Development of tools for the analysis of cell-free DNA in human blood

Application in the noninvasive prenatal testing field

Anna Montaner Domènech

TESI DOCTORAL UPF / 2019

DIRECTORS DE LA TESI

Dr. Jairo Rodríguez Lumbiarres (qGenomics)
Dr. Juan R. González Ruiz (ISGlobal)

TUTOR DE LA TESI

Dr. Luis A. Pérez Jurado (Departament de Ciències Experimentals i de la Salut)

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

Development of tools for the analysis of cell-free DNA in human blood.

Application in the noninvasive prenatal testing field.

Anna Montaner Domènech

PhD Thesis, 2015 - 2019

A la meva mare

Agraïments

Hi ha una pila de persones a qui haig d'agrair que hagin estat al meu costat, o més o menys a prop, o, si més no, no gaire lluny de mi, al llarg d'aquests anys.

En primer lloc, els meus directors de tesi, Jairo Rodríguez i Juan Ramón González. Al Juan Ramón, pel suport bioinformàtic i estadístic, llenguatges poc còmodes (exasperants) al principi que a poc a poc s'han convertit en imprescindibles. Al Jairo, per la mirada exigent i profunda de les qüestions científiques, pel seguiment acadèmic i humà i sobretot, per la paciència.

Tinc la sort de formar part de moltes tribus.

Qgenomics, la que no para de créixer i sempre necessita més espai, la que m'ha proporcionat un entorn d'aprenentatge i treball privilegiat i un munt de persones maques que corren amunt i avall amb molta (però molta) feina entre mans.

La colla d'amics que mai vaig escollir, ni em van escollir, ni ens vam escollir, per ser els de (gairebé) tota la vida. Als del Santa Anna, és una sort anant-me fent gran amb vosaltres.

La gent de circ. Cordistes, telistes, trapezistes, malabaristes, pallassos, elefants i tigres. Les setmanes són grises quan no puc compartir idees, figures, seqüències, números massa curts, massa llargs, músiques i vestuaris amb vosaltres.

La meva petita però ben avinguda família, les històries d'Itàlia del meu avi, l'alegria de la meva àvia, l'estima, suport, i preocupacions (inacabables) del meu pare, la perseverança de la meva mare. Gràcies a tots ells sóc aquí.

El Gerard, gironí de tota la vida, tallador de pernil professional, optimista patològic, usuari de cabines telefòniques, negociador empedreït i enamorat de Messi, sempre al meu costat i per sort, de paciència considerable.

La nombrosa família gironina, la tribu tranquil·la a qui no agraden les cues, amants de les plantes, dels vegetals, de les sobretaules. Sempre amables i acollidors amb una ~~xava~~ barcelonina de ritme accelerat.

A totes, moltes gràcies per ser-hi!

Abstract

The discovery of cell-free DNA (cfDNA) in blood and the advent of massive parallel sequencing technologies has revolutionized the prenatal screening field by providing a better risk assessment of fetal chromosomal alterations than traditional first trimester screening alone and, as a consequence, potentially reducing the number of women unnecessarily undergoing confirmatory invasive diagnostic tests. The so-called noninvasive prenatal testing (NIPT) field is nowadays being exploited by worldwide companies racing to offer cfDNA tests covering an increasing number of fetal conditions, not without ethical concerns. This Thesis explores the use of [redacted] that can be applied in the NIPT field but also in other relevant clinical settings like cancer.

Keywords: cell-free DNA, noninvasive prenatal testing, [redacted], massive parallel sequencing, aneuploidy

Resum

El descobriment d'ADN lliure circulant (ADNcir) en sang i el desenvolupament de tecnologies de seqüenciació massiva han revolucionat el camp del cribratge prenatal mitjançant una millor avaluació del risc d'alteracions cromosòmiques fetals respecte el cribratge tradicional del primer trimestre. Aquest fet, doncs, ha permès una potencial reducció del nombre de dones sotmeses a proves diagnòstiques invasives confirmatòries. L'anomenat camp de l'anàlisi prenatal no invasiu (TPNI) està sent explotat avui en dia per companyies d'arreu del món en una cursa per a oferir proves basades en l'ADNcir que cobreixen un nombre creixent de condicions fetals, no sense preocupacions ètiques. Aquesta Tesi explora l'ús de [redacted] aplicable no només en el camp de TPNI sinó també en d'altres entorns clínics rellevants com són el càncer.

Paraules clau: ADN lliure circulant, anàlisi no invasiu prenatal, [redacted], seqüenciació massiva, aneuploïdia .

Preface

This Thesis has been developed in the context of the Industrial PhD program, an initiative supported by the Generalitat de Catalunya initiated by the end of 2012 that aims at strengthening collaboration and promoting knowledge and technology transfer between university or research centers and industry. Such collaboration usually results in the concretion of the generated knowledge into a technology or product that can provide both economic and social benefits. From the 17 projects started in 2012, a total of 5 have generated results susceptible to intellectual protection through either patents or utility models.

In this case, the collaboration took place between the Barcelona Institute for Global Health (ISGlobal) and qGenomics. qGenomics was founded in 2008 as a spin-off of the Center for Genomic Regulation (CRG) and the Pompeu Fabra University (UPF) and currently has around 50 employees.

Research in this industrial context is very much driven by its utility in improving the efficiency of internal workflows or by its potential translation into a cost-effective commercial assay. In this context, the interest of cell-free DNA (cfDNA) from a commercial point of view is evident: the advent of next-generation sequencing (NGS) technologies around 10 years ago accelerated cfDNA research and has allowed the development of the so-called noninvasive (or minimally invasive) techniques. Several fields of potential application appeared, being the noninvasive prenatal testing (NIPT) the most successful of them. Nowadays, NIPT commercial tests are marketed worldwide covering an increasingly number of fetal conditions.

However, the interest for cfDNA research is not restricted to its commercial value: the biological features of cfDNA require the development of a set of tools to be able to work with low concentrated and partially degraded samples. In the context of qGenomics, this has direct implications in the daily routine, as it is not infrequent to receive low quality samples that pose important challenges to standard sample processing workflows.

The partnership between qGenomics and Hospital Universitari Dexeus allowed to obtain a first reduced group of samples to validate, in the context of a pilot study, the performance of our developed tools for detecting the most relevant chromosomal alterations and evaluating basic parameters typically evaluated in noninvasive prenatal tests. A second phase of the study has recently started that will allow, by collecting a higher number of samples, to explore more refined aspects of both cfDNA biology and our own methodology.

Index of contents

[redacted]

[redacted]

List of Figures

Page

List of Tables

Acronyms and abbreviations

ACMG	American College of Medical Genetics and Genomics
AF	allele frequency
AFP	α-fetoprotein
A414	absorbance at $\lambda = 414$ nm
BAM	binary alignment map file
BCL	binary base call
BCT	blood collection tube
BMI	body mass index
bp	base-pair
β-hCG	β-human chorionic gonadotropin
CAH	congenital adrenal hyperplasia
cfDNA	cell-free DNA
cffDNA	cell-free fetal-derived DNA
chr	chromosome
Cq	cycle quantification value
CRG	Center for Genomic Regulation
ctDNA	tumor-derived cell-free DNA
CVN	copy-number variant
CVS	chorionic villi sampling
d	density
DANSR	Digital Analysis of Selected Regions
ddPCR	digital-droplet PCR
DNase	caspase-activated deoxyribonuclease
dNTP	deoxyribonucleotide triphosphate
DR	detection rate
dsDNA	double-stranded DNA
EB	elution buffer
EVT	extravillous cytotrophoblast
FDR	false discovery rate
FF	fetal fraction
FGA	α-fibrinogen
FGB	β-fibrinogen
FGG	γ-fibrinogen
[redacted]	[redacted]
FNR	false negative rate
FPR	false positive rate
[redacted]	[redacted]
GC	guanosine-cytosine
gDNA	genomic DNA
HBB	hemoglobin beta
HUD	Hospital Universitari Dexeus
ISGlobal	Barcelona Institute for Global Health
Kb	kilobase
[redacted]	[redacted]

MAPQ	mapping quality
Mb	megabase
[redacted]	[redacted]
miRNA	microRNA
MNase	micrococcal endonuclease
MPS	massive parallel sequencing
NGS	next-generation sequencing
NT	nuchal translucency
o/n	overnight
OOR	out-of-range
PAPP-A	pregnancy-associated plasma protein A
PCR	polymerase chain reaction
PE	paired-end
Pol	RNA polymerase
QF-PCR	quantitative fluorescent polymerase chain reaction
qPCR	quantitative real-time PCR
RBC	red blood cell
RhD	rhesus factor D
ROI	region of interest
RPK	reads per kilobase
RT	room temperature
SBS	sequencing-by-synthesis
SCA	sex chromosome aneuploidies
SE	single-end
[redacted]	[redacted]
SLE	systemic lupus erythematosus
SN	supernatant
SNP	single-nucleotide polymorphism
SRY	sex determining region Y
STR	short-tandem repeat amplification
TIRF	total internal reflection microscopy
Tm	melting temperature
TPM	transcript per kilobase million
TSS	transcription start site
T9	trisomy of chromosome 9
T13	trisomy of chromosome 13
T18	trisomy of chromosome 18
T21	trisomy of chromosome 21
uE3	unconjugated estriol
UPF	Pompeu Fabra University
UV-Vis	ultraviolet-visible
WBC	white blood cell
WES	whole-exome sequencing
WGS	whole-genome sequencing
z	z-score

1. Introduction

1.1 General aspects of screening programs

In the clinical setting, the early detection of certain pathological conditions before any symptoms become noticeable is highly desirable to allow an early preventive or diagnostic intervention that can contribute to a better management of that specific condition[1].

To this end, screening programs aiming at identifying affected individuals for a given condition out of a defined population of apparently healthy or asymptomatic individuals have been developed. In this context, if an individual is found to be within a high-risk group according to a specific screening test, he or she will be offered some preventive or confirmatory diagnostic intervention that can provide a definitive an accurate diagnosis[1].

Confirmatory interventions commonly include the invasive extraction of a biological sample of a potentially abnormal tissue or fluid for close examination, which carry a series of risks for the patient. In the context of cancer, for instance, image-based screening methodologies such as colonoscopy or mammography may indicate that an individual is at high risk of colon or breast cancer. If the imaging methodologies suggest that a tumoral mass may be present, a confirmatory tissue biopsy is performed, and a diagnosis may be confirmed. The obtention of a tissue biopsy is an invasive procedure that carries a significant risk of infection, bleeding and even death[2].

In the field of prenatal testing, the risk of potential fetal abnormalities is typically screened at 8 to 13,6 gestation weeks[3] using a combination of blood biomarkers, echographic signatures (nuchal translucency thickness, NT) and maternal age. When a pregnancy is deemed high risk, then fetal genetic material needs to be obtained through invasive methods such chorionic villus sampling (CVS) or amniocentesis. In this context, both CVS and amniocentesis have been associated to a risk of miscarriage of 0.22% and 0.11% respectively[4]. A historic, more in-depth vision of the screening strategies used in prenatal testing is provided in **Section 1.5.1**.

Because screening methods aim at identifying all potential affected individuals in a population, high sensitivity is needed, which is commonly achieved at the expense of specificity. This results in a proportion of false positive individuals that unnecessarily undergo confirmatory invasive procedures[5]. The need to improve the sensitivity and specificity of screening methodologies to minimize both the number of invasive procedures and their associated risks fueled the research of novel more accurate and minimally invasive methodologies, often called noninvasive methodologies.

The rapid expansion of next-generation sequencing (NGS) methodologies over the last decade has created many opportunities for the development of improved screening strategies. In the following paragraphs, the key aspects of an Illumina compatible NGS laboratory workflow are introduced, which is the sequencing technology used throughout this work.

1.2 Next-generation sequencing

The first-generation DNA sequencing techniques were published in 1977 by Fred Sanger and colleagues[6] after nearly a quarter of a century since the discovery of the DNA structure[7]. The original methodology, known as Sanger sequencing, consisted in a sequencing-by-synthesis (SBS) approach where the DNA sequence is determined one nucleotide at a time for a given population of DNA templates. Sanger sequencing allowed, between 1990 and 2004, the development of the Human Genome Project, a $2.7 billion international project aiming at sequencing the entire human genome[8,9].

However, Sanger sequencing presented several limitations in the extent of DNA that could be sequenced, as it was generally restricted to regions of interest (**Figure 1C**). Moreover, the scalability of the method was very limited. The need for faster, more efficient and scalable routine genomic sequencing led to the emergence, starting in

2005, of new sequencing methods, collectively known as NGS, massive parallel sequencing (MPS) or high-throughput sequencing[10].

NGS technologies greatly reduced the time and costs of DNA and RNA sequencing by massively increasing the sequence output compared to Sanger sequencing: while Sanger sequencing is only capable of producing one sequence per template reaction, in NGS, millions to billions of individual sequencing reactions are performed simultaneously[11].

The ability to parallel sequence large portions of the genome allowed the emergence of several genomic approaches that had been impossible before the advent of NGS. One of these approaches is whole-genome sequencing (WGS), which can be performed at a reasonable time and price and allows the study of both coding and non-coding regions (**Figure 1F**).

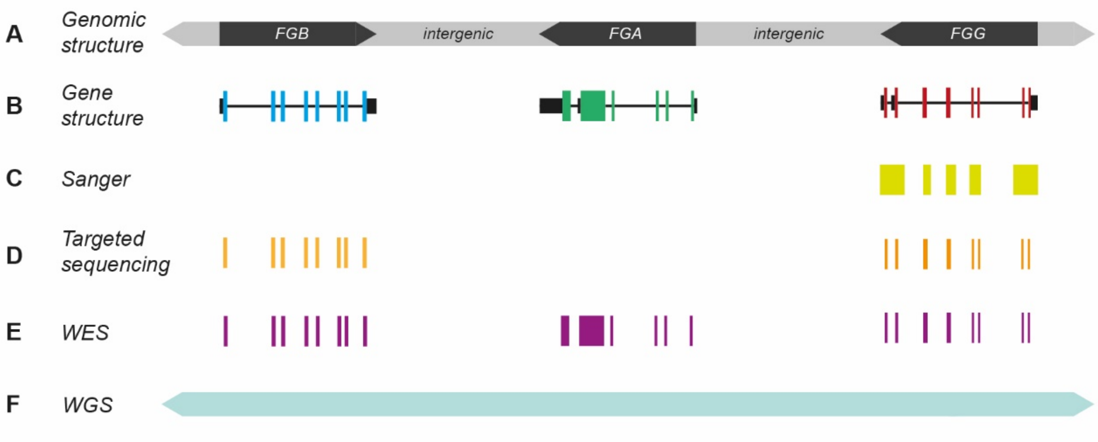


Figure 1. **The horizontal coverage targets of different NGS approaches compared to Sanger sequencing**. Adapted from Kumar et al.[13] **A**. The genomic structure of the fibrinogen gene cluster involves three genes (β-fibrinogen [FGB], α-fibrinogen [FGA] and γ-fibrinogen [FGG]) separated by intergenic regions, with FGB on the plus strand and FGA and FGG on the minus strand of chromosome 4. **B**. Structurally, the genes consist of untranslated regions (small black boxes), protein coding exons (colored boxes), and non-coding introns (horizontal line between exons). **C**. Sanger sequencing is usually restricted to regions within a gene of interest and can span up to approximately 800 bp. **D**. Targeted gene sequencing focuses on genes of particular interest. **E**. Whole-exome sequencing (WES) provides sequence data enriched for all protein coding exons of the entire genome, offering a cost-effective means of considering pathogenic variants throughout the coding regions of the genome. **F**. Finally, whole-genome sequencing (WGS) provides comprehensive sequencing data across the entire genome, including the protein coding exome, intronic regions between exons, and intergenic regions between genes.

Alternatively, targeted sequencing of regions within a single gene or multiple genes of interest can be performed (**Figure 1D**). If targeted gene sequencing is expanded to all protein-coding regions (i.e., exons) the approach is termed whole-exome sequencing (WES) (**Figure 1E**). WES is very useful in the clinical setting, as it is estimated that exonic sequences, which account for only around 2% of the entire genome, contain 85% of currently known disease-causing variants[12].

Among the sequencing platforms developed to date (e.g., Illumina, Roche 454, Ion Torrent, and PacBio), the Illumina system is the predominant platform used today in most NGS-related applications[14] and it is the methodology used in the context of this Thesis work.

1.2.1 The principles of Illumina-compatible library preparation and sequencing

Prior to sequencing, DNA samples are modified in a process known as library preparation (**Figure 2**). In the context of Illumina sequencing, library preparation consists in i) DNA fragmentation, ii) DNA end-repair to generate blunt, 5'-phosphorylated ends, iii) A-tailing of the 3' ends to facilitate ligation to Y-shaped sequencing adapters, iv) ligation of adapters and v) PCR amplification to enrich for library molecules having ligated adapters on both ends. Additionally, samples can be pooled together, and specific regions of the genome can be enriched to reduce sequencing costs. These main steps are further introduced in the following paragraphs.

a) DNA fragmentation

Illumina sequencing chemistry requires fragmentation of the DNA samples to be sequenced. For applications that use genomic DNA (gDNA) as input, samples can be typically fragmented to an average fragment size of 150-300 bp. This step is predominantly done either by physical methods (e.g., acoustic shearing and sonication) or random enzymatic digestion (e.g., non-specific endonuclease cocktails or transposase tagmentation reactions) (**Figure 2A**)[15].

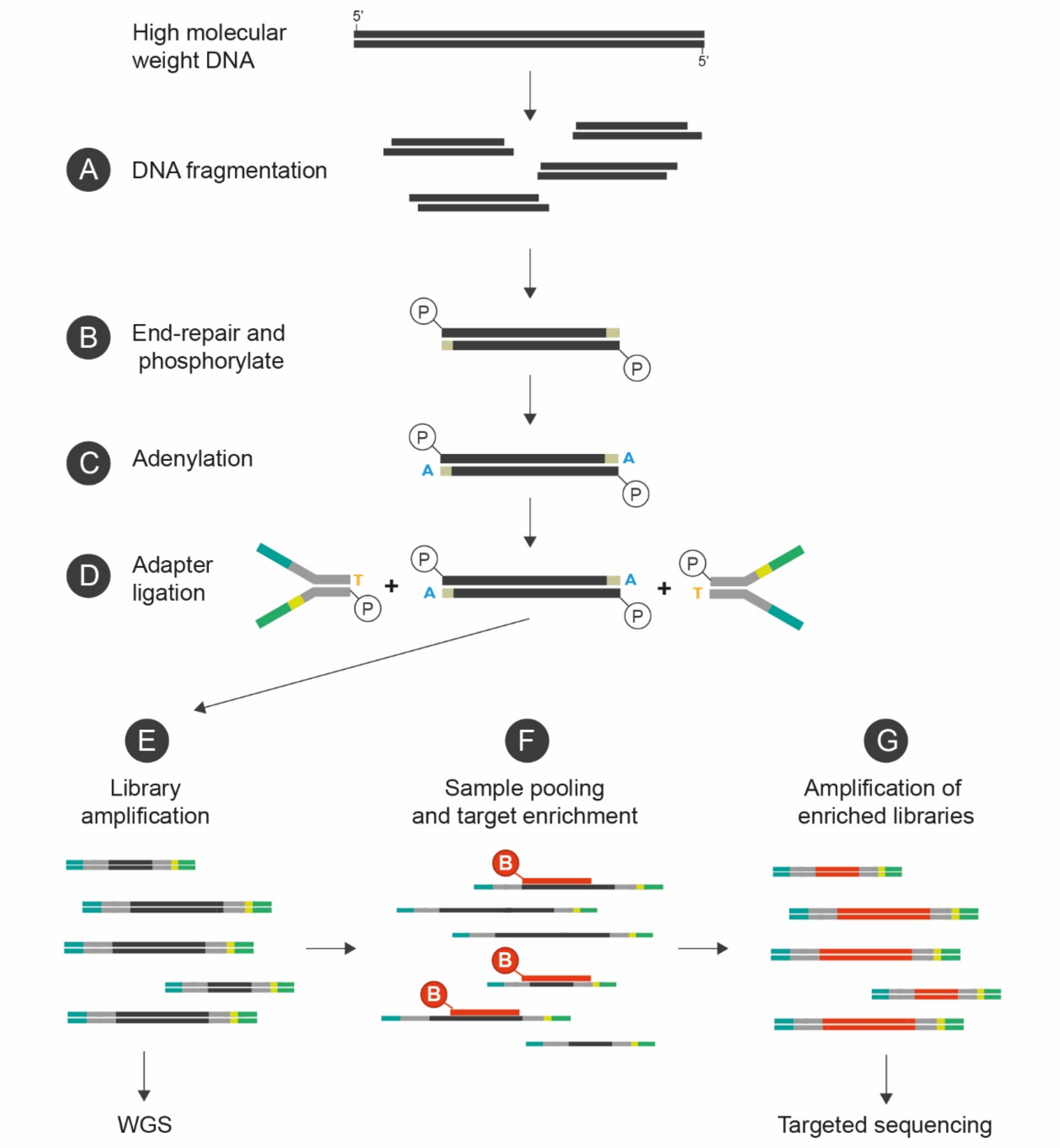

High molecular weight DNA
5'
5'
A DNA fragmentation
B End-repair and phosphorylate
P
C Adenylation
A
D Adapter ligation
T
+
E Library amplification
F Sample pooling and target enrichment
B
G Amplification of enriched libraries
WGS
Targeted sequencing

Figure 2. **Outline of an Illumina-compatible library preparation procedure**. **A**. DNA fragmentation of gDNA by either physical methods or enzymatic digestion. **B**. End-repair of DNA fragments to generate phosphorylated blunt ends. **C**. Addition of an adenine at the 3' ends. **D**. Ligation of Y-shaped sequencing adapters. **E**. PCR amplification of DNA libraries. At this point, one can proceed to WGS. **F**. Sample pooling and enrichment of selected regions. **G**. Amplification of enriched libraries. At this point libraries can be sequenced (targeted sequencing).

b) End-repair

Because the different fragmentation methods leave a mixture of DNA fragments with ends that may be incompatible with the subsequent steps (5' and 3' overhang ends), DNA's ends need to be blunted and 5'-phosphorylated using a mixture of the enzymes T4 polynucleotide kinase, T4 DNA polymerase and Klenow Fragment after fragmentation (**Figure 2B**)[15].

c) Adenylation

Because the efficiency of ligation of blunt ends is known to be remarkably low, a minimal sticky end is created by adding an overhang adenine to the blunted 3’ ends using either Taq polymerase or Klenow Fragment (**Figure 2C**)[15].

d) Adapter ligation

Subsequently, Y-shaped sequencing adapters with a T overhang are ligated to both ends of the A-tailed molecules at specific adapter:fragment ratios, usually determined by the manufacturer. This ratio is a key parameter, since an adapter-to-library excess can favour the formation of adapter dimers (self-ligated adapter molecules) that can be difficult to remove and that dominate in the subsequent PCR amplification steps (**Figure 2D**)[15].

e) Library amplification

Next, PCR amplification is performed to enrich for adapter-ligated DNA fragments. The number of PCR cycles, another key parameter, is usually determined in the manufacturer’s guidelines and is dependent on the initial DNA input. PCR reactions are then purified using either bead or column-based methods prior to quantification and evaluation of library size distribution, usually performed through fluorometric and capillary electrophoresis methods, respectively. At this point, one can proceed to WGS or continue with target enrichment (**Figure 2E**)[15].

f) Sample pooling and target enrichment

Several strategies have been developed to cut the costs associated with NGS. In order to be able to run multiple samples in one sequencing run, adapter molecules tagged with specific sequences known as “barcodes” or “indices” can be used. These are short sequences typically 6-8 bp long that are co-sequenced with the library molecules. During library preparation, each sample is ligated to adapters that have a single barcode, and samples with different barcodes can then be pooled (mixed) together. After sequencing, in a bioinformatic process known as demultiplexing, each sequenced read can be assigned to a specific sample through its barcode[15]. In combination with sample pooling, specific genomic regions can be selected through targeted enrichment. This is typically achieved using hybridization-based methods that use tagged baits or multiplex-PCR (amplicon-based) methods that can amplify thousands of regions simultaneously. Once the libraries are ready, they are diluted to a specific molarity to be loaded onto the sequencing instrument (**Figure 2F**).

g) Amplification of enriched libraries

Finally, enriched libraries are amplified and purified using either bead or column-based methods and evaluated as described in paragraph **e)** prior to sequencing (**Figure 2G**)[15].

h) The Illumina sequencing process

In the Illumina method, which is a SBS method similar to Sanger sequencing, millions of different DNA template strands are loaded into the flow cell, the microfluidics chamber in which the preparatory and sequencing reactions take place (**Figure 3**).

