frac{F}{2}\right) == 0.5$$

Starting from these two values, the pipeline makes p1/p2 selection for each site based on paternal observed allele and allelic imbalance in cffDNA.

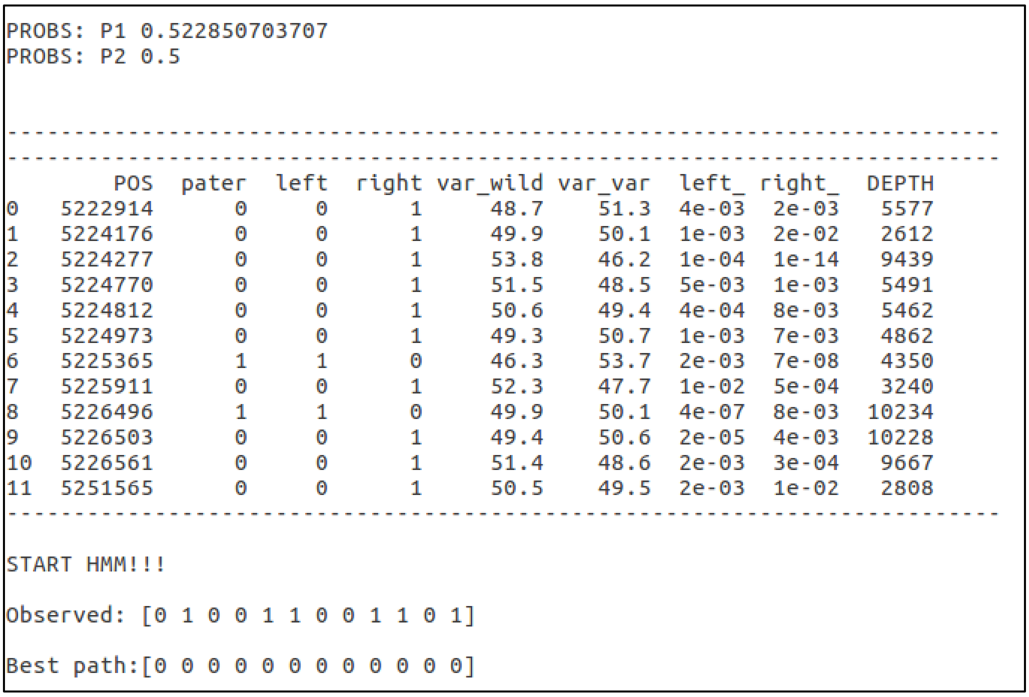
Then the probability that the fetus inherited the left or right haplotype is computed as a binomial function of the number of reads supporting allelic imbalance (*k*), the total number of reads for that site (*n*) and the single-event probability p1 or p2 (*p*) with the formula:

$$BINOM.PMF(k, n, p)$$

Binomial Probability Mass Function determines the probability of getting exactly *k* successes in n trials given that the probability of a success is p.

The maternally transmitted haplotype is site-by-site hypothesized as left or right based on the higher probability value obtained.

The generated path is then corrected, as for paternal haplotype prediction, through the Viterbi Algorithm which establishes the maternally inherited haplotype with higher cumulative probability.



```
PROBS: P1 0.522850703707
PROBS: P2 0.5


---------------------------------------------------------------------------
---------------------------------------------------------------------------
        POS  pater  left  right var_wild var_var  left_ right_  DEPTH
0    5222914     0     0     1     48.7    51.3 4e-03  2e-03   5577
1    5224176     0     0     1     49.9    50.1 1e-03  2e-02   2612
2    5224277     0     0     1     53.8    46.2 1e-04  1e-14   9439
3    5224770     0     0     1     51.5    48.5 5e-03  1e-03   5491
4    5224812     0     0     1     50.6    49.4 4e-04  8e-03   5462
5    5224973     0     0     1     49.3    50.7 1e-03  7e-03   4862
6    5225365     1     1     0     46.3    53.7 2e-03  7e-08   4350
7    5225911     0     0     1     52.3    47.7 1e-02  5e-04   3240
8    5226496     1     1     0     49.9    50.1 4e-07  8e-03  10234
9    5226503     0     0     1     49.4    50.6 2e-05  4e-03  10228
10   5226561     0     0     1     51.4    48.6 2e-03  3e-04   9667
11   5251565     0     0     1     50.5    49.5 2e-03  1e-02   2808
---------------------------------------------------------------------------

START HMM!!!

Observed: [0 1 0 0 1 1 0 0 1 1 0 1]

Best path:[0 0 0 0 0 0 0 0 0 0 0 0]
```

Figure 21. **Maternal Haplotype Prediction**. A section of the report file including the maternal haplotype prediction. After single success probabilities p1 and p2 estimation, for each record is computed a binomial probability for maternal left and right haplotypes (columns “left_” and “right_”) based on the observed allelic imbalance (columns “var_wild” and “var_var”). Then a site-by-site higher probability prediction path is generated (“Observed” line; “0” and ”1” refers to left and right haplotype). The “best path” line represents the output of Viterbi Algorithm correction.

3.2.7 Script #8: FINAL RESULT

Finally the pipeline uses paternally and maternally determined transmitted haplotypes to build fetal haplotypes as a simple “copy and paste”.

Results are stored in a file in which parental and fetal predicted haplotypes are present.

Furthermore, this script generates a text-based report with analysis information, such as the fetal fraction, the number and the IDs of the SNPs used to determine the paternally and maternally transmitted haplotypes, and the predicted fetal genotype for the investigated variant site.

Several additional scripts have been developed for the validation of the obtained results and the estimation of parental constructed haplotypes correctness. These two scripts are not included in the pipeline because they are based on the analysis of fetal DNA obtained via invasive procedures and its comparison with pipeline produced results.

Although necessary for the validation of the analysis method in the early stage of the project, CVS-obtained fetal DNA analysis is not part of the analysis workflow which is entirely based on a non invasive approach.

To underline the above mentioned concept, additional scripts are described separately.

3.2.8 RESULTS VALIDATION SCRIPT

Result validation is done by comparison of fetal genotypes determined by the pipeline with those from the corrispective CVS-obtained fetal DNA analysis.

This script firstly analyzes fetal DNA sequences to produce, starting from a FASTQ file, a VCF with fetal genotypes through the same steps performed by the pipeline for parental analysis (script #1).

Then a site-by-site comparison of obtained fetal genotypes with the pipeline determined ones is performed, resulting in a genotype correctness percentage.

However, the analysis is considered correct only if the predicted fetal genotype for the investigated variant perfectly matches the fetal genotype.

Further considerations are provided in the discussion section of the thesis.

3.2.9 PARENTAL PHASING VALIDATION SCRIPT

The correct determination of fetal haplotypes by the pipeline is strongly related to the construction of parental haplotypes. Errors in this step might result in incorrect fetal genotypes if parental alleles are linked to the wrong haplotype, affecting the result of the analysis.

Given that the pipeline constructs paternal and maternal haplotypes considering parents as single unrelated individuals and with the auxilium of a reference panel of haplotypes, a control step is required to ascertain their correctness.

This script determines the correct parental haplotypes by analyzing and phasing parents jointly with CVS-obtained fetal DNA. The analysis of the family TRIO (father-mother-child) with the specification of sample relationships permits the correction of statistically determined parental haplotypes based on inheritance. In this way, the comparison of obtained haplotypes with those constructed by the pipeline provides an estimation of identity percentage between them.

Finally, the results produced are stored in a text-based report that summarizes analysis information and allows an easy interpretation of results.

3.3 POPULATION-SPECIFIC HAPLOTYPE REFERENCE PANEL CREATION

The construction of parental haplotypes by the pipeline is based on a reference panel of haplotypes.

Ideally, a panel comprising individuals from 26 populations, as the dataset provided by 1000 Genomes Project phase 3, should provide sufficient information to determine individual haplotypes (Delaneau O, Marchini J, The 1000 Genome Project Consortium, 2014) (Sudmant PH, 2015).

Unfortunately, the 2504 included samples does not represent the particular genetic composition of the Sardinian population. More specifically, the investigated variant in the HBB gene (rs11549407), which accounts for 95.7% of the β-thalassemia variants in Sardinians, is present in only one single Mexican individual.

To increase the representation of this variant and, more generally, of Sardinian haplotypes, a population-specific reference panel was constructed by integrating the 1000 Genomes Project phase 3 panel with 117 Sardinian samples, corresponding to the 39 TRIOs (father,mother,child) sequenced and analyzed in the project.

All samples were then phased together by the SHAPEIT algorithms, with the integration of sample relationships and Phase Informative Reads (PIRs) (Delaneau O H. B.-F., 2013).

The importance of PIRs relies on the fact that short sequence reads can contain phase information if they span two or more heterozygous genotypes which can be used to correct statistical phasing.

PIRs can be obtained from BAM files through extractPIRs software, which generates a file containing all PIRs information.

During the phasing process, SHAPEIT software phases all samples present in the file with its algorithms. After several cycles of phasing, SHAPEIT corrects generated phases with PIRs information and correlation between samples provided in a pedigree file.

The final output is a reference panel of haplotypes which is used by the pipeline to construct parental haplotypes.

4. RESULTS

Sequencing data was analyzed through the previously described bioinformatic pipeline to determine in cffDNA the paternal and maternal haplotypes inherited by the fetus and its genotype status for the causative mutation chr11:g.5226774G>A.

As shown in Figure 22 and Figure 23, semiconductor sequencing produced on average 154 494 reads per individual parental sample, with a mean depth of 358. Conversely, 423 105 reads were produced on average for each cffDNA sample, with an increased sequencing depth of 6902.73. Further information in Supplementary Table 1, Supplementary Figure 1Supplementary Figure 1 and Supplementary Figure 2.

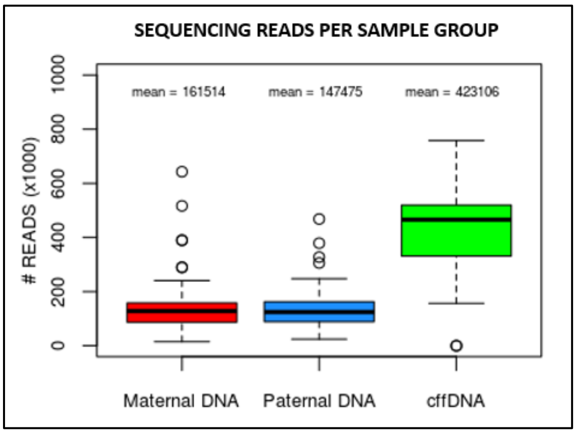


Figure 22. **Sequencing reads**. Boxplots show the distributions of the number of sequencing reads (x1000) per sample group.