Figure 3. **The Illumina sequencing process**. A researcher holds a mid-output flow cell, ready to be loaded onto a NextSeq 500 Illumina platform. (Photo credit: qGenomics).

Inside the flow cell, the library fragments are hybridized to DNA priming sequences that are bound to the flow-cell's surface. These priming sequences are complementary to the adapter regions in the DNA library molecule and allow binding of the DNA polymerase. Each DNA molecule is then amplified into distinct clonal clusters through bridge-amplification to increase the amount of DNA that will provide light emission subsequently, which will increase signal-to-noise ratio later in the process. When cluster generation is complete, the templates are ready for sequencing (**Figure 4A**).

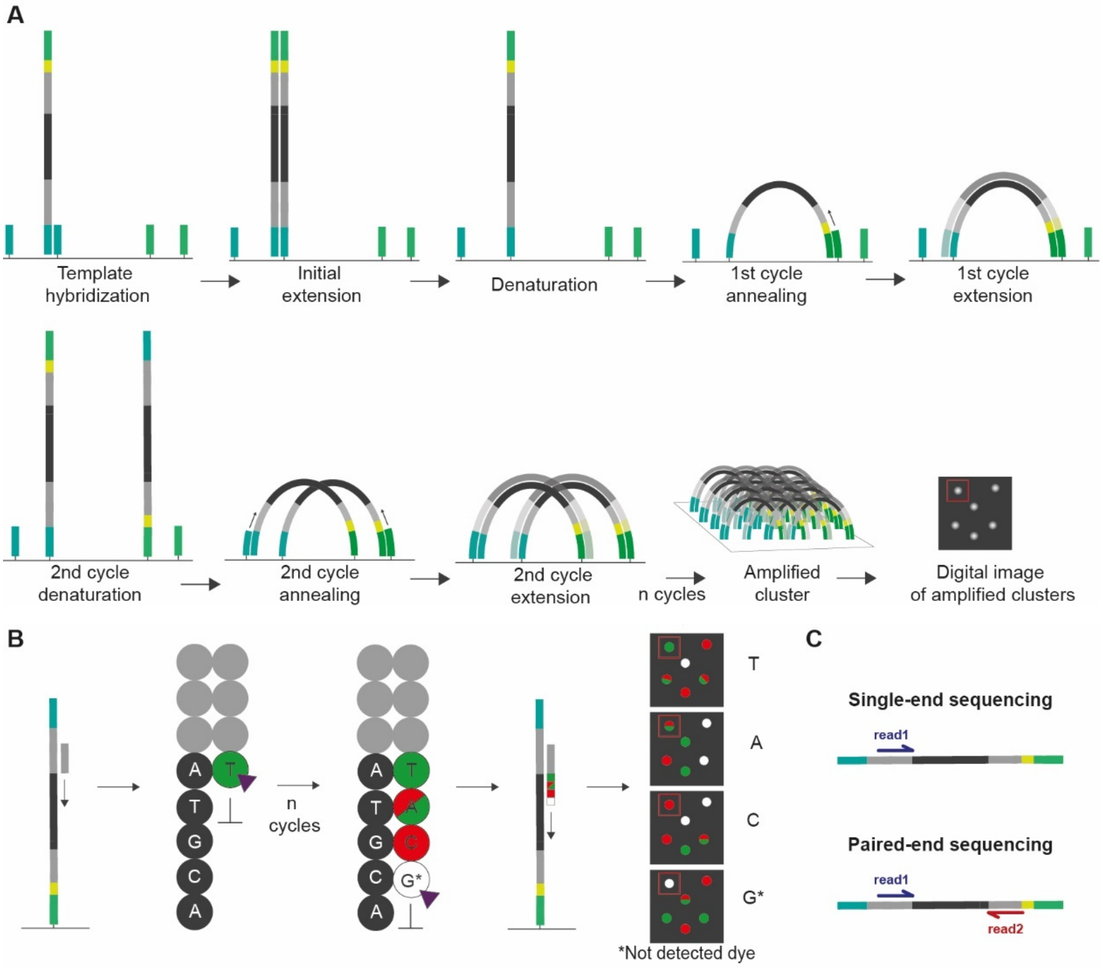


Figure 4. **Schematic view of Illumina sequencing technology in a NextSeq system**. **A**. Cluster amplification. Library fragments hybridize to DNA priming sequences that are bound to the flow-cell's surface. Each DNA fragment is then amplified into distinct clonal clusters through bridge-amplification prior to the sequencing reaction. **B**. Sequencing reaction. For each sequencing cycle, a single base containing a 3'-terminator (here represented as a purple triangle) that blocks further polymerization is incorporated to each growing DNA copy strand. Next, unbound dNTPs are removed, and the flow cell's surface is imaged to identify which dNTP was incorporated. Finally, bound dNTPs are converted to regular bases so that they become extendable and non-fluorescent and the process can be initiated again. **C**. Single and paired-end sequencing options.

During each sequencing cycle, a mixture of all four individually labelled deoxynucleotides (dNTPs) are added. Such dNTPs contain a 3'-terminator that blocks further polymerization so that only a single base can be added by a polymerase enzyme to each growing DNA copy strand. After the incorporation of a single dNTP to each elongating complementary strand, unbound dNTPs are removed, and the flow cell's surface is imaged to identify which dNTP was incorporated at each specific cluster region in the flow-cell. The identification of dNTPs is achieved through total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels. In most Illumina platforms, each dNTP is bound to a single fluorophore that is specific to that base type and requires four different imaging channels, whereas in the NextSeq (the sequencing platform used in this work), the NovaSeq and Mini-Seq systems, a two-fluorophore system is used[14]. Next, in a step unique to NGS, the modified bases are converted to regular bases, such that they become both extendable and non-fluorescent. This restoration process primes them to undergo subsequent rounds of single-base extension and imaging. At the end of a sequencing run with imaging cycles, the fluorescence color at each template position in each image is mapped to a base (i.e., A, T, C, or G) (**Figure 4B**).

The bases from a single template position are concatenated to yield a DNA sequence of length *n*, called a ‘‘read’’. Typically, reads in Illumina sequencing range from 36 bp to 500 bp long[10,13]. Libraries can either be sequenced from one end only (known as single-end sequencing, SE) or from both ends (known as paired-end sequencing, PE). For instance, a sequencing run that can read 300 bp can be set to read the 300 bp from one single end or run so that 150 bp are read from each end (**Figure 4C**). Once the sequencing run is complete, data can be downloaded, processed and analyzed using bioinformatic pipelines.

1.2.2 NGS data processing

Once a sequencing run has finalized, a number of bioinformatic steps must be followed in order to pre-process the raw data and turn it into genomic data that can be analyzed. These steps are usually aggregated into what is commonly known as a bioinformatic pipeline. Next, we briefly introduce the most common steps included in a genetic bioinformatic pipeline. Typically, these generic steps are customized to meet the specific needs of each application.

a) Sample demultiplexing

This step allows to assign each sequence read to a specific sample based on the information provided in the barcode portion of the ligated adapter.

b) Adapter and low-quality base trimming

Bases with low sequencing quality (dubious) and potential adapter sequences are removed from the reads to facilitate their mapping.

c) Sample mapping

The trimmed reads are mapped by an algorithm against a reference genome.

d) Mapped reads processing

The mapped reads go through a number of steps directed at ensuring read quality. For instance, the mapped reads are sorted, and potential read duplicates are removed. Additional quality filters can be applied, for instance, to remove not uniquely mapped reads.

At this point several paths can be taken, for instance to know how many reads fall within or how many genetic variants are found in a given genomic interval.

As mentioned above, the emergence of NGS technologies, bioinformatic tools and procedures has allowed the rapid development of a set of minimally invasive tools that have revolutionized current screening strategies, providing higher sensitivities and specificities compared to classical screening approaches.

1.3 Cell-free DNA as a minimally invasive tool

Minimally invasive methodologies, based on the use of blood biomarkers, have already had a great impact in some key clinical settings. One of the biomarkers used that currently holds more promise in minimally invasive techniques is cell-free DNA (cfDNA). In the following sections, several key aspects of cfDNA's structure, origin and clinical implications are introduced.

1.3.1 cfDNA originates from chromatin

CfDNA is a mixture of short, double-stranded DNA (dsDNA) molecules present at very low levels in a variety of physiological circulating fluids, including blood, lymph, bile, milk, urine, saliva, mucous suspension, spinal fluid, and amniotic fluid. Although different cellular release processes, cfDNA structures, determinants of steady-state levels in plasma and cellular origins have been described for cfDNA (introduced in more detail in **Sections 1.2.2** to **1.2.5)**, the formation of cfDNA ultimately derives from the cleavage of chromatin by nucleases under both physiological and pathological conditions.

Chromatin is the macromolecular protein-DNA complex that packages and regulates the genome within the nucleus of eukaryotic cells[16]. The fundamental unit of chromatin is the nucleosome, which consists of a stretch of ~147 bp of dsDNA wrapped around an octamer of the 4 core histones (two molecules each of histone H2A, H2B, H3, and H4, all highly conserved proteins). Additional histone H1 plays a role in organizing the structure of the 30 nm chromatin fiber formed by coiled nucleosome fibers. Adjacent nucleosomes are connected by a short stretch of linker DNA, whose length varies (estimations range from 8 to 114 bp) both between species, different regions of the genome and cellular types, and is the most susceptible region to deoxyribonucleases (DNase) cleavage (**Figure 5**)[17–19].

When nucleases cleave chromatin, both *in vivo* and *in vitro*, they do so preferentially at the linker spaces between nucleosomes. After nuclease cleavage, multimers of the nucleosome unit of variable length are released from chromatin (**Figure 6**). How these units travel through the body depends on the cellular process that originated them.

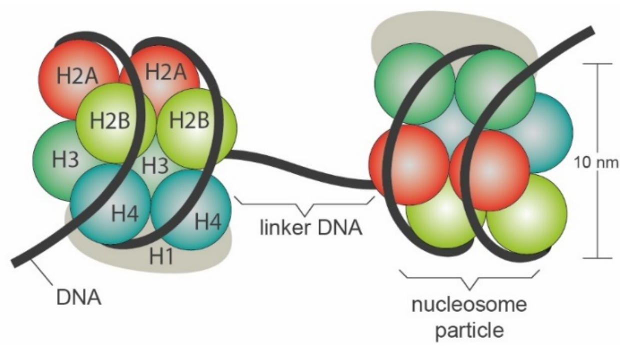


Figure 5. **Nucleosome structure**. Around 147 bp of dsDNA wrap around an octamer of the four core histones (H2A, H2B, H3 and H4) forming a 10 nm thick nucleosome filament that is further coiled into the 30 nm chromatin fiber, which is stabilized by histone H1. Two adjacent nucleosomes are depicted, which are linked by a stretch of linker DNA of variable-length (8 bp to 114 bp), which is not protected by histones from potential cleavage.

1.3.2 Cellular processes that lead to the release of cfDNA

To date, different cellular processes have been identified that lead to the formation and release of a variety of cfDNA complexes (see **Section 1.3.3** for cfDNA complexes). Among those processes, the most well-established include apoptosis, necrosis, and active secretion. Other processes that have been discovered include oncosis, phagocytosis, autophagocytosis, and NETosis[20].

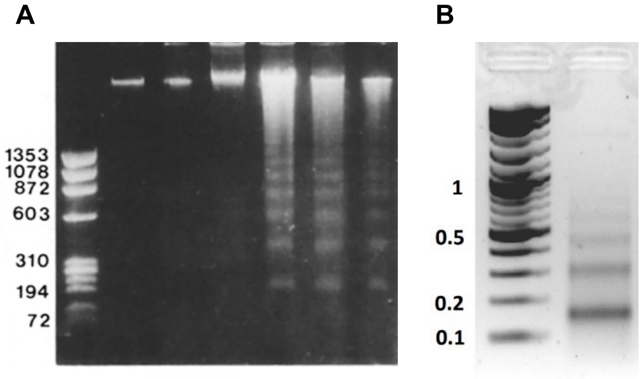


Figure 6. **DNA nucleosomal ladder patterns are generated from chromatin cleavage both *in vivo* and *in vitro*. A**. Agarose gel electrophoresis of DNA nucleosomal ladder patterns from apoptotic cells. Molecular weight marker seizes are expressed in base pairs (bp). Adapted from Gavrieli et al.[21]. **B**. Agarose gel electrophoresis of DNA nucleosomal ladder patterns generated from micrococcal nuclease-cleaved *Saccharomyces cerevisiae*'s chromatin *in vitro*. Adapted from Rodriguez et al.[22].

In apoptotic cells, cell shrinkage, blebbing of the plasma membranes, condensation and fragmentation of nuclei, and chromatin cleavage are observed. Chromatin cleavage, a hallmark of apoptosis, depends on caspase-activated DNases that, as mentioned above, preferentially cut the linked space between nucleosomes[17]. After cleavage, nucleosome particles plus some linker DNA are released into the bloodstream and produce a characteristic ladder pattern of multiples of ~166 bp dsDNA in all kinds of human subjects tested to date, including healthy individuals, pregnant women, organ transplant recipients and cancer patients. Typically, cfDNA fragments released from apoptotic cells are short, mostly enriched for mono- and dinucleosome-sized DNA fragments (**Figure 6A**) which resemble those fragments generated in micrococcal nuclease (MNase) experiments (**Figure 6B**). However, certain conditions can slightly shift the size distributions of the mononucleosome sized band, an observation of great interest for its potential application in the clinical practice. As an example, the estimated size distribution of placental-derived cfDNA (often referred as cell-free fetal DNA, cffDNA) is smaller than that described for cfDNA globally, and presents a main peak at ~143 bp. The smaller average size of cffDNA has been exploited to enrich for fetal-specific signatures[19].

Another important cellular process that can lead to the release of cfDNA is necrosis, which presents some notable differences in comparison to apoptosis: whereas apoptosis is a physiological process that can occur during normal growth and development as well as during disease, necrosis is the sporadic result of cellular injury induced by physical or chemical trauma[23]. In addition, while apoptotic cells spontaneously release endonuclease-cleaved DNA with the characteristic nucleosome-sized pattern, necrotic cells need to be engulfed by macrophages and other scavenger cells to release digested DNA to the extracellular environment. As a result, cfDNA originated from necrotic cells is characterized by presenting a higher molecular weight, typically >10 Kb in size[23].

Finally, the main cellular processes that involve active cfDNA secretion are characterized by the release of vesicles as exosomes, which mainly carry dsDNA both inside the vesicle and bound to the outer side of the vesicle's membrane. The membrane-bound cfDNA has found to be >2.5 Kb in length while the cfDNA inside the vesicles has been found to range from 100 bp to 2.5 Kb.

1.3.3 cfDNA structure and complexes

Depending on the cellular process that originated it, cfDNA can be found either as i) free circulating DNA molecules, ii) as molecular complexes with lipids and proteins or iii) vesicle-internalized or vesicle-bound. It should be noted that the term "cell-free DNA" refers to all extracellular DNA molecules that are present in plasma, regardless of the subcellular structures or complexes in which they are found.

Some of the molecular complexes and subcellular structures in which cfDNA can be found include the nucleosomes[24], the virtosomes[25] (or nucleic acid-lipoprotein complexes), neutrophil extracellular DNA traps[26] (nets of remodeled extracellular

DNA fibers bound to anti-microbial granules), DNA bound to serum proteins[27] (e.g., albumin and immunoglobulins) or DNA bound to the cellular membrane[27].

Some types of molecular complexes, like the nucleosomes, are thought to prevent cfDNA from being further cleaved by nucleases present in the circulatory system[28] while others, like neutrophil extracellular DNA traps, are thought to play a role in the defense against pathogens[26].

Addittionallly, cfDNA can also be found internalized in vesicles, which are composed of proteins and lipids that can also contain mRNAs and microRNAs. Vesicles are secreted by most cells and do not only protect cfDNA from nucleases present in blood but play an important role in intercellular communication and lateral transfer of material. Different types of vesicles have been described based on their size: exosomes (30 to 10 nm vesicles), microparticles or ectosomes (200 to 1000 nm vesicles) and apoptotic bodies (1 to 5 µm vesicles)[29,30].

Studies have shown that the structures in which cfDNA is found have an impact on its distribution in the body and can determine specific functions. For instance, it has been shown that cfDNA is an active carrier of genetic information between cells. This function is carried out through binding of cfDNA and plasma protein complexes to plasma protein receptors[31], or endocytosis of nucleosomes[32] or exosomes[33].

1.3.4 Determinants of steady-state levels of cfDNA in plasma

Under normal physiological conditions, the concentration of cfDNA in blood among healthy individuals is variable and typically very low (1 to 30 ng/mL)[27]. These low steady-state levels depend upon the rates at which cfDNA is released from different cell types into the bloodstream and the rates at which it is cleared from the organism. Under normal conditions, cfDNA is rapidly digested by DNases (typically in 15 to 30 minutes after release)[34] and rapidly cleared from the circulation by the liver and kidneys[35–37]. However, under certain physiological (intense physical exercise, pregnancy) and pathological conditions (cancer, sepsis, autoimmune diseases), the cfDNA release and clearance rates can substantially change, resulting in higher steady-state cfDNA levels in blood. For instance, it is well documented that intense physical activity causes cell death, which in turn results in an increased release of cfDNA into the bloodstream as a product of apoptosis of muscle cells[38]. Additionally, cell death from a growing placenta or a growing tumor can also result in increased release of cfDNA[39,40].

The presence of high steady state cfDNA levels have been shown to have a high autoimmune and inflammatory potential. A well-studied example is systemic lupus erythematosus (SLE), an autoimmune disease in which the presence of large amounts of cfDNA are observed bound to anti-cfDNA antibodies forming complexes that in turn contribute to decrease the accessibility to cfDNA by DNases[41].

On the other hand, high levels of cfDNA in late gestation have been shown to trigger an inflammatory process that leads to maternal, fetal, and placental endocrine events related with the onset of parturition[40].

1.3.5 Cellular origins of cfDNA

In healthy individuals, under normal conditions, the bulk of cfDNA is believed to be originated from hematopoietic cell lines with minimal contributions from other tissues[42]. Snyder and colleagues[18] conducted experiments that strongly support the hematopoietic origin of the bulk of cfDNA in healthy individuals. They demonstrated that it is possible to identify what cell types are majorly dying by looking at key pieces of information in cfDNA such as the nucleosome cleavage patterns (i.e., nucleosome positions), which allowed them to generate genome-wide maps of *in vivo* nucleosome positioning that correlated most strongly with nucleosome positions typical of lymphoid and myeloid cell lineages. More recently it has been determined that 70% of the total cfDNA derived from hematopoietic cells originates from the white blood cell (WBC) lineage whereas the other 30% derives from the erythroid lineage[43].

However, as mentioned above, there are certain physiological or pathological conditions that trigger the release of cfDNA from tissues other than the hematopoietic due to enhanced cellular death in that particular tissue. One such physiological condition is cancer. The authors also showed that, in cancer patients, nucleosome positioning derived from cfDNA samples was not only correlated with nucleosome positions typical of lymphoid and myeloid cell lineages, but also with nucleosome positions from a specific tumor cell line, coinciding with the type of tumor present in these individuals.

In other studies, the analysis of tissue-specific DNA methylation markers in blood allowed the identification of cfDNA molecules derived from pancreatic β-cells in type 1 diabetes patients, oligodendrocytes in relapsing multiple sclerosis patients, brain cells after traumatic or ischemic brain damage, and exocrine pancreas cells in pancreatic cancer or pancreatitis patients[44].

Key to the use of cfDNA in prenatal diagnosis, *in vitro*[45,46] and *in vivo*[47–49] studies have shown that in pregnant women a fraction of cfDNA in the maternal blood is originated in the placenta as a result of the apoptosis that takes place during the process of fusion and differentiation of cytotrophoblast with syncytiotrophoblast cells (**Figure 7**)[50,51].

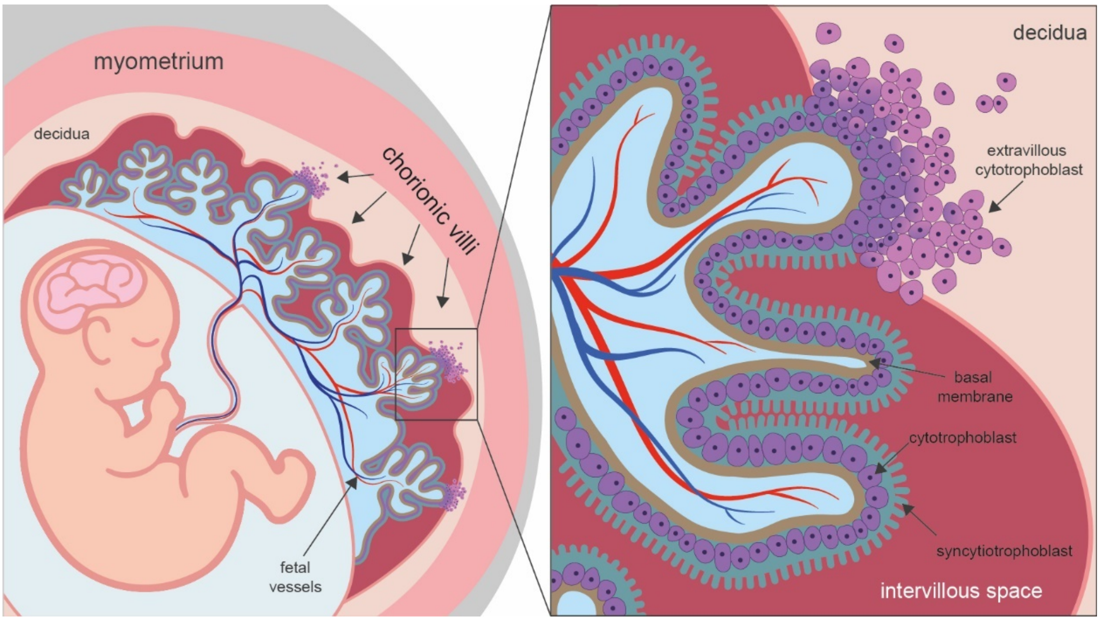


Figure 7. **Schematic representation of chorionic villi in the placenta.** Chorionic villi, delimited by a double layer of trophoblastic cells, project into the intervillous space. The outermost layer (syncytiotrophoblast) prevents the interaction between fetal antigens and the maternal immune system and is responsible for the production of most placental hormones. At the tips of the villi, the cells of the inner layer (cytotrophoblast) undergo an epithelial-mesenchymal transformation, differentiating into extravillous cytotrophoblast (EVT) and invading the decidua.