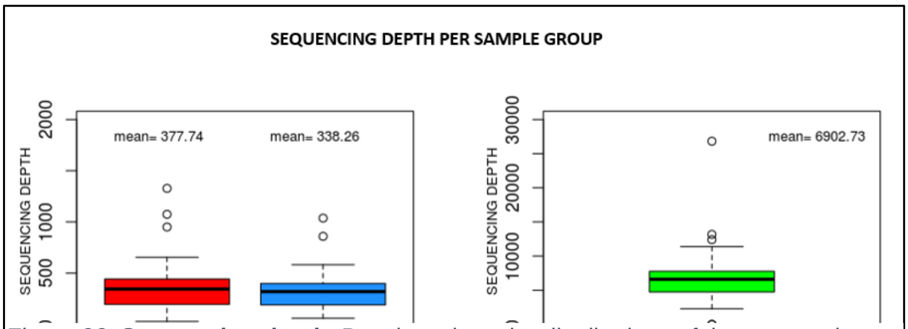


Figure 23. **Sequencing depth**. Boxplots show the distributions of the sequencing depth per sample group. Note that cffDNA boxplot has a different y-axis scale.

Read length distribution examples are showed in Figure 24.

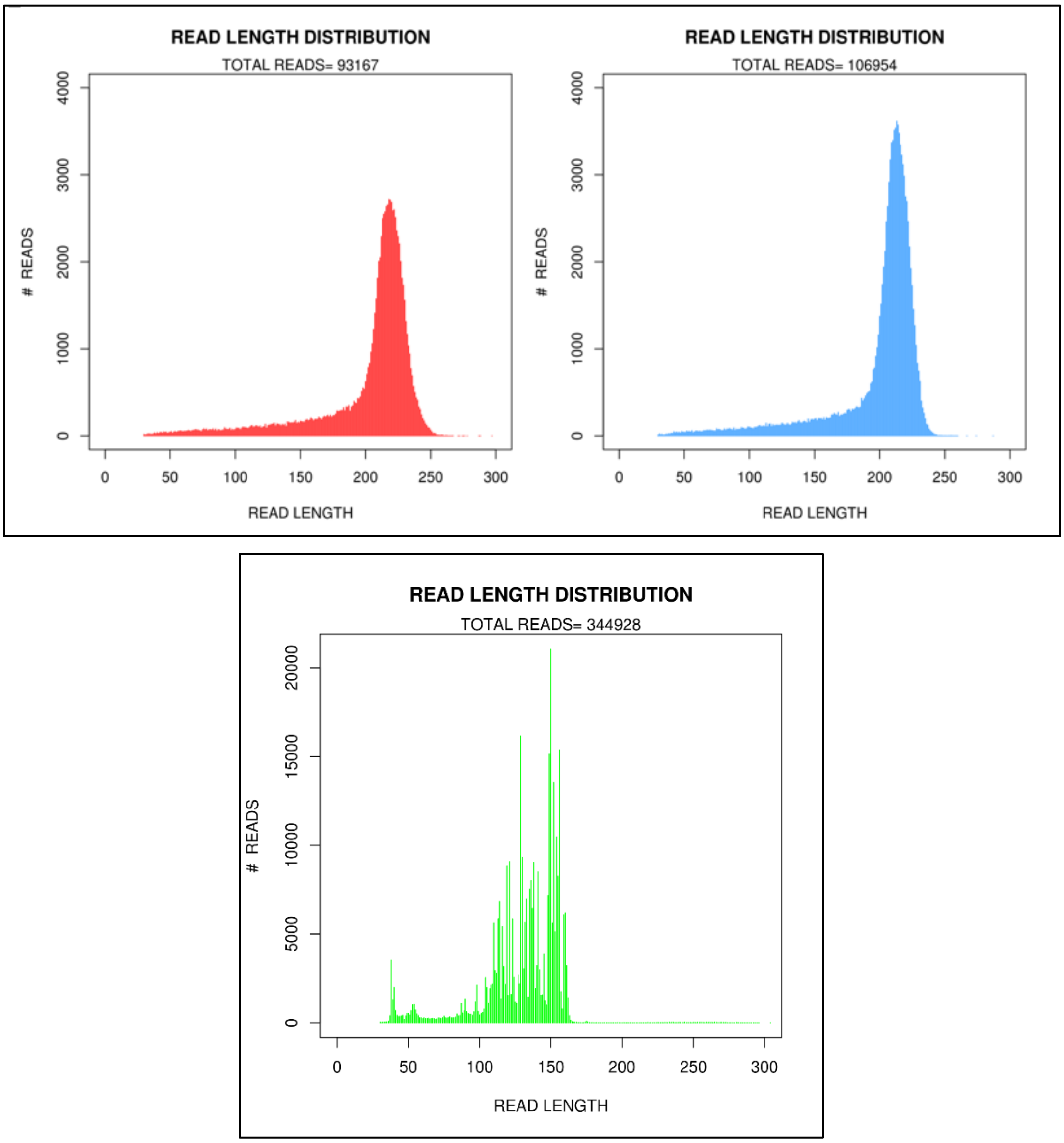


Figure 24. **Read Length Distribution**. Example of read length distributions in a family. Maternal (**red**) and Paternal (**blue**) derived reads present a similar distribution, with a prevalence of 200-250 bp reads determined by library preparation size selection step. cffDNA (**green**) presents a different distribution from parental sequences due to short-amplicon strategy used for library prepatation.

The total number of informative SNPs identified in the processed parental DNAs varied greatly and ranged from 58 to 180, with a mean value of 136, with an overall distribution across SNP categories summarized in Figure 25.

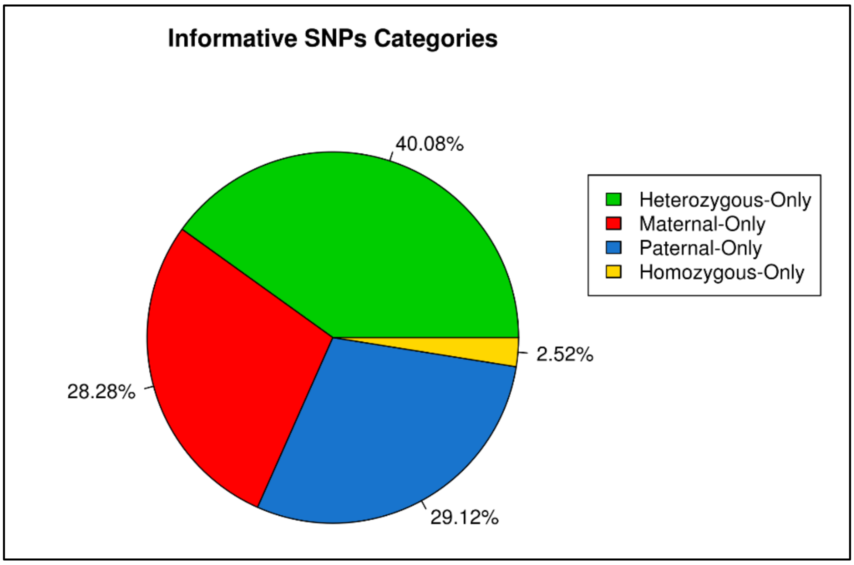


Figure 25. **Informative SNPs Categories**. The pie chart shows the average composition of informative SNPs in 39 Sardinian families.

Frequency of heterozygosity among all Sardinian parental samples was evaluated for all informative SNPs for which at least one individual was heterozygous.

To overcome the absence of genotype information for a given site and in such individuals, the frequency was calculated on the total number of determined genotypes for each SNP, to avoid over or underestimation due to lack of information.

Supplementary Figure 4 shows the first 100 variants with highest heterozygosity in the analyzed cohort of samples.

The analysis was completely performed by the pipeline in 30 out of 37 cffDNA samples. In the remaining 7 samples the analysis was interrupted because they did not respect analysis requirements. Further information is provided in the next section.

In each plasma sample, fetal fraction was determined by the pipeline from the mean fractional read depth calculated in each plasma sample at the paternal-only and homozygous-only SNPs. Figure 26 shows fetal fraction across all samples, which ranged from 3.7 to 12.6%, with a mean value of 6.96%.

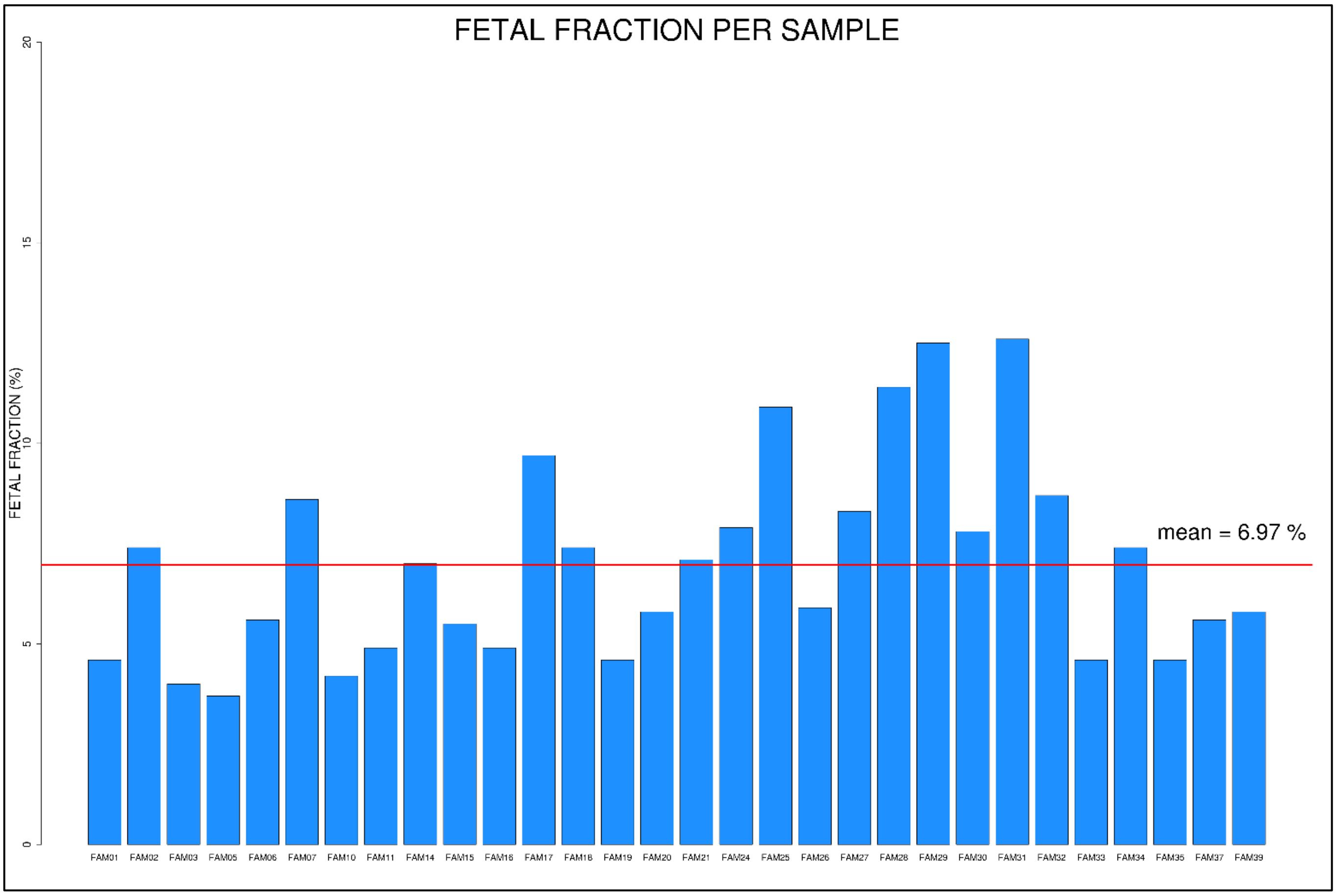


Figure 26. **Determined Fetal Fraction**. Barplot of Fetal Fraction % determined in al cffDNA samples for which the pipeline completed the analysis. The red line represents the average value.