1.4 Historical background of cfDNA

The major findings related to cfDNA, from its biology to its potential applications, have been mainly fueled by the cancer and noninvasive prenatal testing (NIPT) fields. These major findings, which are described below, are summarized in **Figure 8**.

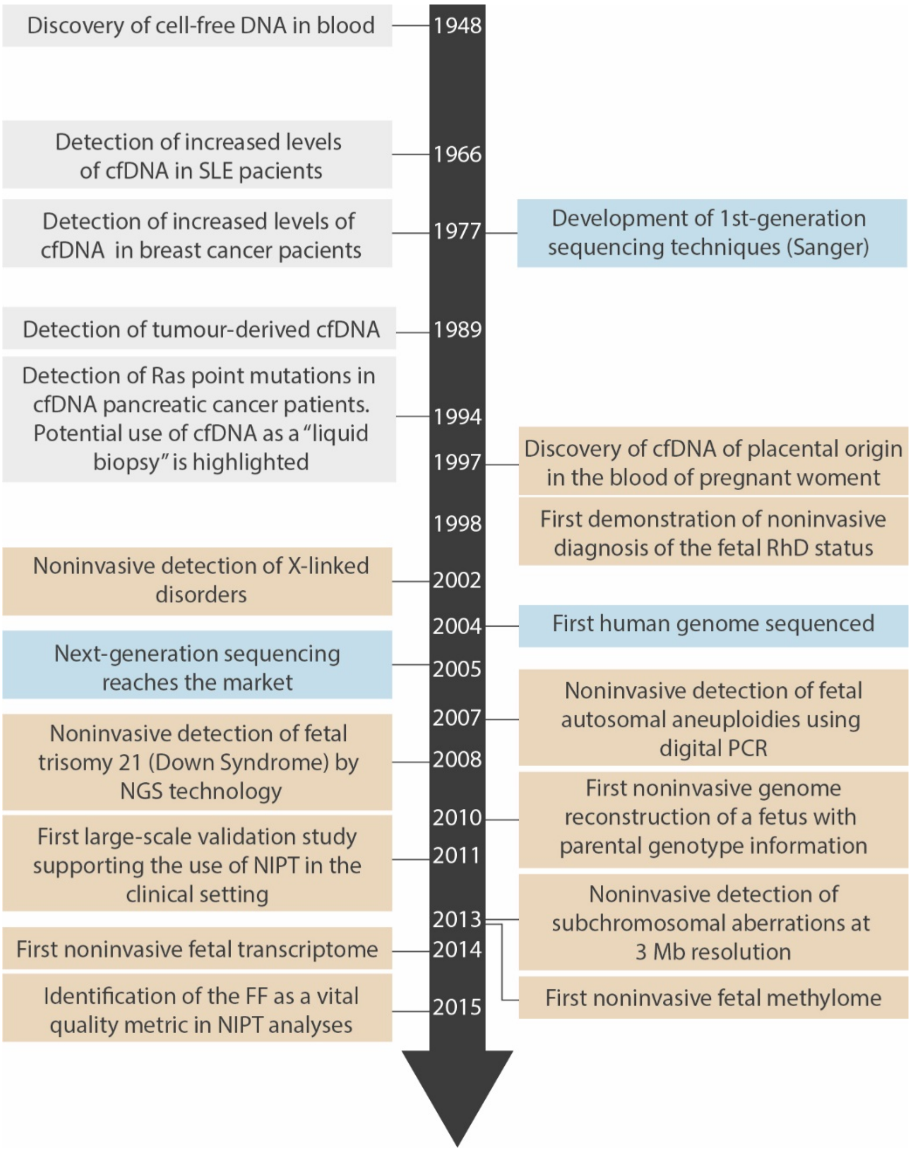

Discovery of cell-free DNA in blood
1948
Detection of increased levels of cfDNA in SLE pacients
1966
Detection of increased levels of cfDNA in breast cancer pacients
1977
Development of 1st-generation sequencing techniques (Sanger)
Detection of tumour-derived cfDNA
1989
Detection of Ras point mutations in cfDNA pancreatic cancer patients. Potential use of cfDNA as a "liquid biopsy" is highlighted
1994
1997
Discovery of cfDNA of placental origin in the blood of pregnant woment
1998
First demonstration of noninvasive diagnosis of the fetal RhD status
Noninvasive detection of X-linked disorders
2002
2004
First human genome sequenced
Next-generation sequencing reaches the market
2005
2007
Noninvasive detection of fetal autosomal aneuploidies using digital PCR
Noninvasive detection of fetal trisomy 21 (Down Syndrome) by NGS technology
2008
2010
First noninvasive genome reconstruction of a fetus with parental genotype information
First large-scale validation study supporting the use of NIPT in the clinical setting
2011
2013
Noninvasive detection of subchromosomal aberrations at 3 Mb resolution
First noninvasive fetal transcriptome
2014
Identification of the FF as a vital quality metric in NIPT analyses
2015
First noninvasive fetal methylome

Figure 8. **Main discoveries related to cfDNA and DNA sequencing achievements from 1948 to 2015**. Sequencing-related achievements are highlighted in blue boxes, cfDNA-related general discoveries are highlighted in gray boxes and NIPT-related discoveries are highlighted in orange boxes.

Initial findings related to the presence of cfDNA in the bloodstream were mainly achieved in the cancer and the autoimmune disease fields and were limited to the observation and analysis of the differential levels of extracellular nucleic acids between healthy and diseased individuals. On one hand, initial observations of the presence of both extracellular DNA and RNA in the blood of healthy and cancer patients were made in 1948 by Mandel and Métais[52]. These early observations were confirmed in 1977 by Leon and colleagues[39] in patients with metastatic breast cancer, who also showed higher cfDNA levels compared to healthy individuals. On the other hand, Tan and colleagues[53] observed, in 1966, high levels of cfDNA in the blood of SLE patients that lead to the formation of anti-double-stranded DNA antibodies, which could explain the set of autoimmune reactions observed in those patients.

From then on, studies related to cfDNA in the cancer field were expanded and more knowledge on the biology of cfDNA was acquired: in 1989, it was first shown that a proportion of the detected cfDNA molecules in the blood of cancer patients contained tumor-derived signatures (double-strand instability)[54], and in 1994, its potential clinical relevance as a "liquid biopsy" was highlighted by the detection of *Ras* point mutations in cfDNA from patients with pancreatic cancer and myelodysplastic syndrome[55,56]. The initial detection of *Ras* mutations led to a wealth of studies that analyzed both genetic and epigenetic alterations, such as microsatellite instability or loss of heterozygosity (LOH)[57] and aberrant methylation in cfDNA extracted from the plasma or serum of patients with cancer[58].

Nowadays, analysis of tumor-derived cfDNA (ctDNA), aims at providing a means of molecular profiling for tumors that are difficult or unsafe to biopsy in a minimally invasive way. Not only that, but it may potentially better capture the molecular heterogeneity harbored by multiple distinct clonal populations in a patient's tumor, as compared with a needle biopsy of a single tumor lesion, and it may allow earlier tumor detection (screening) and monitoring in patients without clinically evident disease as well as assessing sensitivity and resistance to targeted therapies[59].

Although cfDNA has been widely studied in the oncology field, the most evident current commercial application is found in the prenatal testing clinical setting, in which the most important achievements are summarized below.

The discovery in 1997 by Dennis Lo and colleagues [60] of cfDNA of placental origin, named as cell-free fetal DNA (cffDNA), allowed for the first time the noninvasive determination of fetal sex. At the time, this was done through the identification of Y-chromosome-derived sequences in male-bearing pregnancies using real-time quantitative PCR (real-time qPCR). This simple approach rapidly had an impact on the use of CVS and amniocentesis for the detection of X-linked disorders[61], since it potentially allowed to exclude from the test those pregnancies bearing female fetuses. A similar approach, based on the identification of DNA derived from the rhesus factor D (RhD) gene in the blood of RhD-negative pregnant women through real-time qPCR, allowed to determine the fetal RhD status, which opened the possibility to avoid performing CVS and amniocentesis to those RhD-negative women bearing RhD-negative fetuses[62].

The noninvasive detection of fetal chromosomal aneuploidies (that is, an abnormal number of chromosomes in the genome) was initially achieved through the use of digital PCR[63,64]. However, because digital PCR only amplifies DNA molecules in the maternal plasma that have homology for the predetermined PCR primers, most plasma DNA molecules, which do not have homology with the primers, are not analyzed and make this approach not really useful for diagnostic approaches.

Noninvasive fetal aneuploidy detection in maternal plasma became a reality in 2008 thanks to the development of NGS technologies. The first studies used small cohorts (13–86 cases and 34–410 control samples) and focused on the detection of trisomy of chromosome 21 (T21), which is the cause of Down syndrome[65–67]. These initial studies were expanded three years later, in 2011, with the first large-scale clinical evaluation of NIPT for the detection of Down syndrome. The large number of samples analyzed (4,664 pregnancies at high risk for T21), together with the high sensitivity and high specificity, supported the introduction of the noninvasive detection of chromosome T21 in the clinical setting for high-risk pregnancies[67]. From then on, trisomies of chromosomes 13 (T13) and 18 (T18) as well as sex chromosome aneuploidies have been evaluated and current commercial tests are able to detect them.

Since the advent of NGS, there have been numerous innovations and additions aimed at increasing the scope of the tests and improving the methods through which NIPT is performed: for instance, the use of maternal genotype information together with paternal genotype data allowed to reconstruct, in 2010, a genome-wide genetic map of the fetus, which opened up the possibility of a noninvasive genome-wide scanning of potential fetal genetic disorders such as β-thalassemia or congenital adrenal hyperplasia[68–70].

Three years later, in 2013, the feasibility of performing the noninvasive prenatal detection of fetal chromosomal microdeletions and microduplications at 3 Mb resolution genome-wide was demonstrated. This represented the first noninvasive fetal molecular karyotype ever generated[71]. Nowadays, several companies offer extended versions of their tests that cover several microdeletion syndromes of known clinical significance such as the 22q11.2 deletion, which has an estimated prevalence of 1 in 992 in low-risk population defined by chromosomal microarray analysis[72]. More recent achievements in the field include the noninvasive determination of the fetal methylome and transcriptome[73,74].

Other than progressively increasing the scope of the tests, the advances in the field enabled to identify which factors can influence the test performance such as maternal weight, maternal age, gestational age[75] as well as methods of sample collection and shipping conditions that may lead to undesired maternal cell hemolysis[76]. In addition, the fetal fraction (FF), which is the proportion of cffDNA in the total cfDNA in blood, was defined a vital quality metric for NIPT analyses: if the FF is too low, it becomes difficult to accurately distinguish an euploid from an aneuploid fetus, which can lead to false negative results (see **Section 1.7**)[76].

1.5 General applications of cfDNA

Currently, the most widespread and commercially successful application of cfDNA is found in the context of NIPT, which is introduced in detail in the next sections.

In many other fields, cfDNA is regarded as a promising tool to improve or overcome some of the limitations of current screening or diagnostic methodologies. However, to date, none of these applications has reached the routine clinical practice. The most relevant examples are described below.

One important example is found in the field of oncology, where early detection, accurate molecular characterization of tumors and monitoring of disease progression are key to increase cancer patients' survival and to evaluate treatment efficacy[1,59]. However, in this context, several biological and technical limitations challenge the development of assays that can improve the sensitivity and specificity of the current screening and diagnostic invasive techniques. These include the fact that i) tumoral masses are genetically heterogeneous, ii) ctDNA is mixed in the blood stream with cfDNA from other non-tumor sources and iii) most NGS methodologies are error-prone, which hinders the specific identification of low-frequency mutations[77].

Other potential applications of cfDNA include allograft transplant rejection monitoring in transplanted patients through quantitation of the levels of the donor-derived cfDNA in the blood of the recipient[78], diagnosis and prognosis of autoimmune disorders (SLE)[79], and assessment of myocardial infarction and stroke severity and outcome[80].

Finally, another promising example is found in the field of medically assisted reproduction, in which pre-implantation diagnosis is currently performed by aspiration of one or two cells from the developing embryo, potentially risking the successful implantation of the embryos. This risk could potentially be avoided by the analysis of cfDNA released into the cell culture medium by the embryos[81].

1.6 The use of cfDNA in the context of NIPT to improve current prenatal screening strategies

In this section, the strengths and weaknesses of current screening strategies to detect chromosomal aneuploidies and how the use of cfDNA can overcome some of these limitations is introduced from a historic perspective.

Globally, chromosomal aneuploidies are a major cause of perinatal death and birth defects, with an estimated rate of fetal death at 12 weeks (when first trimester screening is performed) about 30% for T21 and 80% for T18 and T13[82]. For a woman aged 35 at 12 weeks gestation, the estimated risks are 1 in 250 for T21, 1 in 600 for T18 and 1 in 1800 for T13, and the risks of delivering an affected baby at term are 1 in 350, 1 in 4000 and 1 in 10000[83]. Several factors modulate the above-mentioned risks, one of the most well-known being maternal age, which increases the overall aneuploidy risk.

1.6.1 First trimester screening for the detection of chromosomal aneuploidies

The use of screening strategies to detect chromosomal aneuploidies dates back to the 1970s, when maternal age was first used to set up a cut-off (35 years) to define a high-risk pregnancy for T21. Using only maternal age, the detection rate (DR) for T21 was 30%, and the false positive rate (FPR) was 5%[83].

In the subsequent years, it was shown that chromosomal aneuploidies are associated with altered maternal serum concentrations of various fetoplacental molecules, including α-fetoprotein (AFP), β-human chorionic gonadotropin (β-hCG), inhibin A, unconjugated estriol (uE3) and pregnancy-associated plasma protein A (PAPP-A)[84–88]. The concentrations of these markers were used to modify the *a priori* maternal age-related risk to derive the patient-specific risk for chromosomal aneuploidies. Screening at the second trimester by combining maternal age with AFP, free β-hCG, uE3 and inhibin A (quadruple test), contributed to increase the DR from 70% to 75% for T21 at a FPR of 5%[83,89,90].

In the 1990 to 2000s, it was shown that combining maternal age, serum markers (free β-hCG and PAPP-A) and nuchal fold thickness during the first trimester was superior than second trimester screening quadruple test. Serum concentrations of these placental products are affected by maternal characteristics (ethnicity, weight, smoking and method of conception as well as the machine and reagents used for the analysis) which are also considered to derive the patient's specific aneuploidy risk. Several studies have shown that the use of first trimester combined screening increases the DR for T21 to 90% at a FPR of 5%[91–97]. The DR for T21 can be further improved to 93-94% when the biochemical testing is performed at 9 to 10 weeks and ultrasound at 12 weeks[98–101]. In the context of the Catalan clinical setting, guidelines currently recommended to perform biochemical testing between weeks 8 and 13.6, and NT measurement between weeks 11.2 and 13.6 [3].

Additional first trimester sonographic markers of T21 include absence of the nasal bone, increased impedance to flow in the ductus venosus and tricuspid regurgitations, which contribute to a small increase of 93% to 96% in the DR and a small decrease in the FPR to 2.5% for T21. However, these additional markers are not still validated enough to be used in the risk calculation[83,102–104].

Screening for T21 also allowed screening for T13 and T18: all three trisomies are associated with increased maternal age, increased fetal NT and decreased maternal serum PAPP-A. However, some markers are different between the detectable trisomies: serum free β-hCG is increased in T21 and decreased in T13 and T18, and T13 is associated with fetal tachycardia while T18 and T21 are not[105–107]. Sonographic markers, some of them shared with T21, are also useful to determine the risk for these trisomies. Based on the first trimester combined screening results, a patient-specific risk for aneuploidy is assigned (**Table 1**).

Table 1. Risk groups and cut-off intervals in first trimester combined screening.

Risk group	Risk intervals
Very high	1/9 - 1/2
High	1/250 - 1/10
Intermediate	1/1100 - 1/251
Low	< 1/1100

The diagnostic accuracy of combined first trimester screening for T21, T18 and T13 was reviewed in a recent prospective validation study by Santorum M et al.[108]. In this study, 108,112 pregnancies at 11+0 to 13+6 weeks' gestation were examined. The DR and FPR of first trimester combined screening were evaluated at different risk cut-offs (**Table 2**).

Table 2. Estimated DR and FPR in screening by a combination of maternal age, fetal NT thickness, serum free β-hCG and PAPP-A at 11-13 gestation weeks. Data is presented as % (95% CI). Adapted from Santorum et al.[108].

Risk cut-off	DR (%)			FPR (%)
	T21	T18	T13	
1 in 10	73 (69 - 77)	91 (86 - 96)	75 (66 - 85)	0.67 (0.64 - 0.71)
1 in 100	90 (87 - 92)	97 (93 - 99.9)	91 (90 - 93)	3.90 (3.82 - 3.99)
1 in 200	93 (91 - 95)	98 (94 - 99.9)	93 (86 - 99.9)	6.46 (6.36 - 6.56)
1 in 1000	98 (96 - 99)	99 (97 - 99.9)	97 (92 - 99.9)	19.26 (19.08 - 19.43)

What the authors found is that DR is maximized at lower risk-cutoffs at the expense of higher FPR. A previous study showed better DR figures at a cut-off of 1 in 100 at 11-13 weeks' gestation: 98% of monosomy X, 97% of triploidies and 55% of other chromosomal abnormalities were also detected[109].

1.6.2 cfDNA screening for the detection of chromosomal aneuploidies

A major improvement in screening for T21, T18 and T13 has been achieved through the analysis of cfDNA in maternal blood. A recent meta-analysis published in 2017 of 35 relevant clinical validation and implementation studies reported that cfDNA screening in singleton pregnancies was able to detect >99% of fetuses with T21, 98% of T18 and 99% of T13 at a combined FPR of 0.13%[110]. It should be noted, however, that the performance of NIPT is better documented in T21 and T18 than in T13, the latter being a much less frequent condition. For sexual chromosome aneuploidies (SCAs), the number of samples tested was very low, and therefore, as stated by Gil and colleagues[110], more validation studies are needed. In addition, the detection of T21 in twin pregnancies showed a DR of 100% at a FPR of 0%, but again, despite these encouraging figures, a reduced number of cases have been examined (**Table 3**).

Table 3. **Estimated DR and FPR of chromosomal aneuploidies by cfDNA analysis**. Data is presented as % (95% CI). Adapted from Gil et al.[110]

Chromosomal aneuploidy	DR (%)	FPR (%)
T21 (1,963 cases)	99.7 (99.1–99.9)	0.04 (0.02–0.07)
T13 (119 cases)	99.0 (65.8–100)	0.04 (0.02–0.07)
T18 (563 cases)	97.9 (94.9–99.1)	0.04 (0.03–0.07)
X0 (36 cases)	95.8 (70.3–99.5)	0.14 (0.05–0.38)
XXX, XXY and XYY (17 cases)	100 (83.6–100)	0.004 (0.0–0.08)
T21 twin pregnancies (24 cases)	100 (95.2–100)	0 (0–0.003)

Most of the studies included in this meta-analysis involved high-risk pregnancies and were not confined to pregnancies in the first trimester. However, subgroup analysis of cfDNA testing in singleton pregnancies for T21 and T18 demonstrated no significant difference in performance of screening between high-risk and routine or mixed populations and between those examined in the first trimester and those examined at any stage in pregnancy. Similarly, there was no obvious difference in performance of screening between the diverse methods for cfDNA testing (detailed in **Section 1.6**).

Because cfDNA testing shows higher DRs at lower FPRs compared to first trimester combined screening, there is an ongoing debate as to whether cfDNA should be applied as a first tier screening tool or whether it should be used subsequent to first trimester combined screening. Reasons against using cfDNA as a first-line screening include high costs and the fact that the first trimester combined test has some unique benefits like the detection of many major fetal defects, diagnosis of multiple pregnancy and its chorionicity, detection of chromosomal defects other than T21, T18 and T13, and early prediction of pregnancy complications (e.g. pre-eclampsia)[108].

In the Catalan setting, a strategy in which cfDNA testing is offered based on the first trimester screening results was presented in June 2018. In pregnancies showing a very high-risk outcome from first trimester combined screening, invasive testing is directly offered. Pregnancies with mid to high risk are offered cfDNA testing as a first option followed by invasive testing if positive results (i.e., high risk of aneuploidy) are obtained, or invasive testing as a first option. Pregnancies falling in the intermediate risk subgroup are offered cfDNA screening, while pregnancies falling in the low risk would are not offered neither invasive nor cfDNA testing (**Table 4**)[3,108].

Table 4. Risk groups in first trimester combined screening test and subsequent screening and diagnostic strategies.

Group (risk intervals)	Strategy followed
Very high (1/9 - 1/2)	Invasive diagnosis
High (1/250 - 1/10)	Invasive diagnosis or cfDNA screening
Intermediate (1/1100 - 1/251)	cfDNA screening
Low (< 1/1100)	None

1.6.3 cfDNA screening for the detection of other alterations

Although in the context of NIPT the main use of cfDNA is found in the detection of chromosomal aneuploidies, starting in 2013, several companies expanded the number of conditions tested by covering a number of clinically important and relatively common subchromosomal aberrations such as the DiGeorge, Cri-Du-Chat, Prader–Willi/Angelman and the 1p36 deletion syndromes, which are all caused by recurrent microdeletions[111].

Other applications of cfDNA in the NIPT field include the RhD status determination, the fetal sex determination and the detection of fetal single-gene disorders through real-time qPCR or digital PCR.

RhD incompatibility describes the situation when a RhD-negative woman is pregnant with a RhD-positive fetus. If prior RhD antigen sensitization has occurred (e.g., due to a previous pregnancy of a RhD-positive fetus or due to fetal injuries during invasive diagnostic techniques), maternal preformed anti-RhD antibodies may cross the placenta and lead to hemolytic disease of the fetus and newborn. To avoid that, anti–D-immunoglobulin injections are systematically offered to RhD-negative pregnant women to reduce the chances of prenatal immunization, even though up to 40% of these women will have a RhD-negative fetus[112]. Because the genome of a RhD-negative woman does not contain the RhD gene, the noninvasive detection of RhD sequences through real-time qPCR in plasma indicate a RhD-positive fetus, and this information can be used to avoid unnecessary anti-D immunoglobulin administration.

Another important application of cfDNA in the context of NIPT is the noninvasive fetal sex determination in cases where the mother is a carrier of X-linked conditions (such as the Duchenne muscular dystrophy) or when both parents are carriers of alterations for congenital adrenal hyperplasia (CAH). In one of the most common forms of CAH, 21-hydroxylase deficiency causes the overproduction of androgens, resulting in virilization of female fetuses, which can be avoided through the administration of steroids during pregnancy. However, steroid administration is not necessary in male fetuses, and as a consequence, knowing the fetal sex in advance can help manage this condition in those fetuses that have shown to have ambiguous genitalia on ultrasound[113].

Finally, another important application is the noninvasive detection of single-gene disorders, the most important being thalassemia, sickle-cell anemia and hemophilia. Using this method, the proportion in maternal plasma of normal and mutant alleles originating from the disease locus is interrogated using digital PCR and the mutational status of the fetus is inferred[114].

1.7 NIPT technologies for the detection of chromosomal and subchromosomal alterations

The release of cffDNA has an effect on the overall concentration of cfDNA in pregnant women. However, the fraction of cffDNA in the total cfDNA amount is considerably low. Studies have shown that cffDNA can first be observed as early as 4 gestation weeks, and the amount of cffDNA increases as the pregnancy progresses up to approximately 30%[43,113]. When measured between 11 and 13 gestation weeks in normal pregnancies (when NIPT is usually performed), the FF has been estimated to be approximately 10% (interquartile range, 7.8–13.0%)[115]. Therefore, the detection of any fetal defect requires highly sensitive technologies as well as robust quality controls.