The determination of the paternal and maternal transmitted haplotypes was performed by analyzing informative SNPs present in cffDNA sequences.

Starting from informative sites identified in parental sequences, the number of available SNPs for the analysis of cffDNA was reduced due to their effective presence in cffDNA sequences and quality filtering.

Figure 27 represents the distribution of the sites used for fetal haplotypes determination across all families (view also Supplementary Figure 5).

SNPs used for paternal prediction ranged from 1 to 34 with a mean value of 13.17. Conversely, maternal prediction was performed on an average of 17.3 maternal-only and heterozygous-only SNPs, ranging from 1 to 33.

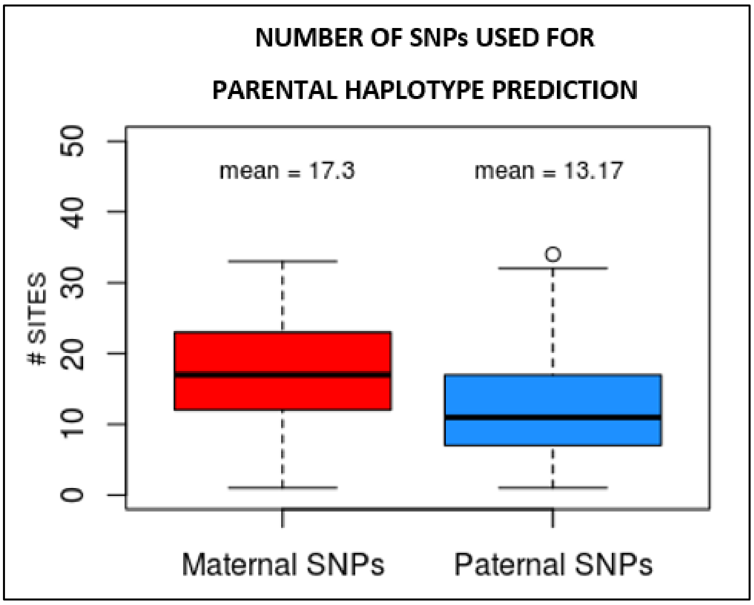


Figure 27. **SNPs used for parental haplotype prediction**. The barplots show the distribution of SNPs used for the determination of the maternal (red) and paternal (blue) transmitted haplotype in all families.

Finally, the pipeline permitted the determination of the correct fetal HBB genotype in 80% of analyzed cffDNA samples (24 out of 30).

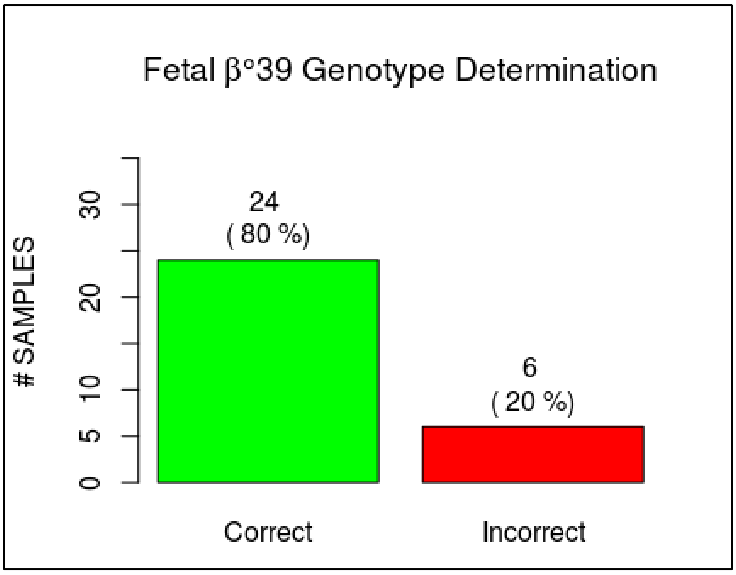


Figure 28. **Fetal $\beta^0$39 Genotype Determination Results**. General resuts of HBB genotype prediction in 30 analyzed cffDNA samples.

No relation was found between analysis results and fetal genotype, suggesting that incorrect genotype determination was not caused by a systematic error specific to the analysis method.

The study of parental haplotypes revealed that the paternally inherited haplotype was correctly determined in 100% of samples, indicating that incorrect genotype outcome was a consequence of incorrect inference of maternal transmitted haplotype (Figure 29).

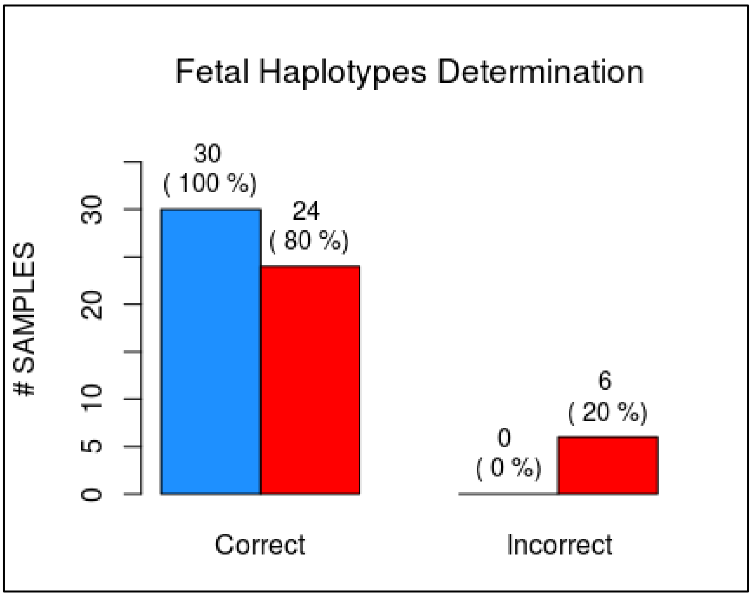


Figure 29. **Fetal Haplotypes Determination Results**. The blue and red bars represent the paternally and maternally inherited haplotypes, respectively.

Furthermore, it was possible to estimate parental haplotypes correctness by comparing them with those determined by TRIO analysis, with a percentage ranging between 94.3% and 100%.

5. DISCUSSION

The study of cell-free fetal DNA present in maternal blood during gestation for the Non Invasive Prenatal Diagnosis of a single gene disorder, such as β-thalassemia, has shown to be a safe and valid strategy for the early determination of the fetal genotype.

Although several improvements are required in terms of accuracy, obtained preliminary results are encouraging, especially in relation to the genetic composition of Sardinian population and the consequent causative mutation frequency in the analyzed cohort.

To test the feasibility and the accuracy of both experimental and bioinformatic analysis workflows in the three groups of samples (wild type, heterozygous or homozygous for the chr11:g.5226774G>A variant), plasma samples were selected and extracted retrospectively after completion of invasive prenatal diagnosis.

Of them, 15 samples were wild type, 15 were heterozygous, and 9 were homozygous.

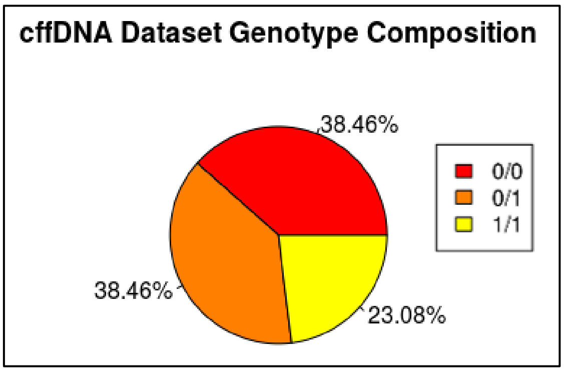


Figure 30. **cffDNA Dataset Genotype Composition**. The pie chart summarizes the fetal $\beta^0$39 genotype representation in all samples included in the study

Two out of 39 plasma samples were not analyzed due to technical issues during the sequencing process.

Both experimental and analysis workflows are summarized in Figure 31 and Figure 32.

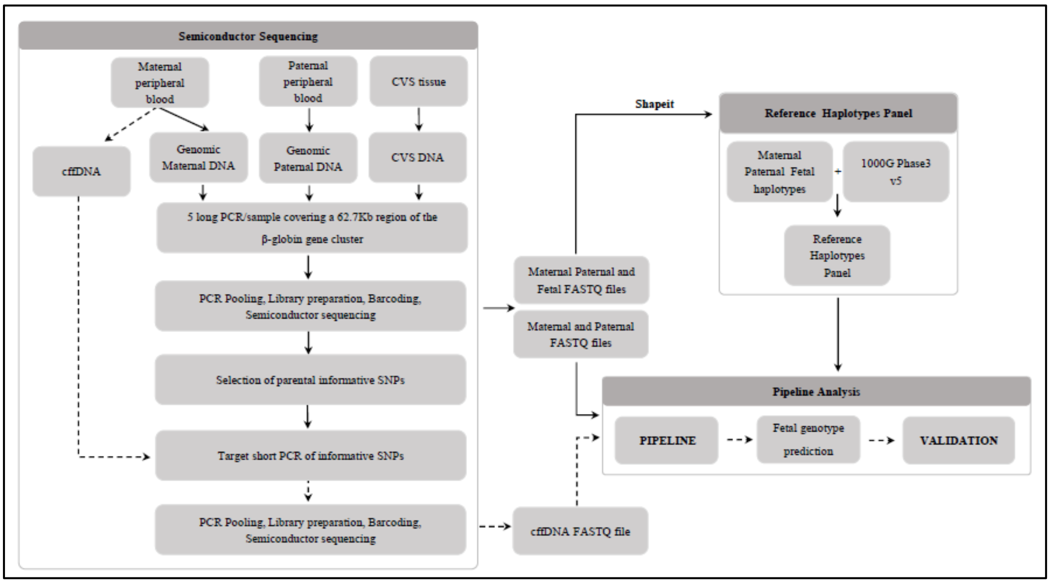

Semiconductor Sequencing
Maternal peripheral blood
Paternal peripheral blood
CVS tissue
cffDNA
Genomic Maternal DNA
Genomic Paternal DNA
CVS DNA
5 long PCR/sample covering a 62.7Kb region of the β-globin gene cluster
PCR Pooling, Library preparation, Barcoding, Semiconductor sequencing
Selection of parental informative SNPs
Target short PCR of informative SNPs
PCR Pooling, Library preparation, Barcoding, Semiconductor sequencing
Shapeit
Reference Haplotypes Panel
Maternal Paternal Fetal haplotypes
+
1000G Phase3 v5
Reference Haplotypes Panel
Maternal Paternal and Fetal FASTQ files
Maternal and Paternal FASTQ files
cffDNA FASTQ file
Pipeline Analysis
PIPELINE
Fetal genotype prediction
VALIDATION

Figure 31. **NIPD of β-thalassemia protocol overview**. A scheme of experimental workflow (left part) and bioinformatic analysis (on the right). *(Saba L, 2017)*

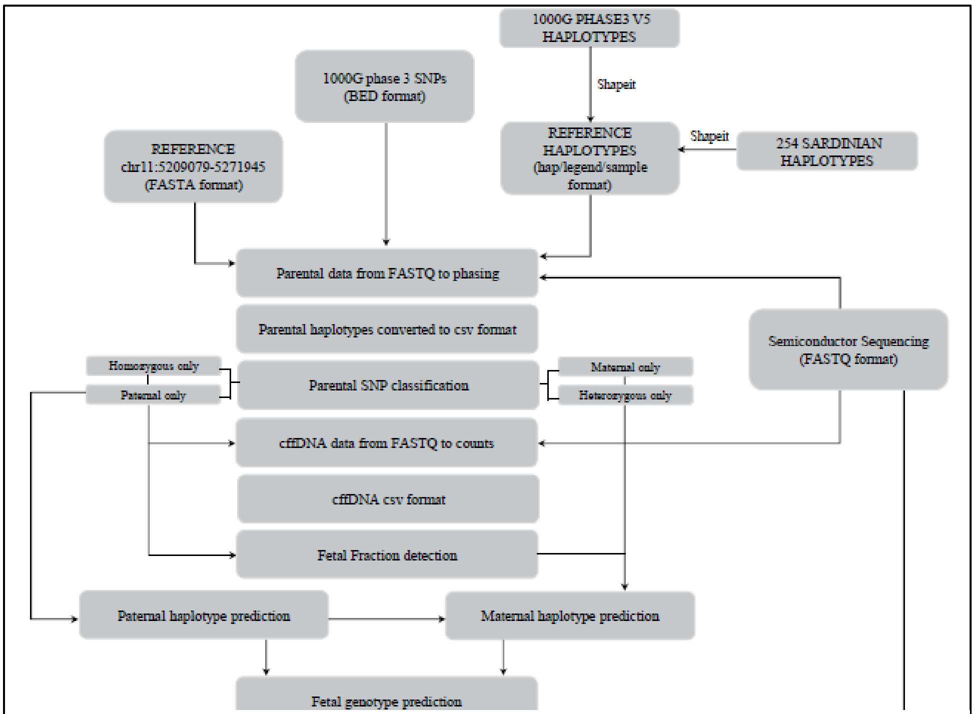

1000G PHASE3 V5 HAPLOTYPES
Shapeit
1000G phase 3 SNPs (BED format)
REFERENCE chr11:5209079-5271945 (FASTA format)
REFERENCE HAPLOTYPES (hap/legend/sample format)
Shapeit
254 SARDINIAN HAPLOTYPES
Parental data from FASTQ to phasing
Parental haplotypes converted to csv format
Semiconductor Sequencing (FASTQ format)
Homozygous only
Paternal only
Parental SNP classification
Maternal only
Heterozygous only
cffDNA data from FASTQ to counts
cffDNA csv format
Fetal Fraction detection
Paternal haplotype prediction
Maternal haplotype prediction
Fetal genotype prediction

Figure 32. **Pipeline overview**. A scheme of the bioinformatic analysis performed by the pipeline to determine the fetal HBB genotype. *(Saba L, 2017)*

Parental DNAs were amplified for a 62.7 Kb target region in the β-globin gene cluster (Figure 16) and then sequenced through the NGS instrument Ion Torrent PGM.

For each family, VCF files produced by the Torrent Variant Caller (TVC) plug-in of sequencing instrument integrated software (Ion Reporter) were analyzed to find informative SNPs.

Corrispective plasma DNA samples were amplified with a short amplicons strategy for genomic regions including one or more informative sites. Target regions were deeply sequenced, on average 6902X, to increase the resolution of sequences present in the fetal-maternal mixture.

The weakness of this strategy relies on the fact that, starting from the number of informative SNPs identified in parental sequences, on average only 52.7% of the SNPs could be effectively sequenced and analyzed in the corresponding plasma samples.

The presence of highly homologous genes, such as HBG1 and HBG2, and an L1 repeat element located 3' to the β-globin gene greatly hindered the design of specific primers that could yield amplicons shorter than 120 bp.

Furthermore, the number of SNPs available for the determination of the paternally and maternally transmitted haplotype was reduced to 50% by the pipeline which filtered out sites with sequencing depths lower than 1000 (paternal-only SNPs) or 1500 (maternal-only and heterozygous-only SNPs).

In addition, only sites with an allelic imbalance greater than 54% were used for the determination of maternal haplotype inherited by the fetus.

These threshold values were established during the pipeline validation as they provided the highest detection rate of the fetal genotypes.

Another reduction of possible informative SNPs is determined by the parental genotypes phasing process. In fact, parental haplotypes are constructed through SHAPEIT software with a reference panel composed of the 1000 Genome Project phase 3 v2 haplotypes and 39 Sardinian TRIOs.

Several variants observed in Sardinian samples were not included in the 2,504 individuals panel. According to the large difference in sample size, these variants were not taken into account and removed from the analysis rather than impute them in the bigger panel.

For this reason, SNPs which have shown to be informative for several families and that were not included in the reference panel were excluded, further reducing the number of available SNPs.

Most plasma samples satisfied basic data requirements and the pipeline was able to perform the analysis. On average the fetal haplotypes determination was performed on 13 and 17 sites, respectively for paternal and maternal prediction.

For 7 out of 37 samples the analysis was interrupted because of a lack of either paternal-only SNPs (two samples) or informative SNPs useful in calculating the fetal fraction (five samples).

Given the dependency of maternally transmitted haplotype determination from paternal ones, fetal fraction estimation represents a fundamental step because the cutoff value used by the pipeline in the inference of presence/absence of the paternal specific allele, suggesting the inheritance of one haplotype instead of another, is fetal fraction-related.

It is clear that a lack of this parameter due to the absence of informative SNPs or a determined value below the pipeline detection limit (set at 2.8%), as the absence of SNPs on which paternal transmitted haplotype is performed, leads to a stop in the downstream analysis.

In the remaining 30 samples, the number of informative SNPs passing the filter step was adequate to perform the analysis.

Fetal fraction was determined for each sample through paternal-only and homozygous-only SNPs analysis. Estimated values ranged from 3.7% to 12.6%, with an average of 6.96% and SD 2.5.

Starting from this parameter, the pipeline determined the most probable paternally transmitted haplotype after correcting the path generated from observations through the Viterbi algorithm.

Due to the reduced number of maternal-only SNPs, paternal haplotypes were used as scaffold for the determination of the maternal haplotype inherited by the fetus. With this strategy, the allelic imbalance evaluation was performed on maternal-only sites together with heterozygous-only SNPs.

For each site the probability that the fetus inherited the maternal-only allele or the allele shared with the father was determined, accounting for respectively heterozygous or homozygous fetal genotype.

Binomial Probability Mass Function was parametrized on the number of reads for a given allele, the total number of reads for that SNP and the fetal fraction-related single event probability p1 or p2 that the fetus is respectively homozygous or heterozygous.

As occurred for paternal haplotype determination, the generated path of states accounting for maternal haplotypes was subsequently corrected through the Viterbi algorithm.

Finally paternally and maternally inherited haplotypes of the fetus were determined as those with the highest cumulative probability.

Obtained results were then compared with CVS obtained fetal DNA to ascertain the correctness of the predicted fetal genotype for the disease-related variant and, more widely, to estimate parental and fetal constructed haplotypes accuracy.

The fetal genotype for the chr11:g.5226774G>A variant was correctly determined in 80% of samples, with a fetal genotype correspondence of 99.8% on average.

Conversely, in 6 samples out of 37 (20%) the pipeline determined an incorrect fetal genotype for the investigated variant. For these samples a genotype correctness of 94.5% was observed.

Generally, no association was found between the fetal genotype status and the analysis correctness, excluding a possible limit of the analysis method related to a specific class of genotype.

To thoroughly analyze obtained results and identify the causes of wrongly predicted fetal genotypes, parental haplotypes analysis was performed.

The availability of fetal DNA obtained from CVS allowed the construction of parental haplotypes with SHAPEIT by including pedigree information.

Each produced haplotype was labelled as “N” or “M” to indicate respectively the presence of wild-type or variant allele for the investigated variant.

Given that the fetus inherits one of the two paternal and maternal haplotypes, this strategy permitted the tracking of variant inheritance and then to establish the composition of transmitted haplotypes.

A simple comparison of these haplotypes with parental haplotypes constructed by the pipeline, allowed the estimation of their percentage of correctness, which should be related to analysis accuracy.

Parental haplotypes were constructed with an average accuracy of 99.6% in samples for which the pipeline completed the analysis, confirming the importance of a population-specific contribution in a reference panel of haplotypes.

Despite the high accuracy observed, even a single error in genotype phasing could lead to a wrong haplotype prediction in samples with a few informative SNPs.

However, the availability of a larger number of Sardinian individuals should provide a better representation of population haplotypes, determining an increase in phasing accuracy.

Once the overall correctness of parental haplotypes was established, fetal inheritance was thoroughly investigated through a “variant tracking” method.

For each family, the trio analysis established which paternal and maternal allele of the disease-associated variant was transmitted to the fetus.

The allele transmission for the same variant was investigated in the fetal haplotypes determined by the pipeline and then a comparison of results was performed.

With this strategy it was possible to establish that the paternal haplotype was correctly predicted in 100% of samples.

On the other hand, the incorrect fetal genotype prediction in 6 out of 30 samples was imputable to the wrong maternally inherited haplotype determination.