Testing for the presence of fetal aneuploidies and subchromosomal alterations can be performed following the combination of different strategies. Based on what portion of the genome is analyzed there are: i) whole-genome approaches, based on the random sampling, amplification and sequencing of cfDNA molecules and ii) targeted-sequencing approaches, based on the specific sampling, amplification, targeted enrichment and sequencing or microarray analysis of specific regions or chromosomes of interest[116,117]. These two strategies are then coupled to diverse methods for the detection of potential chromosomal abnormalities: count-based and single-nucleotide-polymorphism (SNP)-based methods (**Table 5**)[118,119].

Table 5. Summary of the different cfDNA analysis methodologies based on the fraction of genome that is analyzed, technology used and commercial tests.

Fraction of genome analyzed	Aneuploidy detection technology	FF calculation	Commercial test
Whole-genome sequencing	Count-based NGS	Chromosome Y Methylation sensitive-restriction enzymes Size-based SeqFF	MaterniT21 Plus (Sequenom) NeoBona (Synlab) NIFTY™ (BGI) Verifi (Illumina)
Targeted sequencing	Capture enrichment coupled to count-based NGS	Genotype analysis with Bayesian estimation (SNP)	Veracity (NIPD genetics)
	DANSR technology coupled to count-based NGS or microarray	Genotype analysis with maximum likelihood estimation (SNP)	Harmony (Roche)
	SNP-specific amplification coupled to NGS	Maternal and fetal allele ratios (SNP, maternal gDNA required)	Panorama (Natera)

In count-based methods, cfDNA molecules (fetal and maternal) are sampled, amplified, sequenced, and mapped to specific chromosomes. The proportion of sequencing reads mapped to each chromosome is calculated for a set of control reference euploid samples, thus obtaining a null distribution of the copy-number values for each chromosome in euploid pregnancies. Next, the proportion of reads mapped to each chromosome for a given test sample is calculated. Since a trisomic fetus has 50% more genetic material of the trisomic chromosome-of-interest, the proportion of DNA from that chromosome will be higher than in euploid pregnancies. Thus, if the proportion of reads mapped to a chromosome of interest is higher than a predetermined threshold in the null distribution, a high trisomy risk for that chromosome is reported[120].

In contrast, SNP-based methods involve either the use of hybridization-based capture of the genomic regions of interest or the use of highly multiplexed PCR to amplify a set of SNPs on chromosomes of interest followed by NGS[121,122]. SNP methods base their detection power in the fact that the pregnant mother's cfDNA is a mosaic that contains father's variants. These methods incorporate maternal genotype information, which allows a more complex analysis than count-based methods, as it allows the identification of fetal and maternal contribution to the observed sequencing reads followed by a maximum likelihood evaluation that the fetus is either normal, aneuploid, triploid or that uniparental disomy is present.

One important difference between count-based methods and SNP-based methods is that count-based methods are not directly able to discriminate between maternal and fetal-derived cfDNA molecules.

This means, on one hand, that the ability of count-based methods to detect increased chromosomal dosage resulting from fetal trisomy is strictly dependent to the amount of cffDNA in maternal circulation, that is, the FF. As a consequence, at lower FFs, in which small deviations are expected, a high sequencing depth is required to detect chromosomal aneuploidies[123]. Another important consequence is that, with count-based methods, the FF estimation is challenging and usually relies either on cfDNA fragment size differences, methylation differences or coverage differences. Not being able to discriminate the fetal or maternal origin of cfDNA molecules can also lead to confounding results in a number of situations that can be found in the clinical setting such as the presence of vanishing twins, triploid fetuses, dizygotic multifetal pregnancies or maternal aneuploidies, among others[124,125].

In SNP-based methods, the discrimination between maternal and fetal-derived cfDNA allows to evaluate the maternal and fetal contributions to the observed reads which, compared to count-based methods, provides a higher sensitivity at lower FF and a direct and accurate estimation of the FF. These methods also enable to overcome important challenges commonly encountered when performing NIPT in the clinical setting as the presence of confounding situations above-mentioned. However, SNP-based methods also present a series of limitations.

One of the main limitations of SNP-based methods is that maternal genotype information is needed, and therefore, sample preparation and sequencing costs for each assay are increased.

Another limitation of SNP-based methods is that because the maternal genotype is required, egg-donor pregnancies may pose several difficulties due to the presence of additional maternally derived fetal alleles in the surrogate mother.

Finally, the fact that SNPs account for only 1.6% of the human genome, high sequencing depth and high-fidelity amplification are required to identify unambiguously affected pregnancies with small imbalances.

1.7.1 Count-based WGS

Approaches based on WGS (**Figure 9A**) were first described in 2008 and became commercially available between 2011 and 2012[65,126]. Some commercial tests developed using this strategy include MaterniT21®Plus test (Sequenom), the NeoBona test (Synlab), the NIFTY™ test (BGI) or the Verifi® test (Illumina) (**Table 5**)[127]. All of them are able to determine fetal sex, FF and the presence of autosomal and sexual chromosome aneuploidies, as well as a set of subchromosomal alterations in singleton and multiple pregnancies[128].

Because of the rather high production costs associated with WGS[129], targeted sequencing methods have been developed, which amplify and sequence specific genomic regions of interest instead of random regions from all chromosomes. Compared to WGS approaches, in targeted sequencing nearly all sequencing reads are useful in assigning fetal chromosomal copy number, significantly reducing the total number of analyzed reads and increasing efficiency[130]. This also means that only the regions of interest can be studied, and that any chromosomal abnormality out of these regions can be easily overlooked. At least three different types of targeted sequencing methods can be distinguished, which are introduced in the following sections.

1.7.2 Count-based targeted sequencing (solution hybridization enrichment)

In this approach, a set of genomic regions of interest are captured using biotinylated oligonucleotide probes, which allows the accurate detection of autosomal and sexual chromosome aneuploidies fetal sex, FF and several subchromosomal alterations in singleton and multiple pregnancies. One of the NIPT tests using this approach is the Veracity test (NIPD genetics) (**Figure 9B**)[131,132].

1.7.3 Count-based targeted sequencing and microarray-based aneuploidy detection

A count-based method using targeted sequencing, called Digital ANalysis of Selected Regions (DANSR), was developed in 2012. In this method, a set of 384 non-polymorphic loci across chromosomes 13, 18, 21, X and Y and 192 polymorphic loci across chromosomes 1 to 12 are selectively targeted using a set of proprietary oligonucleotide probes (DANSR assays). More specifically, cfDNA molecules are biotinylated, immobilized on streptavidin beads and annealed with the multiplexed DANSR oligonucleotide pool. Appropriate hybridized oligonucleotides (three for each locus of interest) are ligated with Taq ligase, eluted from the bound cfDNA molecules, and amplified using universal PCR primers. DANSR products containing non-polymorphic loci in chromosomes 13, 18, 21, X and Y are used for aneuploidy detection, whereas polymorphic loci across chromosomes 1 to 12 are used to calculate the FF. In addition, two quantitation methods are available, the first one being NGS-based and the other one microarray-based, the latter used in the commercial test[133]. Using the measured FF, the Fetal fraction Optimized Risk of Trisomy Evaluation (FORTE) algorithm is used to calculate the risk of the fetal aneuploidy[117,130,134]. An example of a commercial NIPT test using this technology is the Harmony test (Roche), which offers the detection of T13, T18 and T21 as well as X chromosome monosomy (Turner syndrome), XXY aneuploidy (Klinefelter syndrome) and other SCAs (XXX, XXY, XYY and XXYY) (**Figure 9C**).

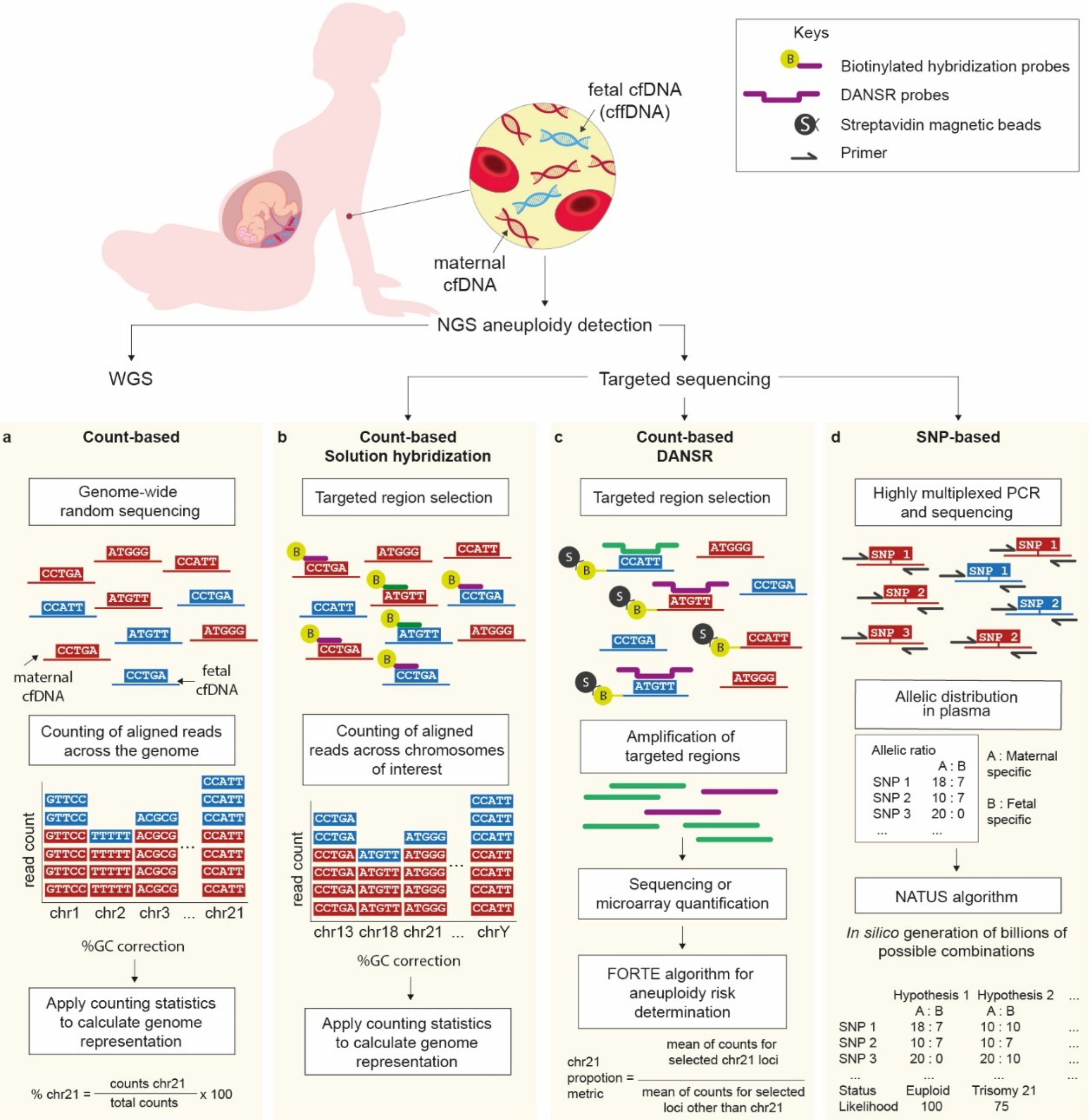

Keys
Biotinylated hybridization probes
DANSR probes
Streptavidin magnetic beads
Primer
fetal cfDNA (cffDNA)
maternal cfDNA
NGS aneuploidy detection
WGS
Targeted sequencing
a
Count-based
Genome-wide random sequencing
maternal cfDNA
fetal cfDNA
Counting of aligned reads across the genome
read count
chr1 chr2 chr3 ... chr21
%GC correction
Apply counting statistics to calculate genome representation
% chr21 = counts chr21 / total counts x 100
b
Count-based Solution hybridization
Targeted region selection
Counting of aligned reads across chromosomes of interest
read count
chr13 chr18 chr21 ... chrY
%GC correction
Apply counting statistics to calculate genome representation
c
Count-based DANSR
Targeted region selection
Amplification of targeted regions
Sequencing or microarray quantification
FORTE algorithm for aneuploidy risk determination
chr21 propotion metric = mean of counts for selected chr21 loci / mean of counts for selected loci other than chr21
d
SNP-based
Highly multiplexed PCR and sequencing
Allelic distribution in plasma
Allelic ratio A : B
SNP 1 18 : 7
SNP 2 10 : 7
SNP 3 20 : 0
A : Maternal specific
B : Fetal specific
NATUS algorithm
In silico generation of billions of possible combinations
Hypothesis 1 A : B
Hypothesis 2 A : B
SNP 1 18 : 7 10 : 10
SNP 2 10 : 7 10 : 7
SNP 3 20 : 0 20 : 10
Status Euploid Trisomy 21
Likelihood 100 75

Figure 9. **Noninvasive prenatal testing technologies for aneuploidy detection**. **A**. Genome-wide random sequencing. **B**. Count-based targeted sequencing of regions of interest captured with hybridization probes. **C**. Count-based targeted sequencing or microarray using the Digital Analysis of Selected Regions (DANSR) approach. **D**. SNP-based targeted sequencing using maternal genotype information.

1.7.4 SNP-based targeted sequencing through multiplexed PCR

The use of highly multiplexed PCR to amplify a set of SNPs on chromosomes of interest became commercially available in 2013 (Panorama test, Natera) and it used to evaluate 19,488 polymorphic loci covering chromosomes 13, 18, 21, X and Y[135,136]. A proprietary algorithm (NATUS) uses the maternal genotypes and recombination frequencies to generate billions of possible fetal genotype results. Based on the sequencing results, the algorithm calculates the relative likelihood for each potential fetal genotype and infers the chance of chromosomal copy number alterations[125]. In 2016, a number of modifications were introduced to this test: the number of interrogated SNP was decreased to 13,392 and the PCR chemistry was optimized to enable a one-step massively multiplexed PCR (as opposed to the previous two-step PCR) and to increase amplification uniformity among PCR targets. The detection algorithm was modified as well to improve overall test performance[137]. Highly multiplexed PCR SNP-based methods have also been shown to be able to detect clinically relevant microdeletion syndromes such as 1p36, Cri-du-Chat or DiGeorge, triploid fetuses and other complex situations like vanishing twins (see **Section 1.3**) (**Figure 9D**)[124,138].

Despite the main advantages and limitations of each technology have been described, studies comparing the performance of the diverse methods on a same set of samples is lacking. Not only that but these studies are often carried out by commercial companies, and thus key pieces of information may not be easily accessible or may vary among the literature. In addition, many NIPT validation studies do not publish the FF, gestational age, or maternal weight distribution of the population studied, despite these are factors known to have a direct impact on measured sensitivities and specificities. Therefore, NIPT performance in populations with higher-than-average maternal weights, or when tests are performed at earlier gestational ages for instance, may not match the results presented in validation studies[137]. The most important NIPT companies worldwide, together with the technology used, the conditions tested, the earliest sampling gestation week, the sensitivities and the costs are presented in **Table 6**. The costs, which vary depending on the number of conditions tested, have been obtained either from the company's website or through direct contact and are updated to July 2019.

Because of the importance an accurate FF calculation has on the overall performance of NITP, we have devoted the following Section to explaining the different approaches that have been developed to measure FF.

Table 6. Examples of several commercially available NIPT tests using the technologies described in Section 1.7 [113,129,139,140].

Test name	MaterniT21 Plus (Sequenom)	Verifi Plus Prenatal (Illumina)	NeoBona (Synlab)	Harmony Prenatal (Roche)	Panorama (Natera)
Technology for aneuploidy detection	WGS	WGS	WGS - Illumina technology and analysis software (VeriSeq)	DANSR microarray	SNP (targeted)
Conditions tested	T21, T18, T13, SCAs, fetal sex, microdeletions	All chromosomes, SCAs, fetal sex, microdeletions	All chromosomes SCAs, fetal sex, microdeletions	T21, T18, T13, SCAs, fetal sex	T21, T18, T13, SCAs, triploidy, fetal sex, microdeletions, vanishing twins
Earliest sampling	10 weeks	10 weeks	10 weeks	10 weeks	9 weeks
Sensitivity range	92- 9%	87-99%	*	80-99%	92-99%
Market cost	465 to 750 €	Not available	445 to 725 €	485 €	550 to 750 €
Delivery time	5 - 9 days	Not available	5 -10 days	3-5 days	5-7 days

* Not specified in the scientific validation paper, Illumina performance assumed.

1.8 Current methods for FF estimation

Various methods of FF measurement have been developed and employed in conjunction with different approaches to cfDNA testing for fetal aneuploidy[141]. It should be noted that these methods are not always described in detail in the literature, especially those used by count-based aneuploidy detection technologies, in which FF calculation is hindered by the fact that these methods do not discriminate between fetal and maternal DNA.

Below are described the methods for FF calculation that are commonly used in NIPT tests. On one hand, WGS count-based methods for aneuploidy detection usually rely on Y-chromosome-based approaches, cfDNA size-based approaches, methylation-based approaches, differences in sequencing counts (seqFF) or in some cases, on small SNP panels for FF calculation. On the other hand, the count-based targeted sequencing approach DANSR for aneuploidy detection (Harmony) relies on SNP-based FF calculation. Finally, methods that rely on SNP detection for aneuploidy detection also use SNPs to calculate the FF (Panorama). A summary of the FF methods described below can be found in **Table 5**.

1.8.1 FF estimation in WGS count-based methods

As mentioned earlier, count-based methods for aneuploidy detection are not able to distinguish fetal-derived and maternal-derived cfDNA molecules through genetic differences. Therefore, besides Y-chromosome-based approaches, which are applicable only to male pregnancies, other strategies are used and combined for FF determination.

a) Y-chromosome-specific sequences

In early works by Lo and colleagues[142], FF was calculated by measuring Y-specific contributions to the total circulating cfDNA when a male fetus was present. More specifically, the amount of the single-copy genetic marker Sex determining Region Y (SRY) located on chromosome Y (chrY) was compared to the amount of hemoglobin beta (HBB) autosome marker by real-time qPCR. Y-specific approaches are simple and accurate methods for FF determination, but the fact that they can only be applied to male-bearing pregnancies has led to the development of a number of alternative methodologies that can be also applied to female-bearing pregnancies[141].

b) Fetal methylation markers

This method is based on the differential DNA methylation profiles among body tissues[143,144]. Thus, regions that are differentially methylated in placenta compared to maternal blood cells can in principle be used to estimate the FF using methylation-sensitive restriction enzymes. This method was initially used by the MaterniT21 Plus test prior to the introduction, in 2015 of seqFF (explained in **paragraph D**). Although methods based on the use of methylation-sensitive restriction enzymes have a high correlation with Y-chromosome-based FF determination ($r = 0.85$, $P < 0.001$, Pearson correlation), they represent a different methodology that must be performed in parallel to WGS and need to be further verified in large-scale datasets generated from different research centers[141].

c) Size-based estimation

Fetal and maternal-derived DNA molecules in a plasma sample have been observed to exhibit different fragmentation patterns. More specifically, it has been reported that the size distribution of the total maternal cfDNA is characterized by a major peak centered at 166 bp with a series of smaller peaks occurring at 10 bp periodicities, suggesting that a predominant population of plasma DNA molecules have a size of 166 bp. In contrast, cffDNA molecules are found to have a dominant population centered at 143 bp[68,145]. The ratio between the count of fragments ranging from 100 bp to 150 bp and from 163 bp to 169 bp can in principle be used to estimate the FF, as it has been shown to have a high correlation with the Y-chromosome-based FF determination (r = 0.827, $P < 0.0001$, Pearson correlation)[141]. One of the commercial tests using cfDNA size distribution is the Neobona test (Synlab)[140]. In addition, this approach is used complementarily to the strategy described below.

d) Differential sequencing counts in regions overrepresented in cffDNA

Different methods attempting to directly estimate FF from routine WGS data have been developed. One of these approaches is seqFF, developed in 2015 by Sequenom with the goal to use it for FF calculation in the MaterniT21 Plus test. In this approach, normalized read counts within 50 Kb bins originating from all chromosomes except chromosomes 13, 18, 21, X, and Y are analyzed to fit a high-dimensional regression model. The model makes use of fragment size differences between fetal and maternal cfDNA to infer discrete regions in the genome that are overrepresented in cffDNA. Within these regions, the presence of a higher proportion of cffDNA results in larger count differences between maternal and fetal-derived cfDNA molecules which can then be used to train the model. This method showed a good correlation with Y-specific methods in two independent cohorts (r = 0.93 in both, Pearson correlation). However, such a high-dimensional model requires the use of a large amount of samples during training of the algorithm, and the performance appeared to be greatly deteriorated when the FF was below 5%, possibly because the number of cases with such FF was not sufficient to train the model[146].

1.8.2 FF estimation in count-based targeted sequencing methods

Methods using count-based targeted sequencing for aneuploidy detection (either solution hybridization or DANSR) use informative SNP loci to infer the FF.

In count-based targeted sequencing using solution hybridization, the FF is calculated from the observed allelic count distribution at heterozygous loci in maternal plasma using a binomial mixture model based on Bayesian inference. Then three possible informative combinations of maternal/fetal genotypes are used within the model to identify FF values that are strongly supported by the observed data[131]. One of the commercial tests using this approach is Veracity (NIPD genetics).

In the DANSR method, used by the Harmony test, a set of 192 loci on chromosomes 1 to 12 shown to contain informative SNPs are analyzed to calculate the FF. These are loci where fetal alleles differ from maternal alleles. For each locus, oligos differing by one base are used to query each maternal/fetal SNP. A maximum likelihood estimation approach is employed to determine the most likely FF based on the measurements from several informative loci[117,130,134]. One of the commercial tests using this approach is the Harmony test (Ariosa).

1.8.3 FF estimation in SNP-based methods

A direct way to estimate the FF is to use the maternal genotype information obtained from WBCs to identify paternally inherited SNPs. This method models a set of hypotheses that represent the different possible fetal genotypes (e.g., monosomy, disomy or trisomy) together with a specific probability. A maximum likelihood estimation analysis selects the most likely hypothesis and calculates the probability of that hypothesis being correct[135,136]. The FF is thus obtained from these estimations. Although this method is a direct and accurate way to assess the FF and generally considered as a gold standard, the total costs of the test are increased as at least two samples must be sequenced and analyzed. One of the commercially available tests that uses this approach is the Panorama test by Natera.

1.9 NIPT limitations: sources of discordant results

In order to better understand NIPT's main limitations, the major sources of erroneous results in NIPT analyses are described.

1.9.1 Sources of false positive results in NIPT

False positive results in NIPT are much more common than false negative results (88 vs. 12%)[147]. Although the consequences of a false positive result are less critical than a false negative result (which could result in the birth of a trisomic baby), a false positive still poses risks and unnecessary stress to the mother as it warrants the performance of the invasive confirmatory procedures it was designed to avoid, such as CVS or amniocentesis. The main sources of false positive results are confined placental mosaicism and vanishing twins.

a) Confined placental mosaicism

In most pregnancies, placental and fetal cells share the same genetic information, which is, in fact, the basis of NIPT. However, in a small proportion of cases, a mutation or chromosomal alteration can occur after the point at which the cells destined to become the fetus have separated from the cells destined to become the placenta. When this genetic discordance between the fetus and the placenta occurs, it is termed 'confined placental mosaicism' or 'confined fetal mosaicism' depending on the location of the mosaic cells[72]. A false positive NIPT result would be obtained by the presence of an aneuploid placenta and an euploid fetus[148].

b) Vanishing twins

A vanishing twin (VT) is a fetus in a multigestation pregnancy that dies in the uterus and is then partially or completely reabsorbed, either by the mother, the surviving twin or the placenta. In VTs, chromosomal abnormalities are common and are, in fact, important contributors of false positive results in NIPT counting methods (15%)[124].