In 5 out of 6 cases, the incorrect prediction of the inherited maternal haplotype was caused by the presence of unexpected allelic imbalances at a number of sites sequenced in the cffDNA samples, which resulted in an incorrect HMM inference of the correct haplotype.

In the sixth sample, the erroneous maternal haplotype inference was determined by the incorrect construction of the maternal haplotypes by the pipeline, with an average correctness of 99.08%.

Even though several improvements are required to increase the predictive power of the analysis method, obtained results permit us to assess the feasibility of cffDNA analysis with an haplotype-based strategy for the NIPD of a monogenic disease such as β-thalassemia in the isolated Sardinian population.

6. CONCLUSIONS

The developed protocol for NIPD of β-thalassemia through target NGS of cffDNA coupled with bioinformatic analysis has shown to be a valid strategy to perform fetal genotype prediction in families where both parents are carriers for the same causative mutation.

The particularity of this approach relies in samples mutational status.

In fact, the literature described studies involved families with parent carriers of different causative mutations and so, while maternally inherited mutation identification is challenging, it is quite easy to ascertain if the fetus inherited the paternal ones (Zafari M, 2016).

Although obtained results are encouraging, several improvements are required for both the laboratory and the analysis part of the workflow, to reach high predictive power and statistical significance compatible with its translation into clinical practice.

To include in the analysis a higher number of informative SNPs, we will extend sequenced region up to 250Kb in the β-globin gene cluster.

The number of paternal-only and maternal-only SNPs is in fact a limiting factor for the analysis, because most of them are filtered out due to low sequencing depth and or base quality. As this filters are necessary to perform the analysis on high quality data, increasing the number of sequenced SNPs will result in the availability of a higher starting number of informative sites, to overcome the significant mean reduction of 70% after filtering.

Another improvement reguards library preparation. Analysis results have shown high variability among allelic balances in cffDNA and differences in sequencing depth in several parts of the sequenced region. This is probably affected by PCR amplification in genomics and cffDNAs, with different amplification performance for some regions or even alleles.

To overcome this bias, a Target Enrichment system will be used with specific probes, ensuring a high specificity and equal representation of DNA regions.

The laboratory workflow improvements will be coupled with several differences in analysis strategy.

The variant calling step will be performed with GATK (Genome Analysis ToolKit - Broad Institute) instead SAMtools. Recent studies have shown a significant increase in genotype determination accuracy that supports its worldwide inclusion in NGS data analysis pipelines (Laurie S, 2016).

The major limit of the pipeline is the dependence of maternally transmitted haplotype determination from paternal ones.

This was adopted as a strategy to overcome the limit represented by the very low number of maternal-only SNPs which, in a lot of samples, were not enough for haplotype determination.

Furthermore, high variability in terms of coverage among reads, due to different amplification efficiency, suggested the filtering out of maternal SNPs with associated allelic imbalance above 54%, which was theoretically expected for a fetal fraction of 8%. This value was chosen because it provided the highest detection rate. Higher allelic imbalance values, sometimes more than what was expected from fetal fraction, were supposed to be amplification/sequencing artifacts.

Major improvements will be firstly the maternal transmitted haplotype determination "unlocking" from paternal ones and the fetal fraction-related allelic imbalance cutoffs setting.

Secondly, pipeline efficiency will be tested on simulated data with different combinations of fetal fraction and informative SNPs number, to identify detection limits and best performance values.

With the availability of a large reference panel of haplotypes, as that produced by the Haplotype Reference Consortium, it will be possible to overcome the need for a population-specific reference panel to construct parental haplotypes, allowing for the analysis of individuals from different populations.

Finally, the new version of the pipeline will be fully written in a different programming language. We will move from Python scripts to R which is globally used among bioinformaticians, principally because it is intuitive, allows a step-by-step “visible” analysis. In this way an analysis package will be constructed and publically released.

Once described planned improvements are performed, the whole platform, laboratory and bioinformatic analysis, needs to be validated on a large set of samples.

For this reason a first pilot stage is required in which the analysis will be parallelly performed with traditional invasive procedures to confirm the results.

If the method reaches high sensitivity and specificity, with a detection rate compatible with diagnostic standards, it could then represent a valid alternative to invasive tests and routinary used for non invasive prenatal diagnosis of β-thalassemia in all couples at risk.

REFERENCES

Alberry M, M. D.-F. (2007). Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. *Prenatal Diagnosis*, 415-8.

Ariga H, O. H. (2001). Kinetics of fetal cellular and cellfree DNA in the maternal circulation during and after pregnancy: implications for noninvasive prenatal diagnosis. *Transfusion, 41*, 1524-30.

Barrett AN, M. T. (2012). Digital PCR Analysis of Maternal Plasma for Noninvasive Detection of Sickle Cell Anemia. *Clinical Chemistry*, 1026-32.

Benachi A, S. J. (2003). Fetal DNA in maternal serum: Does it persist after pregnancy? *Human Genetics, 113*, 76-79.

Bentley DR, B. S. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 465:53-59.

Boon EM, F. B. (2013). Benefits and limitations of whole genome versus targeted approaches for noninvasive prenatal testing for fetal aneuploidies. *Prenatal Diagnosis*, 33:563-568.

Bustamante-Aragones A, d. A.-T.-R.-L. (2012). Non-invasive prenatal diagnosis of single-gene disorders from maternal blood. *Gene*, 144-149.

Bustamante-Aragones A, e. a. (2008). Foetal sex determination in maternal blood from the seventh week of gestation and its role. *Haemophilia*, 593-598.

Cock PJA, F. C. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38:1767-1771.

Costa JM, B. A. (2002). New strategy for prenatal diagnosis of X-linked disorders. *N. Engl. J Med*, 1502.

Danecek P, A. A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27:2156-2158.

Delaneau O, H. B.-F. (2013). Haplotype estimation using sequence reads. *American Journal of Human Genetics*, 93:787-696.

Delaneau O, Marchini J, The 1000 Genome Project Consortium. (2014). Integrating sequence and array data to create an improved 1000 Genomes Projects haplotype reference panel. *Nature communications*, 5:3934.

Delaneau O, Z. J. (2013). Improved whole chromosome phasing for disease and population genetic studies. *Nature Methods* , 10: 5–6.

Dovc-Drnovšek T, K. P. (2013). Reliable Determination of Fetal RhD Status by RHD Genotyping from Maternal Plasma. *Transfusion Medicine and Hemotherapy*, 37-43.

Dressman D, Y. H. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *PNAS USA*, 100:8817-8822.

Drive5. (n.d.). *Drive5 - Bioinformatics Software and Services*. Retrieved from https://www.drive5.com/usearch/manual/fastq_files.html

Fan HC, G. W.-S. (2012). Noninvasive Prenatal Measurement of the Fetal Genome. *Nature*, 487:320-324.

Fedurco M, R. A. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*, 34:e22.

Finning K, M. P. (2004). A clinical service in the UK to predict fetal Rh (rhesus) D blood group using free fetal DNA in maternal plasma. *Ann. N.Y. Acad. Sci.*, 119-23.

Finning KM, C. L. (2008). Non-invasive fetal sex determination: impact on clinical practice. *Semin. Fetal Neonatal Med, 13*(2), 69-75.

Flori E, D. B. (2004). Circulating cell-free fetal DNA in maternal serum appears to originate from cyto- and syncytio-trophoblastic cells. Case report. *Human Reprod*, 723-4.

Gahan, P. (2013). Circulating nucleic acids in plasma and serum: applications in diagnostic techniques. *International Journal of Women's Health*, 177-186.

Gil MM, A. V. (2017). Analysis of cell-free DNA in maternal blood in screening for aneuploidies: updated meta-analysis. *Ultrasound Obstet Gynecol*, 50:302-314.

Goodwin S, M. J. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17:333-351.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860-921.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931-945.

Kitzman JO, S. M. (2012). Noninvasive whole-genome sequencing of a human fetus. *Science Translational Medicine*, 4:137ra76.

Laurie S, F.-C. M. (2016). From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human Mutation*, 37:1263-1271.

Lench N, B. A. (2013). The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made. *Prenatal Diagnosis*, 555-562.

Levy S, S. G. (2007). The Diploid Genome Sequence of an Individual Human. *PLoS Biology*, 5:e254.

Li H, H. B. (2009). The Sequence alignment/map (SAM) format and. *Bioinformatics*, 25:2078–2079.

Li H., D. R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25:1754–1760.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27:2987-2993.

Liao GJW, L. F. (2011). Target Massively Parallel Sequencing of Maternal Plasma DNA Permits Efficient and Unbiased Detection of Fetal Alleles. *Clinical Chemistry*, 92-101.

Lim JH, K. M. (2011). Non-invasive prenatal detection of achondroplasia using circulating fetal DNA in maternal plasma. *J Assist Reprod Genet*, 167-72.

Lo YMD, C. K. (2010, Dic 8). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science Traslational Medicine, 2*(61), 61ra91.

Lo YMD, C. N. (1997). Presence of fetal DNA in maternal plasma and serum. *Lancet*, 485-87.

Lo YMD, C. R. (2011). Plasma nucleic acid analysis by massively parallel sequencing: pathological insights and diagnostic implications. *Journal of Pathology*, 318-323.

Lo YMD, H. N. (1998). Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *N. Engl. J Med*, 1734-38.

Lo YMD, L. T. (1999, Feb). Quantitative abnormalities of fetal DNA in maternal serum in preeclampsia. *Clin Chem, 45*(2), 184-8.

Lo YMD, T. M. (1998). Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *American Journal of Human Genetics*, 768-75.

Lo YMD, Z. J. (1999). Rapid clearance of fetal DNA from maternal plasma. *American Journal of Human Genetics*, 218-24.

Lo, Y. (2013). Non-invasive prenatal testing using massively parallel sequencing of maternal plasma DNA: from molecular karyotyping to fetal whole-genome sequencing. *Reproductive BioMedicine Online*, 593-598.

Lun FM, C. R. (2008). Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. *Clin Chem*, 1664-72.

Lun FMF, T. N. (2008, Dic 16). Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. *Proc. Natl. Acad. Sci. USA, 105*(50), 19920-25.

Ma D, G. H. (2014). Haplotype-based approach for noninvasive prenatal diagnosis of congenital adrenal hyperplasia by maternal plasma DNA sequencing. *Gene*, 544:252–258.

Mackie FL, H. K. (2017). The accuracy of cell-free fetal DNA-based non-invasive prenatal testing in singleton pregnancies: a systematic review and bivariate meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology*, 124:32-46.

Mardis, E. R. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols*, 12:213-218.

Margulies M, e. a. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376-380.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51-56.

Nielsen R, P. J. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews Genetics*, 12:443-451.