Other conditions that can lead to the reporting of a false positive result include maternal chromosomal and subchromosomal abnormalities[149], maternal cancer[148,150], previous organ or bone marrow transplant from male donor[149], and certain maternal medical conditions or treatments[149].

1.9.2 Sources of false negative results in NIPT

In NIPT, false negative results can be mainly caused by two non-mutually exclusive situations: confined fetal mosaicism and low FF[119].

a) Confined fetal mosaicism

As explained above, confined fetal mosaicism refers to the situation that takes place when there is an euploid placenta and an aneuploid fetus, which would lead to a false negative result and the potential birth of a baby with chromosomal anomalies.

b) Low fetal fraction

Another reason that can cause a false negative result in NIPT tests, if not measured, is the FF. As stated in the previous section, the FF refers to the proportion of the total cfDNA molecules in the maternal circulation that originate from the placenta rather than the mother, expressed as a percentage[119]. When the FF is low enough, it becomes difficult or impossible to accurately detect chromosomal alterations or to determine the fetal sex, which can ultimately lead to newborns with unexpected congenital abnormalities. Therefore, determining the FF is paramount to interpret clinical assessments and determining the overall performance of NIPT[76,151]. In fact, in 2016, the American College of Medical Genetics and Genomics stated that it should be reported in all NIPT tests from then on[152]. Current NIPT tests generally require a minimal FF of 3 to 4% for a reportable result[153]. Below these values, a “no call” result will be reported, and a new blood sample has to be obtained.

Several situations or conditions are associated to a low FF. First and foremost, the FF is generally low during the first weeks of a normal pregnancy and tends to increase as the pregnancy progresses[43,113]. Other known determinants of a low FF include maternal body weight[152], aneuploidies of chromosomes 13 and 18[110], multiple gestation causing low FF per fetus[152], certain maternal medical conditions (thromboembolic disorders, heparin use, vitamin B12 deficiency, autoimmune disease) or treatments[149], and assisted reproduction[152].

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

2. Objectives

This Thesis has been developed in the context of an industrial PhD program, and more specifically, in a context in which genetic analyses are routinely performed to provide diagnostic support and genetic counseling for several conditions mainly in the pediatric and cancer fields.

Therefore, the principal objective has been to develop a set of tools to be able to work with cfDNA and to extract the biological information contained in it in a way that allows a further robust clinical application. Two main applications have been considered, cancer and NIPT, although only the latter is presented in this Thesis.

Therefore, the specific objectives of this work are:

a) To identify the main experimental challenges associated to working with cfDNA (mainly low amount and quality) to be able to optimize existing procedures in the laboratory and develop new strategies that allow to obtain quality samples for downstream analyses.

b) To identify relevant biological information contained in cfDNA for different specific applications and develop cost-effective approaches to be able to extract it and analyze it considering the advantages and disadvantages of the existing strategies in the market.

c) To test the performance of our developed approach in the context of NIPT both from a biological and economical point of view.

3. Methods

In the following section I will describe the different sample sources (**Section 3.1**), the optimization of the experimental steps (**Section 3.2**) and the bioinformatic pipelines used for data processing and analysis (**Sections 3.3** and **3.4** respectively).

3.1 Sample sources

Pregnant women attending the Hospital Universitari Dexeus (HUD - Grup Quirónsalut) for routine obstetrics care were recruited in the context of the pilot project *Evaluation of an alternative method for the detection of chromosome aneuploidies in cell-free fetal DNA by massive parallel sequencing*, which was approved by the ethics committee of Grup Hospitalari Quirón (protocol code qGEN-DEX-2017-09).

Potential participants in the study were provided with information on the aims, methods, benefits and potential risks of the study as well as information on personal data confidentiality, access, rectification, cancellation and opposition rights (**Supplementary Document 1**). Written informed consent was obtained from all participants prior to blood draw (**Supplementary Document 2**). For the samples included in the pilot study (batch 1), potentially relevant data regarding maternal habits, maternal body mass index (BMI), maternal diseases or plasma characteristics was not available. In addition, matched gDNA extracted from blood cells was not available either. All relevant information was available for samples in batch 2.

Inclusion criteria

Singleton pregnancies from approximately 11 to almost 17 weeks gestation.

Patient classification

- **Cases**: pregnancies at high risk (>1 in 270) for fetal genetic disorders in the first trimester combined screening (i.e., increased NT in the ultrasound scan and abnormal results in the biochemical screening tests).

- **Controls**: pregnancies at low risk (1 in 1/271 to <1 in 1100) in first trimester combined screening. Controls were confirmed through second trimester screening and physical exploration at birth.

Following the standard diagnostic procedures, controls were offered the noninvasive prenatal test Panorama (Natera) at HUD, while potential cases were directly validated by both quantitative fluorescent PCR (QF-PCR) and karyotype of chorionic villi cells. Because CVS involves some tissue damage that could potentially influence the FF in the maternal circulation, blood samples were drawn after first trimester screening and before CVS was performed.

Known or suspect factors susceptible of having an impact on the FF and thus on the analysis outcome such as maternal age, gestation age, BMI, daily habits as smoking or sport practice and maternal diseases were recorded.

Exclusion criteria

Potential candidates suffering from conditions such as malignancy, hepatocellular damage, trauma, inflammation, obesity or autoimmune diseases were excluded from the study, due to the potential distorting effects that these conditions can have over the levels of cfDNA[115].

3.2 Optimization of laboratory procedures

An outline of the devised experimental workflow can be found in **Figure 11.**

[redacted]

3.2.1 Plasma separation

In the context of NIPT, it is key to keep the FF as constant and high as possible throughout the blood collection and plasma preparation process[167]. This requires robust collection devices and blood processing protocols that prevent an increase of maternal genomic DNA in the collection tube due to WBC lysis. In this context, peripheral blood samples were collected in 10 mL Cell-Free DNA^{TM} blood collection tubes (Streck BCTs) and shipped to qGenomics at room temperature (RT; 15 ºC to 25 ºC) taking advantage of the already established sample collection circuit used for routine samples obtained from HUD. Streck BCTs are pre-coated with cell-preserving reagents to prevent WBC lysis and inhibit nuclease-mediated DNA degradation up to 14 days when stored at RT according to the manufacturer[168] and up to 7 days at RT according to independent recent studies[167].

Plasma was isolated within 72 hours of blood draw by two rounds of centrifugation according to the manufacturer's protocol. Briefly, a first low-speed centrifugation (1,600 ×g for 15 minutes at 16 ºC) was carried out to fractionate blood cells from plasma, which was then carefully transferred to 2 mL fresh tubes without disturbing the buffy coat layer. A second high-speed centrifugation at 16,000 ×g for 10 minutes at 4 ºC was performed to pellet residual debris, and the resultant plasma was transferred to 1.5 mL tubes without disturbing the residual cell pellet and stored at ≤ −80 ºC.

[redacted]

3.2.3 Isolation of cfDNA and quality controls

CfDNA was purified using the Maxwell RSC® ccfDNA Plasma Kit (Promega) in a Maxwell® RSC instrument following the manufacturer's instructions. The Maxwell RSC® ccfDNA Plasma kit is an automated magnetic bead-based method showing one of the highest isolation efficiencies among the commercially available technologies[170]. For each cfDNA extraction, a cartridge containing 8 wells was placed in the Maxwell® RSC Deck Tray with a plunger in well number 8. A 0.5 mL elution tube with 40 µL of elution buffer (EB; 10 mM Tris-Cl, pH = 8.5) was placed into the elution tube position. Up to 1 mL of plasma was added to well number 1 before running the method. The automated Maxwell RSC® ccfDNA Plasma protocol involves an initial step of cfDNA binding to magnetic beads in the presence of detergents and chaotropic agents (guanidine thiocyanate and guanidine hydrochloride), followed by several wash steps with ethanol and isopropanol, and a final elution step with EB. Typically, 30 to 32 µL of EB with the extracted DNA were recovered after running the method.

Fluorometric and electrophoretic methods were used to determine the concentration and size distribution respectively of the extracted cfDNA using 2 µL of sample, following manufacturer's instructions. More specifically, the concentration of cfDNA was determined using the Qubit™ dsDNA HS (High Sensitivity) assay on a Qubit® 3.0 fluorometer instrument (ThermoFisher Scientific) and its size distribution was assessed using the Agilent High Sensitivity D1000 assay on a 2200 TapeStation instrument (Agilent Technologies).

3.2.4 Isolation of gDNA

A volume of 1x phosphate-buffered saline (PBS; Sigma-Aldrich Merck) equal to the extracted volume of plasma was added to the remaining cellular fraction (WBCs and RBCs) in the Streck BCT in order to reconstitute the initial blood's cellular density. Blood was mixed using a Pasteur pipette to homogenize the 1x PBS, the buffy coat and RBCs layers. Isolation of gDNA was done using the Maxwell® RSC Blood DNA Kit following manufacturer's instructions. This gDNA extraction method follows, with some minor changes, the same basic steps in the Maxwell RSC® ccfDNA Plasma protocol (explained in **Section 3.2.3**). Briefly, 300 µL of blood were mixed with 300 µL of lysis buffer and 30 µL of proteinase K and incubated for 20 minutes at 56 ºC shaking at 800 rpm in a Biometra S1 ThermoShaker (Analytik Jena AG). Each DNA extraction required loading a total of 630 µL in well number 1. Samples were eluted in 50 µL of EB, from which 40 µL were typically recovered.

Similar to cfDNA, the concentration and size distribution of the extracted gDNA was determined by fluorometric and electrophoretic methods. However, in this case, the Qubit™ dsDNA BR (Broad Range) and the Agilent D1000 assays were used on a Qubit® 3.0 fluorometer instrument (ThermoFisher Scientific) and on a 2200 TapeStation instrument (Agilent Technologies), respectively.

3.2.5 gDNA fragmentation

The first main step in preparing gDNA for NGS with an Illumina platform is fragmentation which, as explained in the **Introduction** chapter, is typically done by either physical methods (e.g., acoustic shearing and sonication) or random enzymatic digestion (e.g., non-specific endonuclease cocktails and transposase tagmentation reactions)[15].

A total of 1.1 µg gDNA was enzymatically sheared using the NEBNext® dsDNA Fragmentase® (New England Biolabs) following manufacturer's instructions. Briefly, NEBNext dsDNA Fragmentase was thawed on ice immediately prior to use, vortexed for 3 seconds, and kept on ice. A reaction mix for n+1 reactions was prepared by combining 2 µL of 10x Fragmentase Reaction Buffer with 2 µL of NEBNext dsDNA Fragmentase per reaction. Hence for each reaction, 4 µL of the reaction mix were distributed per well on a 96-well plate and kept on ice. In addition, for each reaction, 16 µL of DNA (a total of 1.1 µg) were also distributed to a 96-well plate and kept on ice. A multichannel pipette was then used to transfer each DNA sample to a well containing the reaction mix avoiding bubble formation. This step is crucial to avoid differences in the incubation time between samples. The plate was sealed with adhesive film, vortexed for 3 seconds and briefly centrifuged. Samples were incubated from 30 to 42 minutes at 37 ºC in the TruTempTM DNA Microheating System (Robbins Scientific Corporation) to generate fragments centered around 230 to 260 bp. The reaction was stopped with 5 µL of 0.5 M EDTA (Sigma-Aldrich Merck) and DNA was purified with 60 µL of Agencourt AMPure XP Beads (Beckman Coulter) at a 1.2 beads-to-DNA volume ratio to maximize the recovery of smaller DNA fragments. DNA was allowed to bind the beads at RT for 15 minutes and the tubes were then placed on a magnetic stand for 10 minutes. After two freshly prepared ethanol 80% washes, DNA was eluted in 45 µL of EB, from which 40 µL were recovered. The remaining 5 µL were used to assess both the DNA concentration and fragmentation profiles by using Qubit dsDNA HS assay and TapeStation HS D1000, respectively.

When working with cfDNA, this step was skipped since cfDNA is already fragmented into molecules of around 185 bp (range =160-700 bp).

3.2.6 DNA library preparation

Both cfDNA and gDNA Illumina-compatible libraries were prepared with the NextFlex® Rapid DNA Sequencing Kit (Bioo Scientific). This kit uses the conventional double-stranded library construction method (see **Section 1.2.1** in the **Introduction** chapter).

NGS libraries were prepared according to the manufacturer's protocol with slight modifications considering the challenges and consequences (mainly low library complexity and sequence bias) of working with low DNA inputs.

[redacted] The reactions for each library preparation reaction, which are outlined in **Figure 2** in the **Introduction** chapter), are described in detail below:

a) End-repair and adenylation

The DNA end-repair and adenylation (phosphorylation of the 5' ends and A-tailing of the 3' ends) steps were done by mixing in a PCR plate 32 µL of DNA (be it cfDNA or gDNA) with 15 µL of the NEXTflex™ End-Repair & Adenylation Buffer and 3 µL of the NEXTflex™ End-Repair & Adenylation Mix. The reaction was performed in a SimpliAmp™ Thermal Cycler (Applied Biosystems) with the following thermal profile: incubation at 22 ºC for 20 minutes followed by incubation at 72 ºC for 20 minutes with a final 4 ºC incubation step. The lid was set at 105 ºC throughout the procedure.

b) Adapter ligation

The adapter ligation reaction was performed by adding 2.5 µL of NEXTflex™ barcoded adapters and 47.5 µL of NEXTflex™ Ligase Enzyme Mix to the 50 µL of end-repaired and adenylated DNA obtained in the previous step. At this point, the adapter-to-DNA molar ratio is key: while an excess of adapters versus DNA favors the formation of adapter dimers (which can be difficult to completely remove during purification and thus can affect subsequent PCR steps), too little amount of adapters versus DNA will reduce the efficiency of the ligation step. Thus, adapter titration experiments were performed in order to find the optimal adapter-to-DNA molar ratio.

c) Ligation purification

The ligation reaction was purified by performing two consecutive rounds of bead clean-ups using the Agencourt AMPure XP Beads (Beckman Coulter). [redacted] Briefly, 100 µL of ligated DNA was mixed with 60 µL of beads and vortexed to ensure thorough mixing of the solution and binding of the DNA to the magnetic beads. Samples were incubated at RT for 5 minutes on the bench and 5 minutes on a magnetic stand. After the incubation on the magnet, the supernatant (SN) was carefully removed without disturbing the beads pellet. Then the beads were washed by adding 200 µL of freshly prepared 80% ethanol. The plate was incubated at RT for 30 seconds, after which the 80% ethanol was carefully removed. After a second 80% ethanol wash, the plate was briefly spun on a Tabletop centrifuge (VWR™ Galaxy Centrifuges) and placed on the magnet again to allow the removal any residual 80% ethanol. The plate was left uncovered on the bench for 1 to 3 minutes until the bead pellet was visibly dry. In order to release the DNA from the beads, 52 µL of EB were added to the dried beads, vortexed until homogenized and incubated at RT for 5 minutes on the bench. The plate was then placed on the magnetic stand for 5 minutes and the SN was transferred to a new well. A second bead clean-up was performed in which 50 µL of beads were added to the resulting 50 µL of sample. The final pellet was eluted in 22 µL of EB, from which 20 µL were transferred to a new plate for PCR amplification.

d) Determination of PCR amplification cycles

[redacted]

[redacted]

[redacted]

[redacted]			
[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	

[redacted]			
[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]	

e) PCR amplification

For PCR amplification, 17.5 µL of adapter-ligated template were mixed with 2 µL of NEXTflex™ Primer Mix, 12 µL of NEXTflex™ PCR Master Mix, and nuclease-free water to a final volume of 50 µL in a PCR plate an run in a thermocycler with the same thermal profile used in the real-time qPCR experiment (**Table 7**) except for the melting step, which was substituted for a final 5 minute extension at 72 ºC. Samples were amplified for as many cycles as determined by qPCR. The PCR reaction was purified using a single bead clean-up [redacted]. Finally, amplified libraries were eluted in 20 µL of EB.

Qubit™ dsDNA HS (High Sensitivity) and Agilent High Sensitivity D1000 assay were used to determine the concentration and size distribution respectively of the final libraries using 2 µL of samples, following manufacturer's instructions.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted] A total of 7.5 µL of 2x Hybridization Buffer and 3 µL of Hybridization component A from the SeqCap EZ Hybridization and Wash Kit (Roche Sequencing) were added.

The hybridization procedure involved i) DNA denaturation at 95 ºC for 10 minutes, after which [redacted] and nuclease-free water up to 15 µL were added, and ii) DNA-probe annealing through [redacted] on a SimpliAmp™ Thermal Cycler equipment (Applied Biosystems) with the lid set to [redacted] The hybridization reactions were purified using the SeqCap EZ Hybridization and Wash Kit (Roche Sequencing) together with the magnetic streptavidin beads Dynabeads™ M-270 Streptavidin (ThermoFisher Scientific) following manufacturer's instructions. The streptavidin beads were tempered at RT for 30 minutes and resuspended in 100 µL of 1x Bead Wash Buffer after two rounds of wash, after which they were transferred to a 0.2 mL PCR tube placed in a magnetic stand. The SN was discarded, and the beads were mixed and vortexed briefly with the hybridization reaction to allow binding of the biotinylated heteroduplexes to the streptavidin beads. The reaction was incubated for 45 minutes at [redacted] in a thermocycler with brief vortex every 15 minutes. After the incubation, the unbound molecules were washed off with a series of more stringent washes as follows: 100 µL of 1x Wash Buffer I (preheated at [redacted] were added to the reaction, which was briefly vortexed, centrifuged and incubated on a thermocycler at [redacted] for 5 minutes. The reaction was then placed on a magnetic stand and the SN was quickly discarded once the solution was completely clear to avoid any significant temperature loss. This step was performed twice and repeated with 200 µL of Stringent Wash Buffer preheated at [redacted] Next, 200 µL of 1x Wash Buffer I at RT were added. The sample was vortexed for 2 minutes, placed on the magnetic stand and SN was discarded once the solution was completely clear. The same procedure was carried out with Wash Buffer II and Wash Buffer III with vortex for 1 minute and 30 seconds respectively. Finally, enriched libraries were eluted in 30 µL of EB.

The enriched library pool was PCR amplified, purified and quantified as described in **Section 3.2.5**. However, at this point, beads were not discarded as the biotinylated library heteroduplexes were still bound to the streptavidin moiety of the beads. It is not until libraries are denatured at 95 ºC during the PCR amplification that the library molecules are released from the beads.

3.2.9 Illumina next-generation sequencing

The [redacted] library pools were loaded onto a NextSeq™ 500 Sequencing System (Illumina) following the standard normalization method described by the manufacturer. Briefly, libraries were diluted to 4 nM with the appropriate volume of EB and denatured for 5 minutes at RT with freshly prepared 0.2 N NaOH (Merck®). After denaturation, libraries were diluted to the 1.8 pM loading concentration with prechilled hybridization buffer HT1 and kept on ice until they were loaded. As a sequencing control, a denatured, 1.8 pM PhiX control v3 library (Illumina) was added at 1% (v/v) to the final pool. Samples were sequenced using 2 x 150 bp paired-end protocol. The instrument was set to read 6 bp indices.

3.2.10 Identification of fetal sex by real-time qPCR

The DYZ1 region, which consists of a 3.4 Kb array of 2000 to 4000 of highly repetitive sequences located on the long arm of chromosome Y[176,177], is of great interest for forensic sex determination assays of highly degraded or minute samples[178]. In the context of a pregnant woman, the detection of the DYZ1 region in a pool of cfDNA molecules would distinctly indicate the presence of a male fetus, while its absence would be compatible with the presence of a female.

However, the absence of DYZ1 amplification could also be due to insufficient cfDNA (PCR failure) or insufficient cffDNA template in the sample due to low FF, which would lead to an incorrect fetal sex determination. In order to control for potential PCR failure due to insufficient or low quality cfDNA, we simultaneously assayed an autosomal region with a pair of primers targeting the HBD and HBB genes in the β-globin locus (chr11:5,248,147-5,248,248 and chr11:5,255,559-5,255,660 respectively, hg19 human genome reference build).

Primers for the DYZ1 region (forward 5'-TCCTGCTTATCCAAATTCACCAT-3', reverse 5'-ACTTCCCTCTGACATTACCTGATAATTG-3') were designed as described by Wataganara et al.[179], and β-globin primers were designed as described by Vasavda et al.[180] (forward 5'-GTGCACCTGACTCCTGAGGAGA-3', reverse 5'-CCTTGATACCAACCTGCCCAG-3').

Real-time qPCR was first performed in serial dilutions of gDNA samples in a QuantStudio® 12K Flex Real-Time PCR System (ThermoFisher Scientific). A reaction mix was prepared as follows: 5 µL of the Power SYBR™ Green PCR Master Mix kit (Applied BiosystemsTM) were mixed with 2 µL of forward and reverse primers (1 µM final concentration), 2 µL of nuclease-free water and 1 µL of template DNA to a final volume of 10 µL. Samples were amplified with the following thermal profile: 95 ºC for 10 minutes followed by cycles of denaturation at 95 ºC for 15 seconds and annealing and extension at 60 ºC for 1 minute. The quantification cycle (Cq) differences between DYZ1 and β-globin were used to relatively quantify them using the following the formula (**eq. 1**). The specificity of the amplified product was evaluated through electrophoresis on an 1.3% agarose gel (w/v).

$$\frac{\text{DYZ1}}{\beta\ \text{globin}} = 2^{\left(\text{Cq}\ _{\beta\text{globin}}\ -\ \text{Cq}\ _{\text{DYZ1}}\right)} \qquad (1)$$

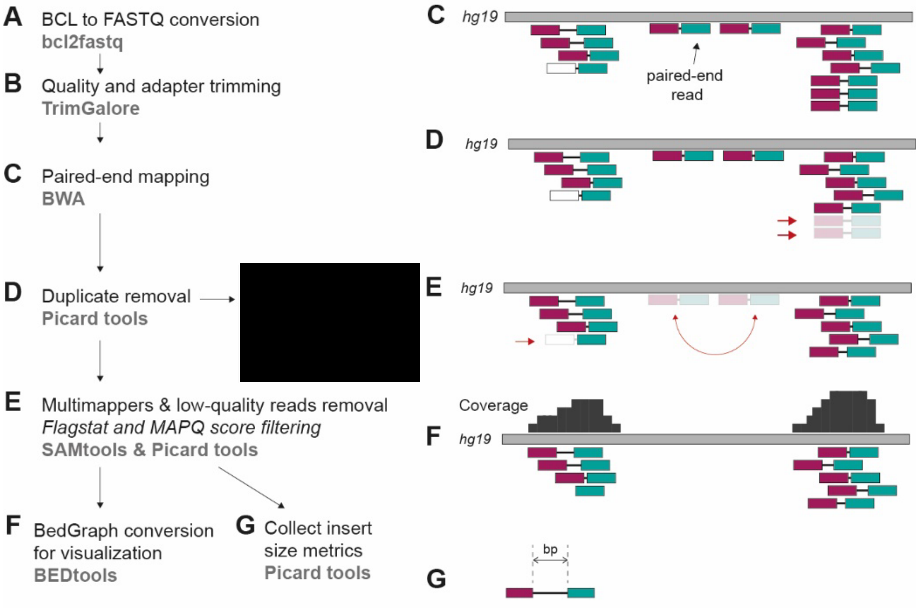
3.3 Bioinformatic data processing

The data processing steps, involving read pre-processing and [redacted] calculation are described below. All bioinformatic procedures were performed at qGenomics using custom bash shell scripts.

3.3.1 Read pre-processing

Raw sequencing data in the binary base call (bcl) format generated by the NextSeq 500 Sequencing System were demultiplexed and converted to standard FASTQ files with the Illumina bcl2fastq2 Conversion Software v2.20[181] (**Figure 13A**). The unmapped read sequences in FASTQ format were trimmed using TrimGalore software (version 0.4.4)[182] (**Figure 13B**). First, low-quality base calls were trimmed off the 3' end of the reads (quality trimming) followed by adapter removal from the 3' ends of reads (adapter trimming). Trimmed reads were aligned to the reference human genome, build 37 (GRCh37/hg19)[183] using the Burrows-Wheeler Aligner's (BWA version 0.7.5a) maximal exact matches (BWA-MEM) algorithm[184,185] with standard parameters (**Figure 13C**). The generated sequence alignment map (SAM) file was then sorted (Picard tools [v2.14.1][186]) and saved as a binary alignment map (BAM) file, which was then subject to marking and removal of duplicated reads (Picard tools v2.14.1) (**Figure 13D**). [redacted] In addition, reads with a mapping quality value (MAPQ) score below 20 (which corresponds to >1% probability of being wrongly mapped and includes multimapper reads) and a flagstat below 256 were also removed using a combination of SAMtools (v0.1.19)[187] and Picard software tools (**Figure 13E**). The Picard function CollectInsertSizeMetrics was used to calculate the average insert size from BAM files (**Figure 13G**).

BAM alignment files were then converted to bedGraph track files using BEDtools[188] (v2.25.0) to allow visualization of sequencing coverage using the WashU Epigenome Browser (v48.5.0)[189,190] (**Figure 13D**).


A BCL to FASTQ conversion
bcl2fastq
B Quality and adapter trimming
TrimGalore
C Paired-end mapping
BWA
D Duplicate removal
Picard tools
E Multimappers & low-quality reads removal
Flagstat and MAPQ score filtering
SAMtools & Picard tools
F BedGraph conversion for visualization
BEDtools
G Collect insert size metrics
Picard tools
C
hg19
paired-end read
D
hg19
E
hg19
Coverage
F
hg19
G
bp

[redacted]

Figure 13. **Summary of read pre-processing**. **A**. BCL files generated by the sequencing equipment are converted to FASTQ files to allow read processing. **B**. Quality and adapter trimming are followed by I mapping of reads to the reference hg19 genome. **D**. Next, duplicated reads, I multiple mappers and low-quality reads are filtered out. At this point, files are ready for further analysis, as for example, for (**F**) data visualization or (**G**) insert size metrics collection. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

3.3.3 Data matrix building

[redacted] The final matrix was imported to R[193] to perform downstream analyses concerning relative coverage changes among samples such as fetal sex determination, trisomy detection or FF calculations.

3.4 Bioinformatic data analysis

[redacted]

[redacted]

[redacted] [redacted]

[redacted]

[redacted] [redacted]

[redacted]

The distance to TSS for each interval in the sample was annotated using the Homer software package. The average coverage as a function of the distance to TSS was calculated using a set of windows. The window size and hence the number of windows used was dependent on the total distance upstream and downstream TSS that was being evaluated: for a distance of ±1.73 Mb, ±150 Kb and ±3 Kb, the window size was set at 500 bp, 100 bp and 25 bp respectively.

[redacted]

The DNase signal from blood primary cells in each genomic interval in the matrix was annotated using Homer software. DNase data was obtained from the ENCODE databse[194,195] and signal was averaged from all samples downloaded. A summary of the downloaded data can be found in **Table 9**.

Table 9. **Blood primary cells DNase datasets downloaded from the ENCODE database**. The specific identifiers for each downloaded sample are provided.

Biosample summary	**ENCODE identifiers**
B cell female adult (34 years)	ENCFF469VSO
B cell male adult (37 years)	ENCFF203BEH
CD14-positive monocyte female adult (34 years)	ENCFF175LYA
CD14-positive monocyte male adult (37 years)	ENCFF418HYD
CD4-positive, alpha-beta T cell female adult (33 years)	ENCFF938VVP
CD4-positive, alpha-beta T cell male adult (37 years)	ENCFF162GBJ
CD8-positive, alpha-beta T cell female adult (34 years)	ENCFF736QAD
CD8-positive, alpha-beta T cell male adult (37 years)	ENCFF154RQQ
Common myeloid progenitor, CD34-positive male adult (36 years)	ENCFF800ZLW
Common myeloid progenitor, CD34-positive female adult (33 years)	ENCFF352KRQ

3.4.2 Detection of chromosomal aneuploidies

The identification of chromosomal aneuploidies was assessed by comparing differences in the [redacted] sequencing read density of each chromosome between control and test samples, which was calculated as described above (**eq. 2** and **3**).

Then, the expected normal variation for each chromosome was defined by calculating the mean (**eq. 4**) and standard deviation (**eq. 5**) for each chromosome in the control population of samples.

[redacted]

(4)

(5)

Then the z-score (z-test) for each chromosome in test samples (that is, potentially aneuploid samples) was calculated to assess how many standard deviations each chromosome in the test sample deviates from the baseline normal average (**eq. 6** and **Figure 15E**). [redacted]

[redacted]

[redacted] (6)

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

3.4.3 Fetal sex discrimination through relative coverage differences in chrY

For chrY, fetal sex determination was first performed by comparing raw chrY read counts in male and female-bearing pregnancies using custom-designed scripts with the R software. More specifically, the percentage of reads mapping to chrY (%chrY) (**eq. 7**) was calculated as:

$$\%\mathrm{chrY} = \frac{\sum \mathrm{counts}_{\mathrm{chrY}}}{\sum \mathrm{counts}_{\mathrm{sample}}} \cdot 100 \qquad (7)$$

Previous knowledge of fetal sex information was used to exclude any Y-specific regions present in 2 or more female-bearing pregnancies (regardless of its presence in male-bearing pregnancies). The percentage of reads after chrY filtering ($\%\mathrm{chrY}_{\mathrm{filt}}$) (**eq. 8**) was then calculated as:

$$\%\mathrm{chrY}_{\mathrm{filt}} = \frac{\sum \text{counts}_{\ \mathrm{chrY}_{\mathrm{filt}}}}{\sum \text{counts}_{\ \mathrm{sample}}} \cdot 100 \qquad (8)$$

Heatmaps were generated in R to visually compare filtered and unfiltered %chrY signal in male and female-bearing pregnancies.

The R software was also used to apply the unpaired two-samples Wilcoxon test (alternative hypothesis: true location shift is less than 0) to test whether the ratio differences between males and females was statistically significant when using either unfiltered or filtered chrY.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

3.4.5 Statistical parameters

The results of our data analysis approaches to the detection of fetal aneuploidies and determination of fetal sex were compared to actual Panorama NIPT test results (aneuploidies) and fetal sex as provided by HUD. To this end, contingency tables were generated (**Table 10**).

Table 10. **Contingency table for the calculation of sensitivity, specificity and positive predictive value**.

	True positive	True negative	Total
Test positive	A	C	A + C
Test negative	B	D	B + D
Total	A + B	C + D	A + B + C + D

From the contingency table, several parameters that are of interest in this study can be calculated. These include the sensitivity (often referred also as detection rate [DR]; **eq. 9**), the specificity (**eq. 10**) and the positive predictive value (PPV; **eq. 11**).

$$\text{Sensitivity} = \frac{\text{A}}{\text{A} + \text{B}} \tag{9}$$

$$\text{Specificity} = \frac{\text{D}}{\text{C} + \text{D}} \tag{10}$$

$$\text{PPV} = \frac{\text{A}}{\text{A} + \text{C}} \tag{11}$$

3.4.6 Fetal fraction determination

[redacted]

[redacted] (12)

[redacted]

[redacted]

[redacted] (13.1)

(13.2)

[redacted]

[redacted]

[redacted]

Heatmaps for data visualization were constructed using the R software. Values were scaled by row for visualization purposes.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted] (14.1)

[redacted]

(14.2)

(14.3)

(14.4)

[redacted]

[redacted] (15)

[redacted]

[redacted] (15)

[redacted]

[redacted]

[redacted]

All scripts for data processing and analysis are available upon request.

4. Results

4.1 Samples

Samples collected from HUD were shipped to qGenomics using the already established sample collection circuit used for routine samples. Full patient information for all samples is shown in **Supplementary Table 1A** and **1B** respectively. A summary is shown in **Table 11**. Karyotype results of confirmed cases NIPT1_S4, NIPT1_S5 and NIPT3_S3 are shown in **Supplementary Figure 2**. Karyotype images from sample AM14_S2 were not available.

Table 11. **Summarized patient information from collected samples**. Maternal age, maternal BMI and gestational age are expressed as average (minimum value–maximum value). In addition, gestational age is expressed as week.day. Maternal BMI for samples in batch 1 was not available (NA).

	Batch 1	**Batch 2**
Samples		
Controls	7	13
Aneuploidies	1	3
Total	8	16
Maternal age	40.38 (35–46)	36.75 (30–43)
Maternal BMI	NA	23.10 (17.99–32.05)
Gestational age	13.93 (12–17.5)	13.27 (11.5–16.6)
Pregnancy type		
Simple	6	16
Simple VT	2	0
Informed fetal sex		
XX	4	6
XY	4	10

4.2 Optimization of laboratory procedures

4.2.1 Plasma separation and cfDNA extraction

Typically, 6 to 8 mL of whole blood were collected from each pregnant woman in a single Streck BCT, which yielded an average of 4.15 mL of plasma (range 3.6–5 mL). The average time elapsed from blood draw to plasma separation at qGenomics was 1.38 days (range 0–2 days) for the first batch of samples and 1.25 days (range 0–3 days) for the second batch, well below the manufacturer's recommended limit of 14 days. Full sample information is shown in **Supplementary Table 2A** and **2B** respectively. A summary of the plasma parameters is shown in **Table 12**.

Table 12. **Summarized plasma information**. Days between blood collection and plasma separation, plasma, volume and [redacted] are shown. Values are expressed as average (minimum–maximum). NA indicates non available.

	Batch 1	Batch 2
Time elapsed between sample collection and plasma separation (days)	1.38 (0–2)	1.25 (0–3)
Plasma volume obtained (mL)	NA	4.15 (3.6–5)
[redacted]	[redacted]	[redacted]

None of the plasma samples showed evident signs of hemolysis as shown by the yellowish plasma coloration (see **Supplementary Tables 2A** and **2B** for plasma color codes and **Supplementary Figure 1** for reference color palette), which, in NIPT commercial tests, is deemed acceptable to proceed with downstream analyses. [redacted]

No correlation was observed between the days elapsed between sample collection and plasma separation and [redacted], which suggests that under the shipping and processing conditions described here no measurable hemolysis took place in plasma.

The extraction of cfDNA from 1 mL of maternal plasma from gestational ages 11.5 to 17.5 yielded an average of 4.92 ng/mL (range 1.99–11.1 ng/mL) for 22 out of the 24 samples collected. These values are well in agreement with previously published data[27]. Two samples yielded out-of-range (OOR) results, meaning their concentration was below the 10 pg/µL detection limit of the instrument (High sensitivity assay) which would correspond to less than 0.5 ng of cfDNA per mL of plasma.

Similarly, no correlation was observed between days elapsed between sample collection and plasma separation and the amount of DNA extracted (r = -0.10, P =

0.662) (**Figure 21B**), suggesting that WBC lysis, if present, did not increase during sample processing.

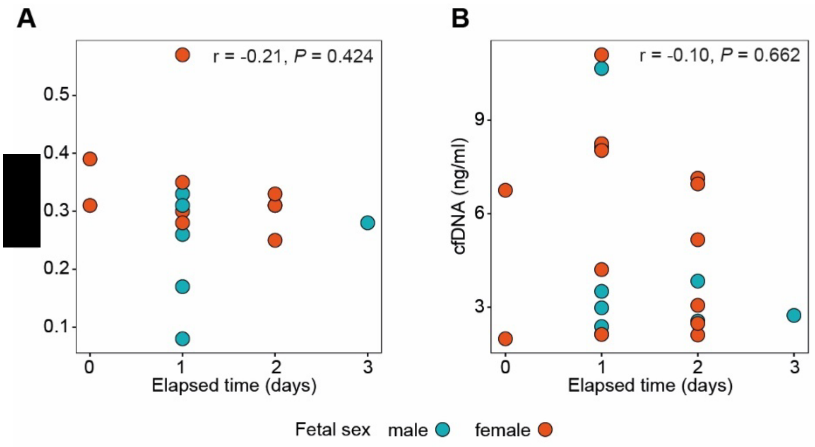


Figure 21. **Effect of plasma storage on** [redacted] **and cfDNA concentration.** No correlation is observed between time elapsed from blood draw until plasma separation (days) and (**A**) [redacted] (r = -0.21, *P* = 0.424) or (**B**) cfDNA concentration (r = -0.1, *P* = 0.662). Red and blue dots represent male and female-bearing pregnancies respectively.

Additional correlation analyses were performed regarding the amount of cfDNA extracted and diverse maternal conditions. In our set of samples, a weak negative correlation was observed between the amount of cfDNA extracted and maternal BMI (r = -0.29, *P* = 0.273) (**Figure 22A**), and no correlation was observed with gestational

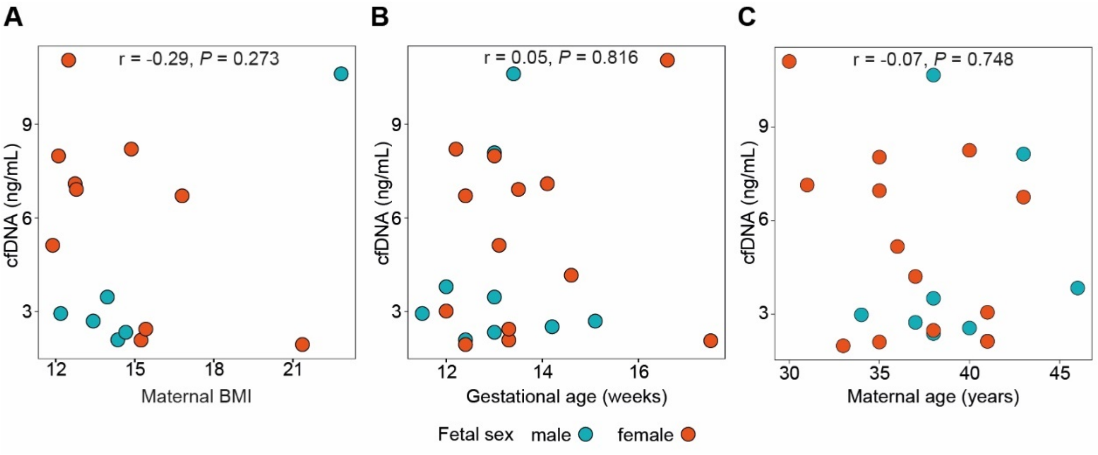


Figure 22. **Effects of maternal factors on the amount of cfDNA extracted**. **A**. A weak negative correlation is observed between maternal BMI and cfDNA concentration (r = -0.29, *P* = 0.273). **B** No correlation between gestational age and cfDNA concentration is observed (r = 0.05, *P* = 0.816). **C**. No correlation between maternal age and cfDNA concentration is observed (r = -0.07, *P* = 0.748).

age (r = 0.05, P = 0.816) (**Figure 22B**), or maternal age (r = -0.07, P = 0.748) (**Figure 22C**).

Taken together, the above results suggest that hemolysis, and therefore, WBC lysis, if present, were low within the timeframe in which samples were preprocessed.

4.2.2 cfDNA extraction quality controls

The assessment of the size distribution of the extracted cfDNA fragments showed a main peak centered at the mononucleosome plus linker DNA size range (166 to 237 bp) and a secondary peak at the dinucleosome plus linker DNA size range (332 to 474 bp) well in agreement with the cfDNA size range reported in the literature[20,126]. **Figure 23** shows the size distribution illustrative of an extracted cfDNA sample at 0.171 ng/µL (6.84 ng cfDNA per 1 mL of plasma) presenting a main peak at 185 bp and a secondary peak at 386 bp.

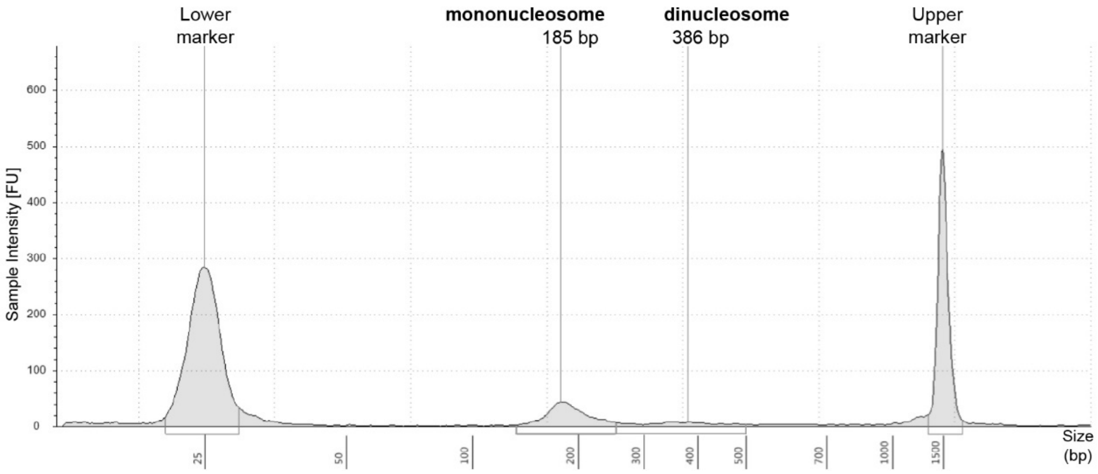


Figure 23. **Fragment size distribution of cfDNA fragments extracted from a maternal plasma sample at 0.171 ng/µL final concentration**. A main peak and a secondary peak around 185 bp (mononucleosome) and 387 bp (dinucleosome) respectively are observed. Upper (1.5 Kb) and lower (25 bp) molecular weight markers are run together with the sample to provide size reference.

The absence of high molecular weight DNA in the fragment size distribution profiles of the extracted cfDNA (**Figure 23**) provides additional proof that no evident hemolysis took place prior or during sample processing.

Then, we attempted to identify the presence of cffDNA in the maternal circulation as a clear evidence of an effective isolation of cfDNA from plasma. In this context, one of the simplest assays that can be performed is to detect, by quantitative PCR (qPCR), the presence of Y-derived molecules in male-bearing pregnancies.

The specificity of the DYZ1 and β-globin primers was first tested by *in silico* PCR. The β-globin primers yielded two amplicons of 102 bp each corresponding to the HBB and the HBD genes. The DYZ1 primers produced 12 and 4 amplicons of 85 bp and 2,462 bp each (**Supplementary Table 3**). Since cfDNA is highly fragmented, only the 85-bp products are expected to be generated. Consequently, a 3-fold ratio of DYZ1 (12 amplicons for a single chromosome) to β-globin (2 amplicons each chromosome 11, 4 amplicons total) was expected, which translates into an approximate Cq difference (ΔCq) of 1.5 when assayed by qPCR.

The above calculations were confirmed using 10-fold serial dilutions of a male reference gDNA from 8.67 ng/μL (measured with fluorometric methods) down to a theoretical 8.67 pg/μL (not measurable with fluorometric methods). Results show that both β-globin and DYZ1 could be linearly detected through all the dilution points and that the ΔCq was maintained through all the dilution points (ΔCq = 1.66 ± 0.1) (**Figure 24A**).

Next, real-time qPCR was performed in 3 maternal plasmas from women at gestation weeks 11.3, 14.3 and 11 bearing female (XX), female (XX) and male (XY) fetuses respectively. As expected, β-globin was amplified in all three samples before the blank control. On the other hand, DYZ1 was detected before the blank control only in the male sample. In female samples, amplification with DYZ1 primers was observed only after the blank control (**Figure 24B**). The specificity of the amplified products was confirmed by gel electrophoresis, which showed a 102-bp band (β-globin) in all samples and a 85-bp band (DYZ1) only in the male sample. Additional bands of a lower molecular weight observed in the β-globin wells which could be due to primer-dimer formation (**Figure 24D**).

The ΔCq between β-globin and DYZ1 for each sample was used to calculate the amount of relative chrY signal among samples. Results show that this signal in the male fetus is 60-fold higher than that of the reference female fetus (**Figure 24C**).

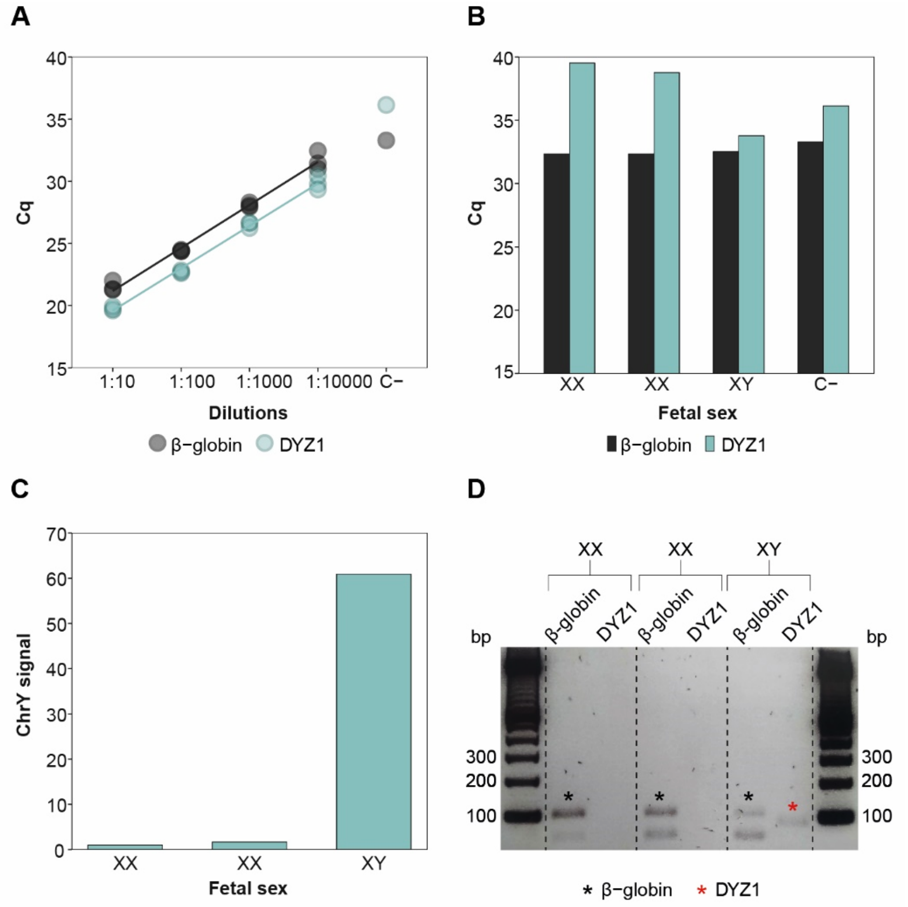


Figure 24. **Y-derived sequences detection in plasma**. **A**. Cq values for β-globin and DYZ1 at different point dilutions (1:10, 1:100, 1:1000 and 1:10 000) in a reference male gDNA sample. **B**. Cq values for β-globin and DYZ1 in cfDNA samples of two female and one male-bearing pregnancies. **C**. Relative amount of chrY-derived signal among cfDNA samples. **D**. Gel electrophoresis showing the specificity of the amplified products. Amplicons corresponding to β-globin (102 bp band, highlighted with a black asterisk) are observed in all samples. Amplicon corresponding to DYZ1 (85 bp band, red asterisk) is only observed in the male-bearing pregnancy. Additional bands of less than 85 bp are observed in all samples and may be attributed to primer-dimer formation.