Python Software Foundation. (2010). *Python Language Reference, version 2.7*. Retrieved from http://www.python.org

Quail MA, S. M. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*(341).

Rijnders RJ, v. d. (2001). Fetal sex determination from maternal plasma in pregnancies at risk for congenital adrenal hyperplasia. *Obstet Gynecol, 98*, 374-78.

Robinson JT, T. H. (2011). Integrative Genomics Viewer. *Nature biotechnology*, 29:24-26.

Rothberg JM, H. W. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475:348-352.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall/CRC.

Saba L, M. M. (2017). Non-invasive prenatal diagnosis of beta-thalassemia by semiconductor sequencing: a feasibility study in the sardinian population. *European Journal of Human Genetics*, 25:600-607.

Saito H, S. A. (2000). Prenatal DNA diagnosis of a single-gene disorder from maternal plasma. *Lancet*, 356:1170.

Scheffer PG, v. d.-C. (2010). Reliability of fetal sex determination using maternal plasma. *Obstet Gynecol*, 117-26.

Schmidt B, H. A. (2017). Next-generation sequencing: big data meets high performance computing. *Drug Discovery Today*, 22:712-717.

Shendure J, e. a. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728-1732.

Shendure J, e. a. (2011). Overview of DNA Sequencing Strategies. *Current Protocols in Molecular Biology*, 96:7.1.1-7.1.23.

Shendure J, J. H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26:1135-1145.

Sherry ST, W. M. (2001). dbSNP: the NCBI database of genetic variation . *Nucleic Acids Res.*, Jan 1;29(1):308-11.

Stroun M, A. P. (1987). Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol*, 318-22.

Sudmant PH, R. T. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75-81.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526:68-74.

The Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48:1279–1283.

The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526:82-90.

Thorvaldsdóttir H, R. J. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14:178–192.

Tsui NBY, K. R. (2011, Mar 31). Noninvasive prenatal diagnosis of hemophilia by microfluidics digital PCR analysis of maternal plasma DNA. *Blood, 117*(13), 3684-91.

Van der Walt S, C. S. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13:22-30.

Venter JC, e. a. (2001). The sequence of the human genome. *Science*, 291:1304-1351.

Veritas Genetics. (2016). *Veritas Genetics Launches $999 Whole Genome And Sets New Standard For Genetic Testing*. Retrieved from Veritas Genetics: https://www.veritasgenetics.com/content/veritasgenetics-launches-999-whole-genome-and-sets-new-standard-genetic-testing

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory*, vol. IT–13:260–269.

Vogelstein B, K. K. (1999). Digital PCR. *Proc. Natl, Acad. Sci. USA*, 9236-41.

Wald, Á. (1947). *Sequential Analysis.* New York: John Wiley and Sons.

Wetterstrand, K. A. (2016). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Retrieved from www.genome.gov/sequencingcostsdata

Wheeler DA, S. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872-876.

Zafari M, K. M. (2016). Non-invasive prenatal diagnosis of β-thalassemia by detection of the cell-free fetal DNA in maternal circulation: a systematic review and meta-analysis. *Annals of Hematology*, 95:1341-50.

APPENDIX A - SUPPLEMENTARY TABLES

Supplementary Table 1. Sequencing reads and depth

Sample ID	*Number of reads*	*Sequencing depth*
PGM001	117936	296.25
PGM002	118591	334.77
PGM004	364679	5486.51
PGM005	225021	597.55
PGM006	305677	859.12
PGM008	456172	3811.08
PGM009	289122	563.24
PGM010	218641	334.21
PGM012	393910	6613.37
PGM013	101405	295.6
PGM014	99430	241.46
PGM016	0	0
PGM017	74030	194.49
PGM018	247865	379.01
PGM020	265142	3719.67
PGM021	128569	320.12
PGM022	125176	314.54
PGM024	523368	7736.38
PGM025	126846	363.15
PGM026	24205	59.71
PGM028	652616	11391.1
PGM029	121837	345.08
PGM030	115479	339.95
PGM032	467013	6903.12
PGM033	123127	299.76
PGM034	115679	276.36
PGM036	0	0
PGM037	139198	422.49
PGM038	147592	452.36
PGM040	482872	8242.94
PGM041	165867	462.77
PGM042	70140	177.28
PGM044	546005	7937.48
PGM045	14910	28.42
PGM046	88029	255.93
PGM048	270095	4695.82
PGM049	89333	208.64
PGM050	65979	124.34
PGM052	606611	7676.62
PGM053	52354	69.11
PGM054	101694	160.26
PGM056	444307	4436.49
PGM057	240605	654.42
PGM058	75689	189.75
PGM060	482717	4973.15
PGM061	389838	654.9
PGM062	104239	319.95
PGM064	156663	2295.52
PGM065	290177	638.39
PGM066	147271	406.12
PGM068	423776	5600.85
PGM069	152341	420.33
PGM070	124661	341.67
PGM072	324872	5312.82
PGM073	111054	315.1
PGM074	68139	190.39
PGM076	471501	7809.06
PGM077	79240	213.21
PGM078	82651	173.92
PGM080	548014	26819
PGM081	389478	1074.28
PGM082	177968	528.85
PGM084	317274	4663.62
PGM085	146266	372.41
PGM086	158326	406.96
PGM088	478115	6878.28
PGM089	25808	60.32
PGM090	148328	393.16
PGM092	630248	10273.9
PGM093	516566	950.27
PGM094	378855	508.97
PGM096	757657	9427.1
PGM097	142819	357.01
PGM098	130643	366.94
PGM100	482397	7569.79
PGM101	127906	310.91
PGM102	165598	498.74
PGM104	366492	6522.64
PGM105	84457	147.63
PGM106	199480	582.65
PGM108	245714	4795.84
PGM109	60197	99.81
PGM110	159364	385.31
PGM112	432029	6915.02
PGM113	642769	1327.51
PGM114	327350	519.22
PGM116	599465	13169.7
PGM117	153039	369.31
PGM118	123556	307.39
PGM120	533003	8666.96
PGM121	116078	150.85
PGM122	65050	148.27
PGM124	497591	7696.39
PGM125	149734	351.59
PGM126	145930	353.14
PGM128	621245	12398.4
PGM129	70230	83.92
PGM130	166983	147.13
PGM132	384253	7133.64
PGM133	46311	114.73
PGM134	68317	162.01
PGM136	465995	7156.9
PGM137	163068	487.05
PGM138	103536	287.61
PGM140	337708	5459.61
PGM141	133810	352.45
PGM142	156219	253.36
PGM144	276154	5642.42
PGM145	59271	141.21
PGM146	71527	164.11
PGM148	515799	4592.3
PGM149	148091	419.97
PGM150	468072	1036.95
PGM152	171388	2461.33
PGM153	90335	197.65
PGM154	89593	210.35
PGM156	508264	6321.65

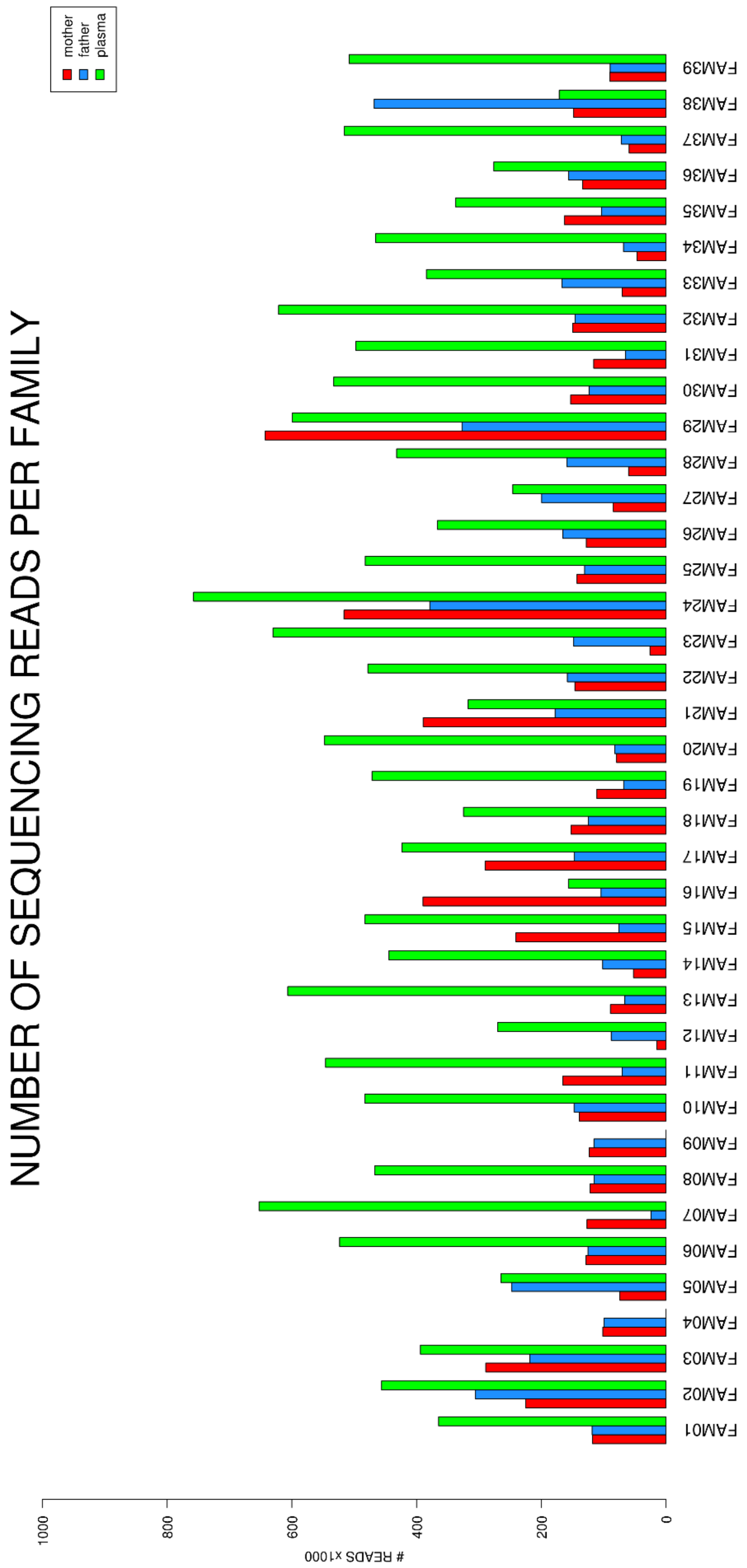
Supplementary Table 2. Informative SNPs Categories per family