Overall, these results show that we have successfully extracted cfDNA as well as cffDNA and that we can specifically detect the presence of male cffDNA in blood as early as 11 weeks by qPCR.

4.2.3 gDNA extraction and fragmentation

Regarding gDNA sample processing prior to library preparation, gDNA samples were extracted and enzymatically digested. As expected, the size distribution of fragmented gDNA is found between ~100 and 700 bp (**Figure 25**).

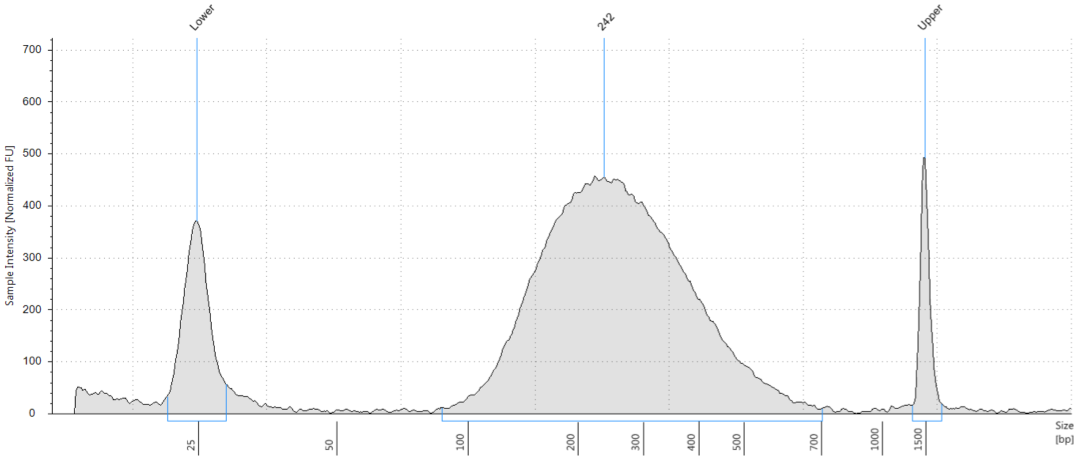


Figure 25. **Example of a fragmented gDNA sample**. Size distribution of fragments is centered at 242 bp (range ~100–700)**.**

4.3 Generation of low input libraries

The optimized generation of low input dsDNA libraries of both cfDNA and gDNA involved a series of modifications to the manufacturer's protocols that are detailed in the following Sections and summarized in **Table 13**.

4.3.1 Adapter molarity and PCR cycle optimization for low input DNA

In library construction, in order to maximize the number of ligated molecules, the amount of adapter with respect to the DNA input as well as the number of PCR cycles to perform have to be optimized.

a) Optimization of adapter molarity

If adapter molecules outnumber the DNA molecules to ligate, the formation of adapter dimer, this is, the ligation of two adapter molecules through their double-stranded ends, can be favored. The formation of adapter dimers during library preparation is unwanted because adapter dimers are the smallest molecules within the library that can be formed and as such, will be preferentially amplified and sequenced in the subsequent steps[15]. It is therefore critical to find the optimal adapter-to-DNA molar ratio that will maximize library formation while minimizing adapter dimer formation.

Therefore, we first aimed at evaluating manufacturer's guidelines of adapter molarity for inputs as low as 1 ng of gDNA and cfDNA (**Table 13**).

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Because adapter dimers are preferentially amplified during the PCR reaction, the total number of PCR cycles performed has an effect on the final adapter dimer to DNA library ratio. Therefore, the amount of adapter should be ideally optimized considering not only the DNA input, but also the number of cycles that will be applied to DNA libraries. In turn, the total number of cycles that will be applied to DNA libraries is dependent on the amount of ligated molecules available.

When amplifying ligated DNA libraries, PCR cycles must be kept to the minimum necessary to obtain enough amplified product and avoid PCR biases and errors. In the context of this work, in which the amount of cfDNA is determined by its amount in plasma, the amount of PCR cycles required needs to be adjusted for each sample (described below). Therefore, in our context, we considered the optimal adapter concentration as the one in which adapter dimer formation is avoided regardless of the number of PCR cycles.

[redacted]

[redacted]

b) Optimization of PCR cycles

In targeted sequencing experiments, PCR amplification cycles have to be performed to generate enough material to allow the precise quantification of the amplified product by fluorometric methods and generate enough template for hybridization. Importantly, PCR overamplification has to be avoided as this can introduce biases due to polymerase errors and sequence base composition. This is especially critical when working with low inputs as in the case of cfDNA.

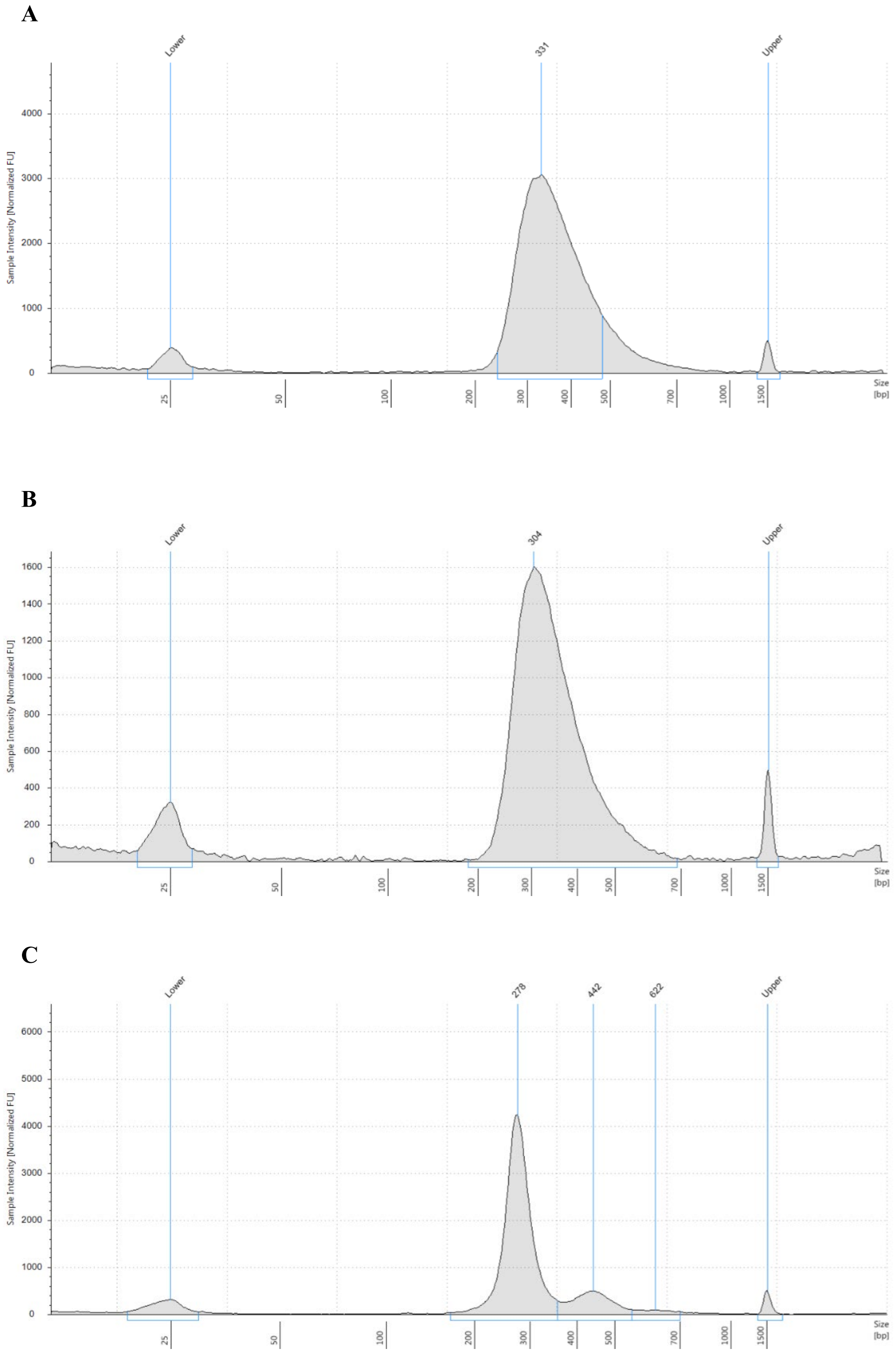
To find the optimal number of PCR cycles, qPCR was used. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]


A
Lower
331
Upper
Sample Intensity [Normalized FU]
4000
3000
2000
1000
0
25
50
100
200
300
400
500
700
1000
1500
Size [bp]
B
Lower
304
Upper
Sample Intensity [Normalized FU]
1600
1400
1200
1000
800
600
400
200
0
25
50
100
200
300
400
500
700
1000
1500
Size [bp]
C
Lower
278
442
622
Upper
Sample Intensity [Normalized FU]
6000
5000
4000
3000
2000
1000
0
25
50
100
200
300
400
500
700
1000
1500
Size [bp]

[redacted]

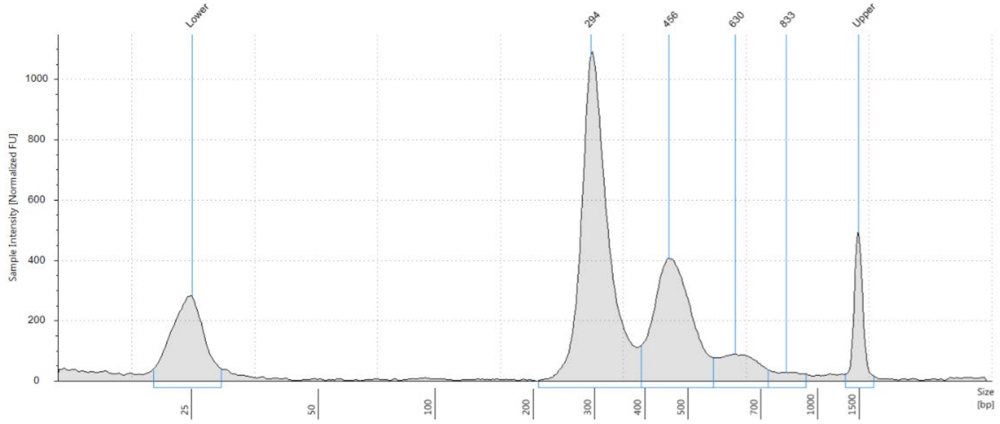


Figure 28. **Size distribution of cfDNA and gDNA libraries before and after solution hybridization**. **A**. Size distribution profile of a gDNA library at 11.3 ng/µL before solution hybridization. **B**. Size distribution profile of a final gDNA library pool at 4.89 ng/µL. **C**. Size distribution profile of a cfDNA library at 11.1 ng/µL before solution hybridization. **D**. Size distribution profile of a final cfDNA library pool at 3.93 ng/µL.

4.3.2 Optimization of the hybridization probe

Targeted sequencing through library hybridization is a procedure that is routinely performed at qGenomics and therefore, general conditions for this step were mainly obtained from qGenomics internal protocols. [redacted] Hence, we first evaluated i) the feasibility of hybridizing low DNA inputs using our custom hybridization probes and ii) the optimal amount of probe that should be used. [redacted] The hybridization and washing reactions were performed as described in the **Methods** chapter and the amount of recovered material after hybridization was quantified using real-time qPCR. [redacted]

[redacted]

[redacted]

[redacted]

Final gDNA and cfDNA library pools show the size distributions observed prior to hybridization. In the case of the gDNA pool, a continuous size distribution of fragments centered at ~304 bp is observed, of which ~120 bp correspond to adapter sequences and ~184 bp correspond to the fragmented DNA insert (**Figure 28B**).

Regarding cfDNA libraries, a main peak around 297 bp, of which ~120 bp correspond to adapter sequences and ~177 bp correspond to the nucleosomal DNA length plus a variable-length linker DNA can be observed. Additional peaks of higher molecular weight corresponding to di-, tri- and also tetra-nucleosomes are observed (**Figure 28D**).

4.4 Raw and pre-processed data quality metrics

A total of 8 unpaired cfDNA samples (batch 1) and 16 cfDNA and gDNA paired samples (batch 2) from pregnant women were sequenced on the NextSeq500 Illumina platform using 150-bp paired-end reads. A number of different metrics relative to the sequencing and processing parameters are summarized for each sample in **Supplementary Table 4**.

Because the first unpaired 8 cfDNA samples that were sequenced were the ones that had been used during the library preparation optimization and initial sequencing tests, the total initial number of reads obtained varies considerably among samples: an average [redacted] reads (range [redacted] reads) was obtained. The average percentage of duplication was [redacted] (range [redacted]). These values are considerably higher than the ones obtained in the subsequently sequenced cfDNA and gDNA paired samples. This high variability in some of the sequencing metrics of the first batch of sequenced samples can be explained by a technical error in 5 out of the 8 samples analyzed. The error occurred during the recovery of cfDNA after the hybridization step (AM14_S1, AM14_S2, AM3_S43, AM3_S46 and AM7_S1). More specifically, it was due to not including the streptavidin beads containing most of the sample in the PCR amplification step after hybridization (see the **Methods** chapter, **Section 3.2.7** for information on the detailed procedure of DNA recovery after hybridization and successive PCR amplification). Despite this error, which was solved in the subsequent analyses, these samples were included in downstream analyses, as the number of final reads was comparable to the samples that were processed properly. From the remaining non-duplicated reads, an average of [redacted] (range [redacted]) corresponded to unique sequences with a MAPQ score above 20 and a flagstat below

256, which yielded a final average of [redacted] unique reads (range [redacted] reads) per sample. This corresponded to an average of [redacted] (**Supplementary Table 4A**).

For paired cfDNA and gDNA samples, an average of [redacted](range [redacted]reads) and [redacted] (range [redacted] reads) total reads per sample was obtained respectively. An average of [redacted] (range [redacted]) and [redacted] (range [redacted]) of duplicated reads was obtained for each cfDNA and gDNA sample respectively, from which an average of [redacted] (range [redacted]) and [redacted] (range [redacted]) (cfDNA and gDNA samples respectively) of them corresponded to uniquely mapped reads to the hg19 human reference genome with a MAPQ score above 20 and a flagstat below 256. The final unique sequence reads per sample averaged [redacted] (range [redacted] reads) and [redacted] (range [redacted] reads) for cfDNA and gDNA samples respectively (**Supplementary Tables 4B** and **4C**). The average number of genomic regions captured in cfDNA and gDNA samples was [redacted] (range [redacted] regions) and [redacted] (range [redacted] regions) respectively. Regarding horizontal coverage, an average of [redacted] (range [redacted]) were covered by cfDNA samples and an average of [redacted] (range [redacted]) were covered by gDNA samples.

In final uniquely mapped reads, the median fragment size for cfDNA samples was very close to the size of the nucleosome-protected DNA. **Figure 30** shows the distribution of insert sizes for sample NIPT2_S5, with a median fragment size of ~164 bp. It's important to note that cfDNA samples, unlike gDNA samples, are not fragmented prior to sample preparation (see **Methods**). The sequence size of the matched, *in vitro* fragmented gDNA was a little higher, with a median value of 178 bp.

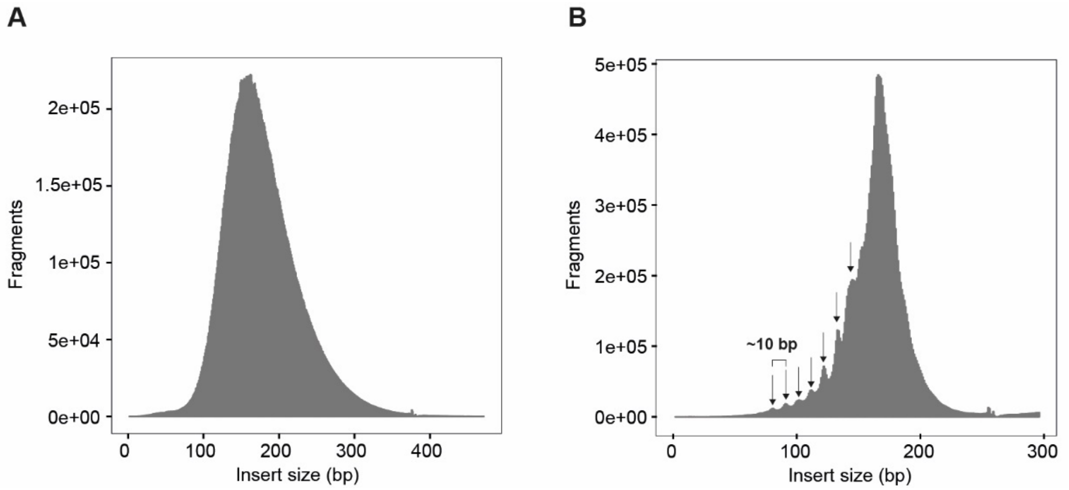


Figure 30. **Fragment length of gDNA and cfDNA libraries inferred from alignment of paired-end reads**. **A**. In gDNA libraries, the median insert size in gDNA libraries was found at ~178 bp. **B**. In cfDNA libraries, the median insert size was found at ~164 bp, consistent with the length of DNA associated to a nucleosome. Additional small peaks at ~10-bp periodicities are also observed, which have been attributed to the to the helical pitch of DNA on the nucleosome core[18].

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted].

[redacted]

[redacted]

[redacted]

4.6 Data matrix building

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

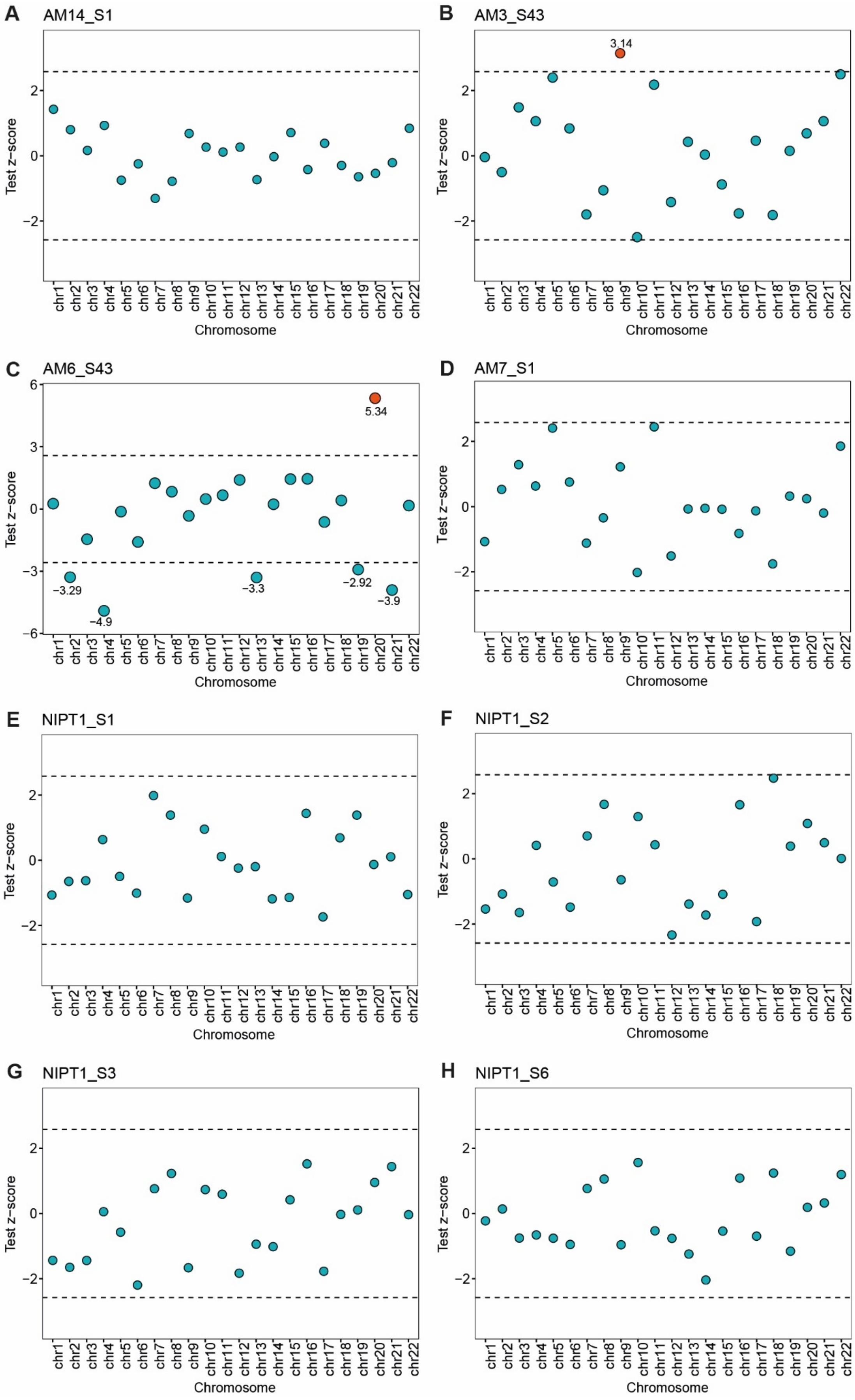
[redacted]

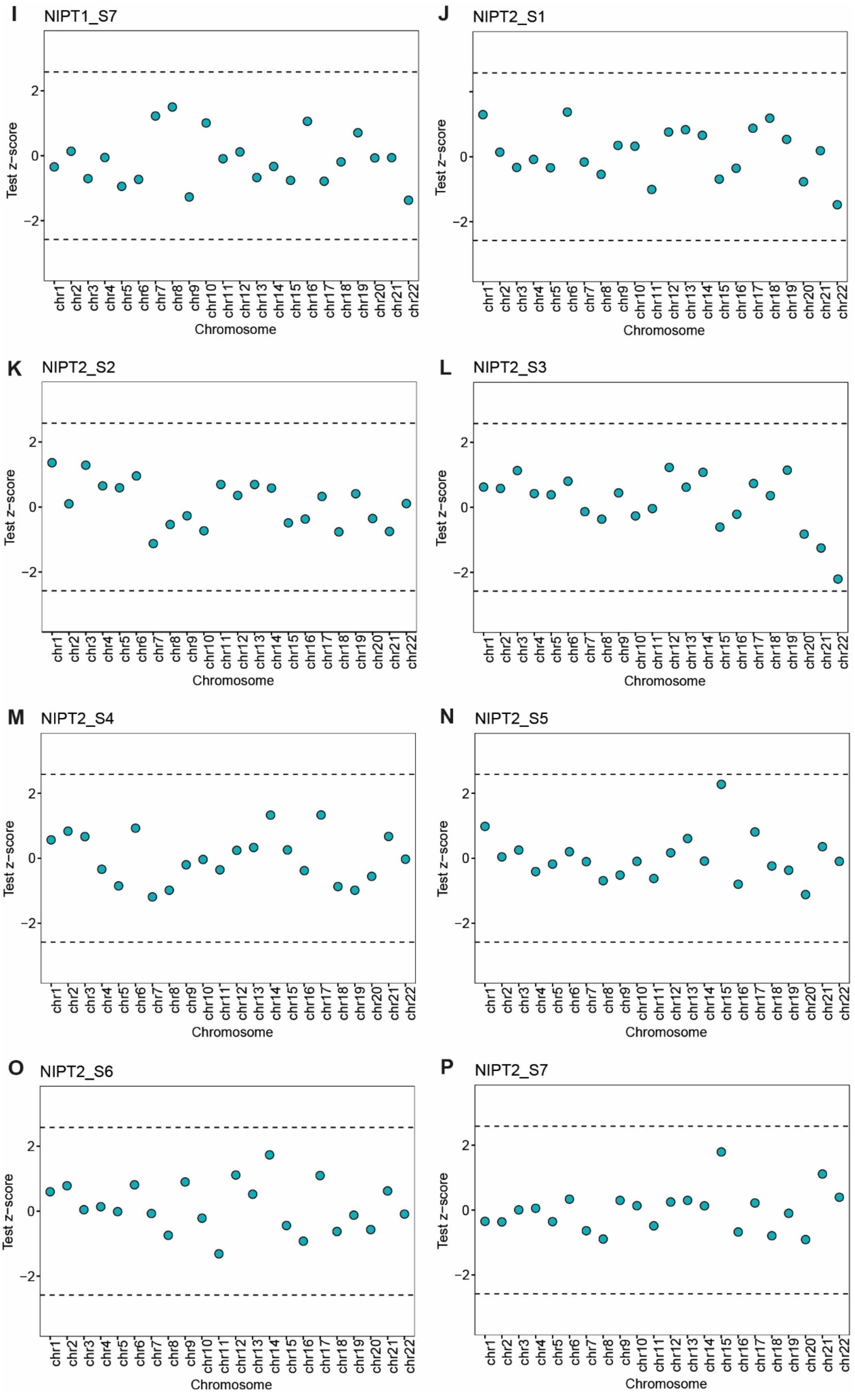
[redacted]