Family ID	*Heterozygous-Only*	*Maternal-Only*	*Paternal-Only*	*Homozygous-Only*
FAM01	69	13	66	0
FAM02	66	28	46	0
FAM03	73	63	7	0
FAM04	73	67	11	0
FAM05	111	12	10	0
FAM06	107	5	14	0
FAM07	46	32	45	0
FAM08	92	31	30	0
FAM09	62	28	64	0
FAM10	32	51	24	0
FAM11	69	26	57	0
FAM12	57	19	23	0
FAM13	76	50	11	0
FAM14	26	68	49	1
FAM15	42	61	56	0
FAM16	17	81	51	20
FAM17	49	66	65	0
FAM18	3	74	52	0
FAM19	25	50	95	0
FAM20	9	13	57	1
FAM21	22	67	53	4
FAM22	50	6	2	0
FAM23	88	32	16	0
FAM24	84	24	53	1
FAM25	95	26	32	0
FAM26	64	9	38	0
FAM27	6	10	112	0
FAM28	67	17	64	0
FAM29	45	64	20	0
FAM30	74	7	46	0
FAM31	90	54	32	1
FAM32	66	8	11	0
FAM33	28	76	27	1
FAM34	59	44	52	1
FAM35	4	54	7	59
FAM36	10	24	45	2
FAM37	18	63	50	43
FAM38	72	51	7	0
FAM39	87	31	50	0
mean	55	39	40	3

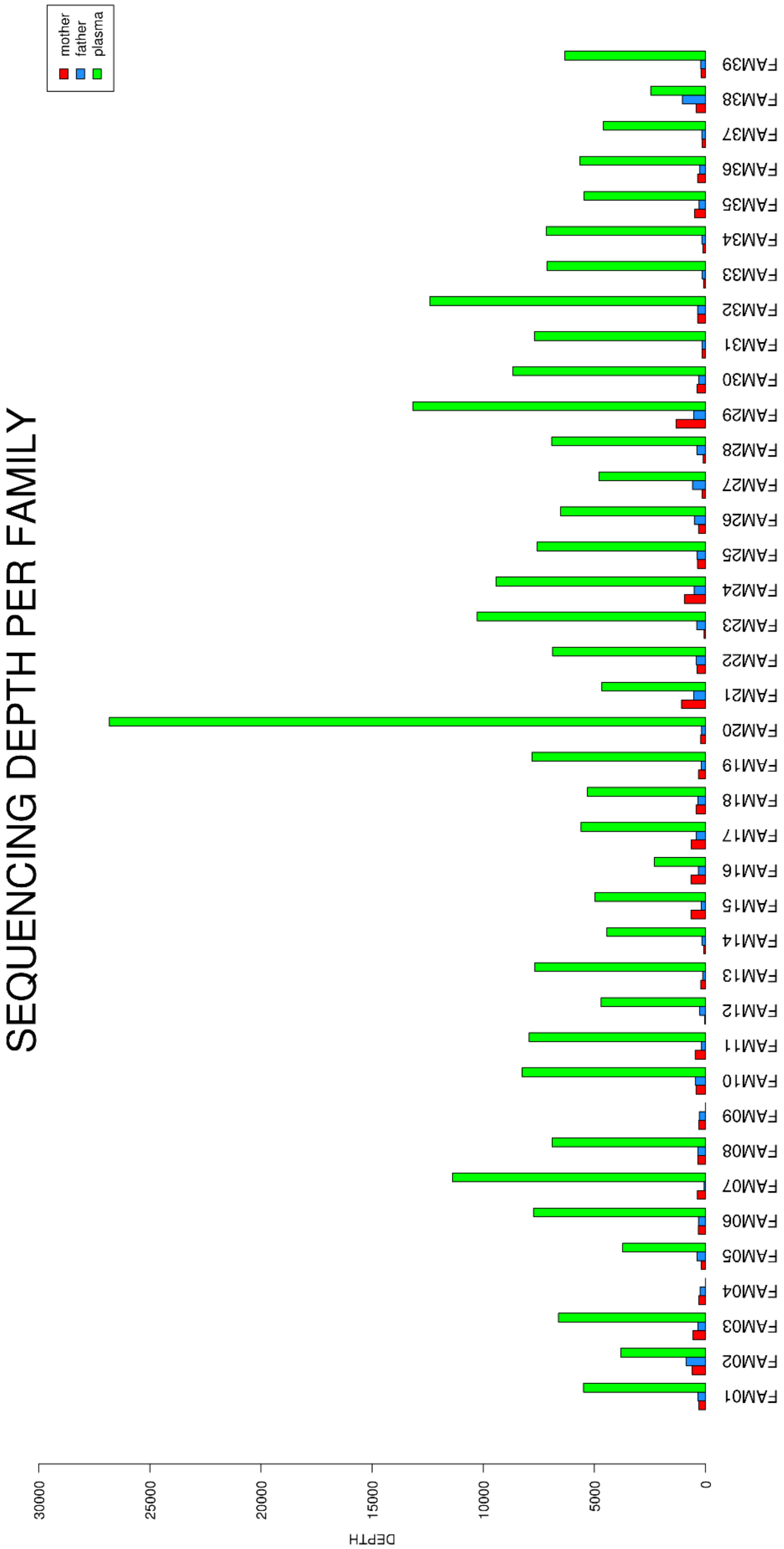
Supplementary Table 3. SNPs used for parental haplotype prediction and cffDNA predicted HBB genotypes

Family ID	*Sequenced SNPs*	*Paternal SNPs*	*Maternal SNPs*	*Fetal Fraction (%)*	*Predicted Genotype*	*CVS Genotype*	*Analysis Result*
FAM01	69	32	12	4.6	0/0	0/0	Correct
FAM02	58	11	16	7.4	0/1	0/1	Correct
FAM03	60	1	25	4	0/1	0/1	Correct
FAM04	-	-	-	-	-	0/0	-
FAM05	57	3	15	3.7	1/1	1/1	Correct
FAM06	58	5	26	5.6	0/0	0/0	Correct
FAM07	58	14	20	8.6	1/1	1/1	Correct
FAM08	58	-	-	-	-	1/1	-
FAM09	-	-	-	-	-	1/1	-
FAM10	55	7	27	4.2	0/0	0/0	Correct
FAM11	66	30	15	4.9	0/0	0/0	Correct
FAM12	44	-	-	-	-	0/1	-
FAM13	56	-	-	-	-	1/1	-
FAM14	76	26	11	7	0/1	0/1	Correct
FAM15	75	15	19	5.5	0/1	0/1	Correct
FAM16	61	10	14	4.9	0/0	0/1	Incorrect
FAM17	80	17	21	9.7	0/1	0/1	Correct
FAM18	58	10	17	7.4	0/0	0/0	Correct
FAM19	64	34	14	4.6	0/0	0/1	Incorrect
FAM20	13	11	1	5.8	0/1	0/1	Correct
FAM21	66	26	10	7.1	0/0	0/0	Correct
FAM22	23	-	-	-	-	0/1	-
FAM23	57	-	-	-	-	0/1	-
FAM24	69	15	23	7.9	0/1	0/1	Correct
FAM25	67	7	20	10.9	0/1	0/0	Incorrect
FAM26	50	7	26	5.9	1/1	0/1	Incorrect
FAM27	44	29	3	8.3	1/1	1/1	Correct
FAM28	61	13	17	11.4	0/1	1/1	Incorrect
FAM29	45	3	16	12.5	0/0	0/0	Correct
FAM30	56	12	26	7.8	0/1	0/1	Correct
FAM31	67	4	33	12.6	1/1	1/1	Correct
FAM32	47	5	21	8.7	0/0	0/0	Correct
FAM33	56	7	18	4.6	0/1	0/0	Incorrect
FAM34	58	18	8	7.4	0/0	0/0	Correct
FAM35	56	2	8	4.6	0/1	0/1	Correct
FAM36	29	-	-	-	-	0/0	-
FAM37	81	9	12	5.6	1/1	1/1	Correct
FAM38	54	-	-	-	-	0/0	-
FAM39	86	12	25	5.8	0/0	0/0	Correct

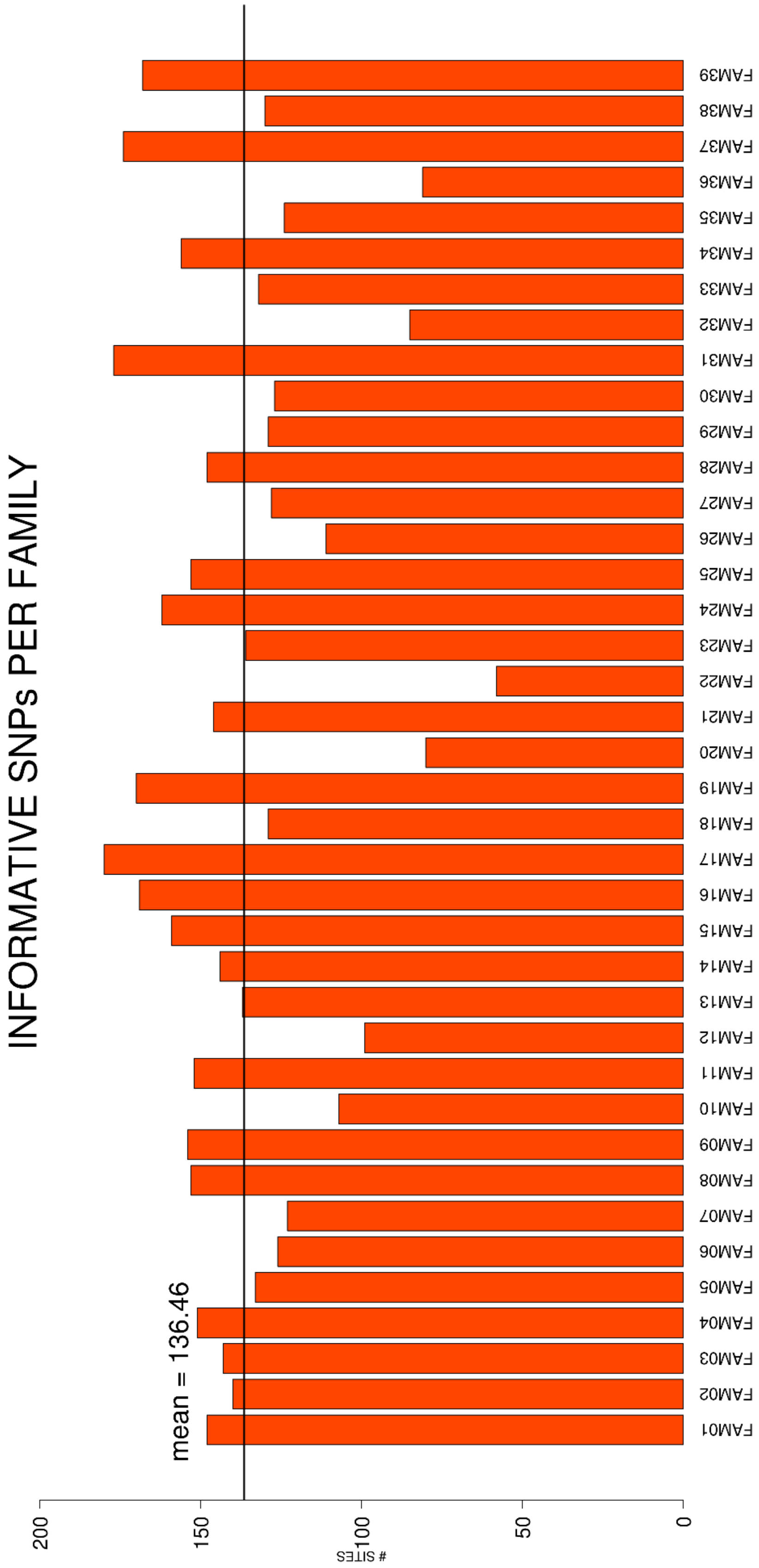
APPENDIX B - SUPPLEMENTARY FIGURES



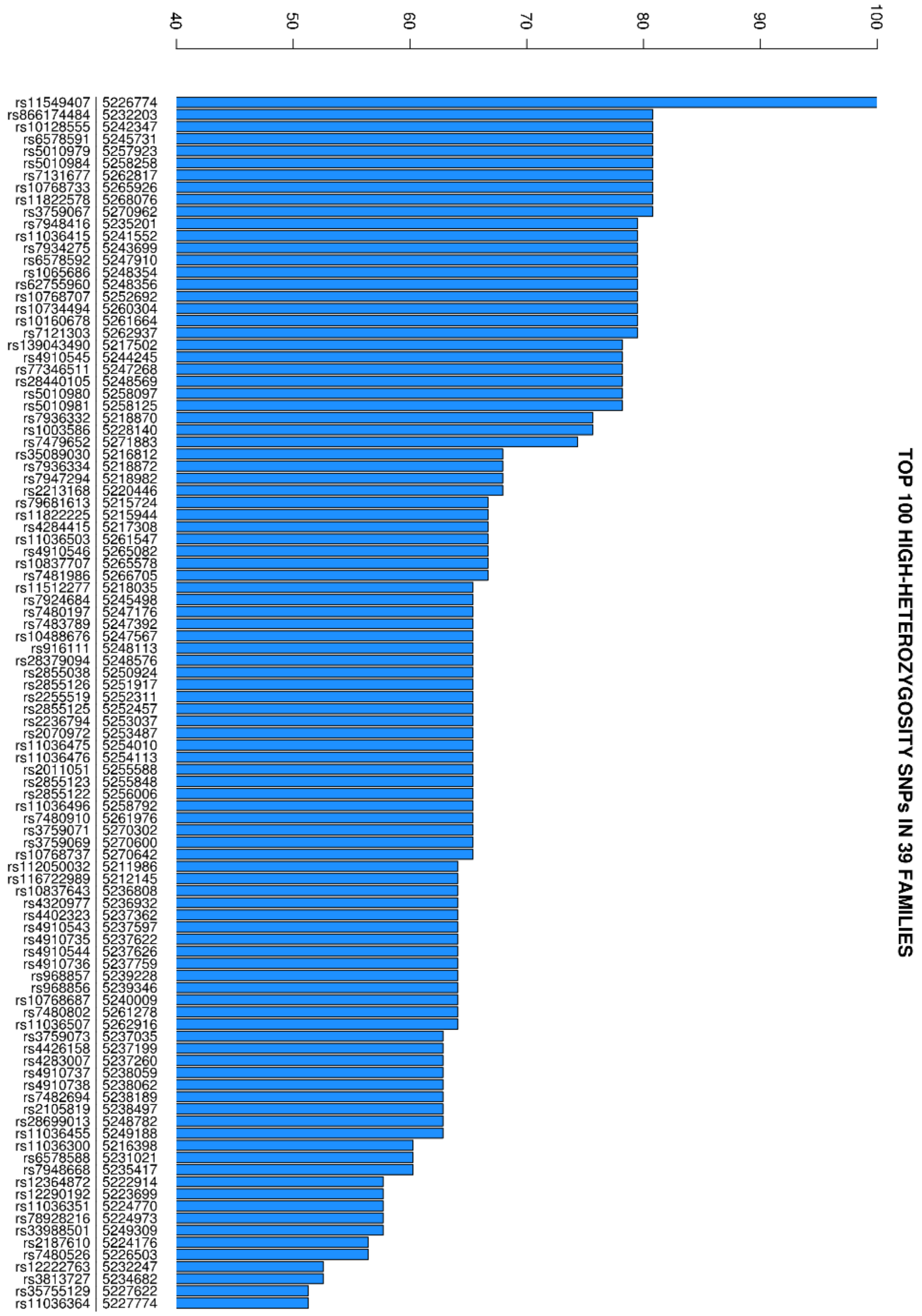
Supplementary Figure 1. **Number of sequencing reads per family**. For each family are grouped maternal (red), paternal (blue) and cffDNA (green) sample reads (x1000).



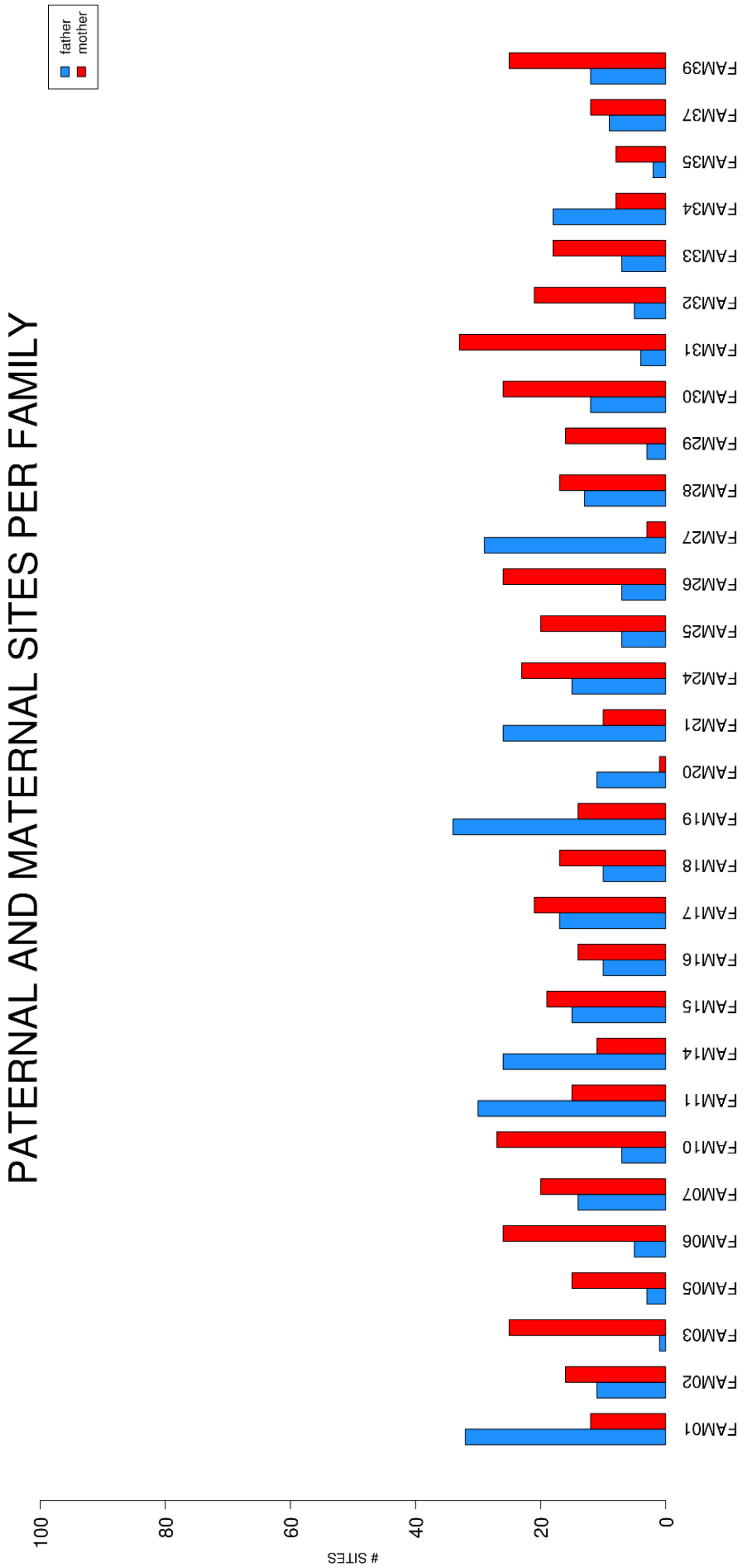
Supplementary Figure 2. **Sequencing depth per family**. For each family are grouped maternal (red), paternal (blue) and cffDNA (green) sample sequencing depth.



Supplementary Figure 3. **Number of informative SNPs per family**. For each family is shown the number of informative SNPs present in the sequenced region of parental genomes.



Supplementary Figure 4. **Top 100 high-heterozygosity SNPs in 39 families.** Frequency of heterozygosity of 100 SNPs in 39 Sardinian couples. SNPs are sorted from higher to lower heterozygosity frequency. For each site are shown both genomic coordinates (GRCh38) and dbSNP IDs (dbSNP150).



Supplementary Figure 5. **Number of paternal and maternal sites per family**. For each family are grouped the number of maternal (red) and paternal (blue) sites used for the fetal haplotypes prediction.

ACKNOWLEDGEMENTS

I would like to thank all the people who helped and supported me in these years:

- **Prof.ssa Maria Cristina Rosatelli**, my supervisor of the thesis, for accepting me in her laboratory, for her professional guidance and for believing in me in these years. Especially for firmly encouraging my stage abroad.
- The **NIPD group** of the Genetics and Genomics Lab, for the collaboration, the support and the patience in these years.
- **Dr. Holger Heyn**, my supervisor at CNAG-CRG during my stage abroad. For having me welcomed in his team and for the amazing research experience.
- The **SCG team** at CNAG-CRG for helping me in improving my bioinformatics skills.
- The **12th community of S. Joaquina** in Barcelona, for the friendship and the support for me and my family when we were alone.
- **My Parents**, for always supporting and encouraging me, for all the sacrifices they did in their life to allow my studies.
- My sons **Daniele** and **Nicola**, for giving me the strength to move on every day with their smiles, even in difficult situations.
- Finally my wife, **Miky**, to be part of my life, to be The One, to take every day the worst part of me and despite everything being with me. For come with me in Spain with a 5 month baby, pausing your studies and insisting for me to finish the PhD when I wanted to give up. This title is mine as yours. Thank you very much. I love you.