4.8 Detection of fetal chromosomal aneuploidies

The major goal of the methodology described here is to sensitively and specifically detect chromosomal aneuploidies at the earliest stages of fetal development. The basis of the approach is one that is widely used for the detection of chromosomal aneuploidies in the context of NIPT[65,127,130]. For each chromosome, a distribution of expected reference amounts of relative coverages [redacted] is generated using a set of control samples (euploid fetuses according to first trimester screening). [redacted]

[redacted]


A
AM14_S1
B
AM3_S43
3.14
C
AM6_S43
5.34
−3.29
−4.9
−3.3
−2.92
−3.9
D
AM7_S1
E
NIPT1_S1
F
NIPT1_S2
G
NIPT1_S3
H
NIPT1_S6
Test z-score
Chromosome
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
2
0
−2
6
3
−3
−6


I
NIPT1_S7
J
NIPT2_S1
K
NIPT2_S2
L
NIPT2_S3
M
NIPT2_S4
N
NIPT2_S5
O
NIPT2_S6
P
NIPT2_S7
Test z-score
2
0
−2
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
Chromosome

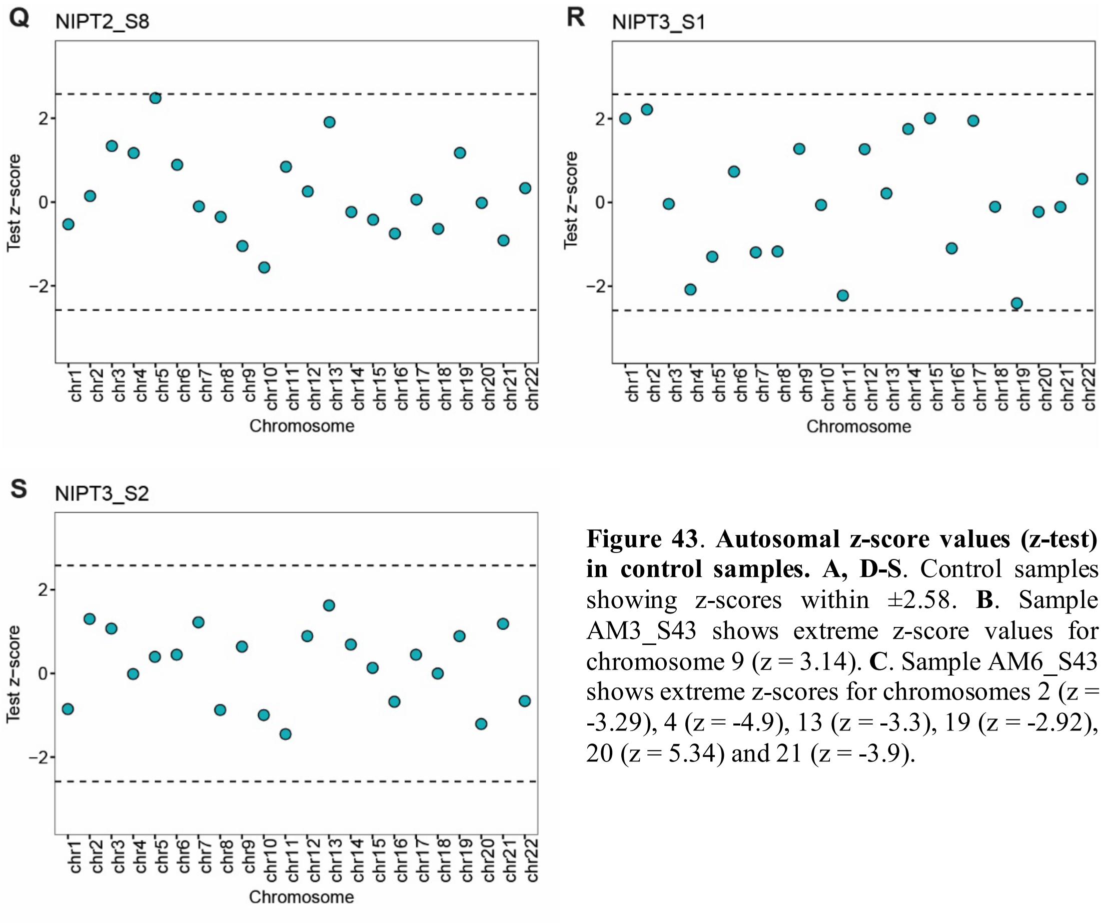
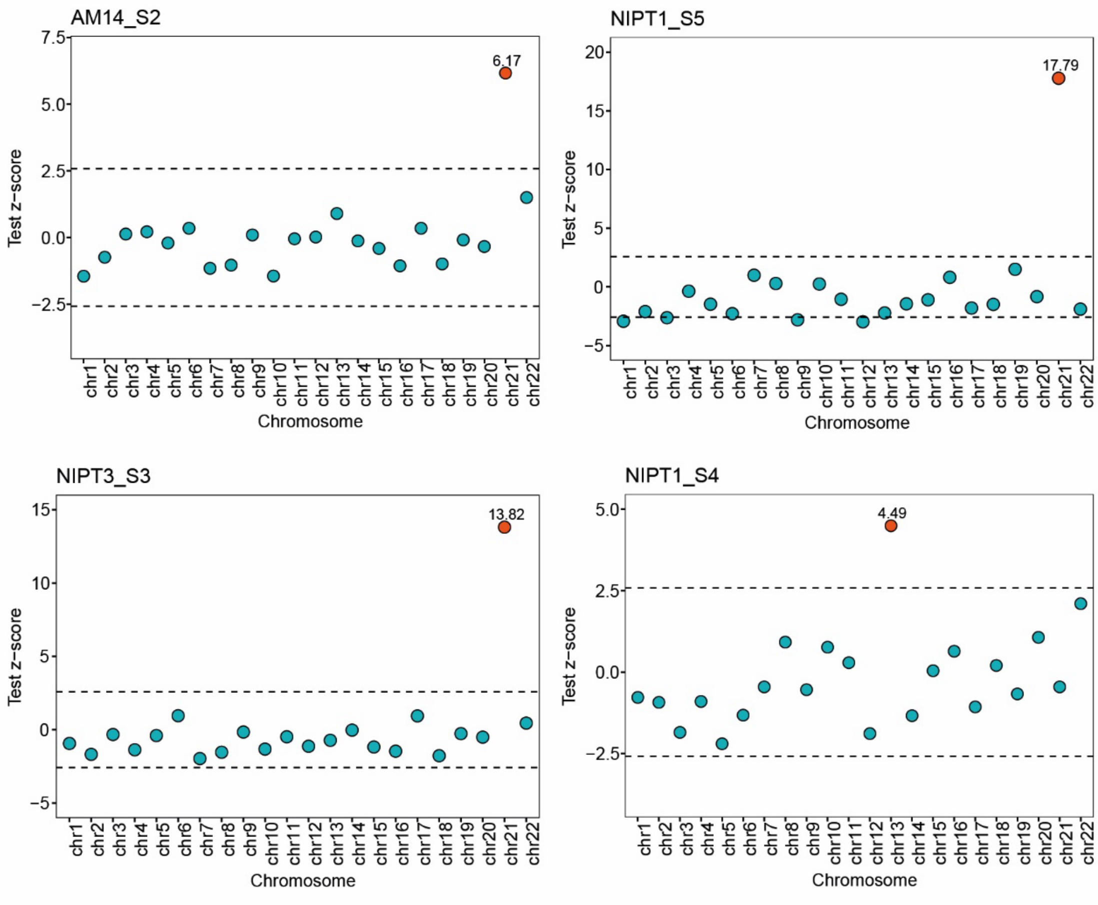


Figure 43. **Autosomal z-score values (z-test) in control samples. A, D-S**. Control samples showing z-scores within ±2.58. **B**. Sample AM3_S43 shows extreme z-score values for chromosome 9 (z = 3.14). **C**. Sample AM6_S43 shows extreme z-scores for chromosomes 2 (z = -3.29), 4 (z = -4.9), 13 (z = -3.3), 19 (z = -2.92), 20 (z = 5.34) and 21 (z = -3.9).

This procedure was repeated until all control samples were evaluated. In all of them, except in samples AM3_S43 (**Figure 43B**) and AM6_S43 (**Figure 43C**), chromosomal z-scores (z-tests) were found between the ±2.58 threshold, which means that the chromosomal read densities evaluated are found within the expected range of values for true control samples. Sample AM3_S43 showed a z-score of 3.14 for chr9, meaning that the read density in this chromosome was higher than expected compared to the rest of control population. In addition, sample AM6_S43 showed extreme z-scores for chromosomes 2, 4, 13, 19, 20 and 21: the read density was lower than the reference read densities in the case of chr2, 4, 13, 19 and 21 (z-scores of -3.29, -4.9, -3.3, -2.92 and -3.9 respectively) and higher in the case of chr20 (z-score = 5.34).

We next tested whether we were able to detect the presence of chromosomal aneuploidies by comparing differences in the relative coverage of all chromosomes between control and potential aneuploid samples, as determined in the first trimester screening.

Using the same method, we were able to correctly identify 4 aneuploid cases (three T21 and one T13) that had been previously confirmed by CVS and QF-PCR. Z-scores (z-tests) for samples AM14_S2 (T21), NIPT1_S5 (T21), NIPT3_S3 (T21) and NIPT1_S4 (T13) were 6.17, 17.79, 13.82 and 3.89 respectively (two-tailed test, p-value < 1e-4). An additional sample (NIPT1_S8) classified as a potential case by first trimester combined screening was found to be a false positive, as none of the analyzed chromosomes showed a z-score above the threshold set at 2.58 (**Figure 44**). Our findings were confirmed by karyotyping (samples NIPT1_S4, NIPT1_S5 and NIPT3_S3) and QF-PCR (all of them).



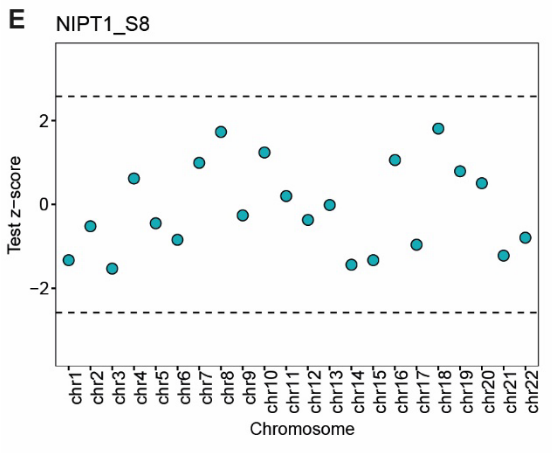


Figure 44. Chromosomal z-score values (z-test) in potential aneuploid samples. A. AM14_S2 (T21). **B.** NIPT1_S5 (T21). **C.** NIPT3_S3 (T21). **D.** NIPT1_S4 (T13). **E.** NIPT1_S8 (control with abnormal screening results). Z-score values for chr21 in samples AM14_S2, NIPT1_S5 and NIPT3_S3 were 6.17, 17.79 and 13.82 respectively. Chr13 showed a z-score of 3.89 in sample NIPT1_S4 and for NIPT1_S8 all z-scores were found within the predefined normality range.

The sensitivities and specificities were calculated considering two different situations. The first situation is one in which only chr21, 18 and 13 with a z-score >2.58 is reported (**Table 15**). In this case, the sensitivity and specificity would be both 100% (**Table 17**). The second situation is one in which z-score >2.58 in chromosomes other than 21, 18 and 13 are also reported (**Table 16**). In this case, the resultant sensitivity would still be 100%. However, the specificity would decrease to a 90% (**Table 17**).

Table 15. Contingency table in which only potential trisomies (z-test >2.58) for chr21, 18 and 13 are considered.

	Actual trisomies	Actual controls	Total
Predicted trisomies	4	0	4
Predicted controls	0	20	20
Total	4	20	24

Table 16. Contingency table in which potential trisomies (z-test >2.58) for any chromosome are considered.

	Actual trisomies	Actual controls	Total
Predicted trisomies	4	2	6
Predicted controls	0	18	18
Total	4	20	24

Table 17. Sensitivity and specificity for the detection of either i) chr21, 18 and 13 (Table 15) or ii) all chromosomes (Table 16).

	Sensitivity	Specificity
Chromosomes 21, 18 and 13	100 %	100 %
All chromosomes	100 %	90%

4.9 Noninvasive fetal sex determination

In the context of NIPT, the simplest strategy for fetal sex determination in maternal blood is the identification of chrY-derived cffDNA in male-bearing pregnancies, which has been shown to be reliable as early as gestation week 8[141,142,204-206], earlier than CVS and ultrasound.

During a normal, single pregnancy, the presence of chrY-derived cfDNA in the mother's bloodstream is a clear indication of the presence of a male fetus. [redacted]

We first aimed at evaluating whether we could [redacted] determine fetal sex through the detection of chrY-derived sequences in male-bearing pregnancies. To this end, we first simply calculated for each sample the percentage of reads mapping to [redacted] chrY [redacted] (%chrY) with respect to the total number of reads, in both male and female-bearing pregnancies. As expected, in male fetuses the %chrY is higher than in female fetuses (Wilcoxon rank-sum test *P* = 4.70e-5; **Figure 45A** and **Supplementary Table 8**, unfiltered reads). However, the male and female distributions almost overlap, which suggested that a significant number of reads fall into chrY intervals also in females. To further explore this, we created a heatmap with the %chrY [redacted] both in males and females (**Figure 45B**). From this visualization it becomes clear that there are a number of chrY intervals that are sequenced in both male and female bearing pregnancies. [redacted]

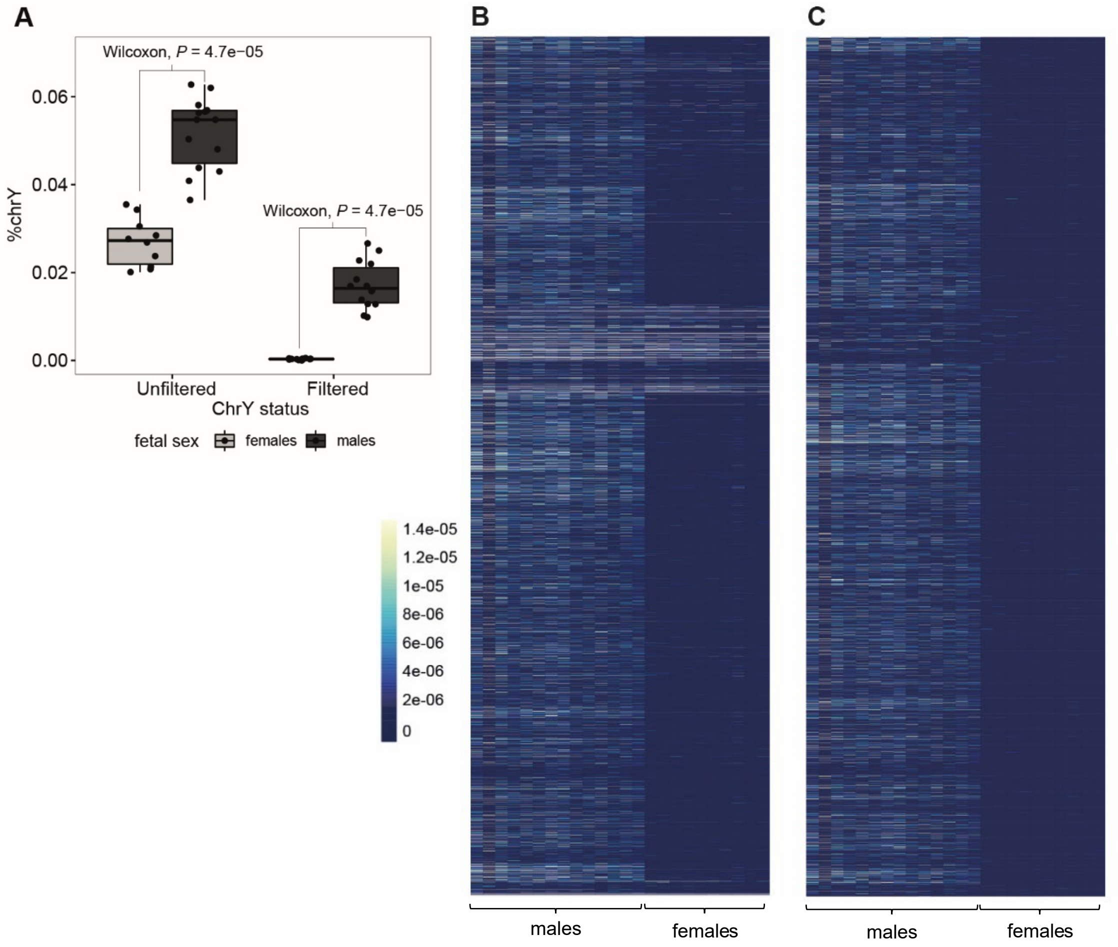


Figure 45. **ChrY-derived reads in male and female pregnancies before and after chrY filtering**. **A.** Distribution of the percentage of reads mapping to unfiltered chrY (left) and filtered chrY (right) in male and female fetuses. **B.** Heatmap showing unfiltered read counts mapping to chrY in male and female pregnancies. Genomic regions are shown row-wise and samples are shown column-wise. **C**. Heatmap showing filtered read counts mapping to chrY in male and female pregnancies. Heatmap data sorted by chromosomal location.

Because chrY reads observed in both male and female fetuses are not informative of fetal sex, we next attempted to increase the sensitivity for the detection of male pregnancies by filtering out those regions containing read counts in more than one female fetus sample. In other words, only those chrY regions that rarely contained reads in female samples were kept, while regions that were frequently covered among female samples were filtered out.

Filtering had consequences on both the horizontal (total number of genomic regions in chrY) and the vertical (total number of reads mapped to chrY) coverage available for further analyses. [redacted]. Because part of the filtered regions were also present in male-bearing pregnancies, filtering resulted in a decrease in the %chrY in both female and male fetuses. [redacted] The Wilcoxon rank-sum test showed that the differences in coverage in %chrY between male and female fetuses were significant regardless of chrY being filtered or unfiltered (*P* = 9.40e-5 in both comparisons). However, filtering resulted in a relative chrY signal increase in male-bearing pregnancies with respect to female-bearing pregnancies (calculated as the average %chrY in males divided by the average %chrY in females) from 1.92-fold to 60.90-fold (**Figure 45A** and **Supplementary Table 8**, filtered reads). [redacted]

After establishing the need to filter chrY reads to better discriminate male from female pregnancies, we focused on establishing a test that would allow us to predict fetal sex and identify aneuploidies affecting sexual chromosomes. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

4.10 Fetal fraction estimation

Next, we aimed at determining the FF using [redacted]. To this end, several strategies based on the methods mentioned in the **Introduction** chapter were explored in order to devise a method that can accurately measure FF at the lowest possible cost. The first strategy that we explored was the Y-chromosome-based approach for being one of the simplest and most accurate methods for FF determination in male fetuses. It should be noted that all the results in **Section 4.10** are validated using an external FF measure obtained from the ffPanorama test which had been performed in these samples in HUD. Nevertheless, this measure was not available in the beginning of this work, which prompted us to explore all the described strategies below.

4.10.1 ChrY-based approach (ffChrY)

To calculate the FF based on the amount of Y-derived sequences (ffChrY) in our male-bearing pregnancies, the percentage of sequencing reads mapping to filtered chrY (%chrY$_{\text{filt}}$) in male fetuses (n = 14) was first calculated, which resulted in an average of 0.017% (range 0.009–0.027%) of reads mapping to chrY. For a subset of them (n = 7), an external FF measure obtained from the SNP-based Panorama® commercial test (Natera), was available (ffPanorama). Comparison of the %chrY to the ffPanorama revealed a statistically significant strong positive correlation (r = 0.96, *P* = 0.003) (**Figure 49A**).

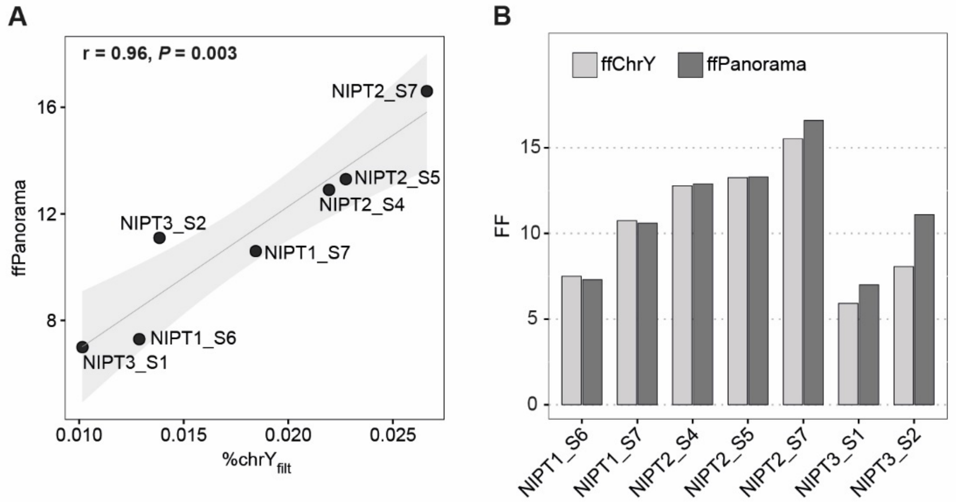


Figure 49. **ChrY-based FF estimation**. **A**. A strong positive correlation between the %chrY$_{\text{filt}}$ and the external FF measure ffPanorama for a set of male-bearing pregnancies is observed (r = 0.96, *P* = 0.003). Shaded region corresponds to a 95% confidence band. **B**. Comparison of ffChrY and ffPanorama FF values.

We next aimed at translating the %chrY$_{\text{filt}}$ values to FF values (ffChrY). To do this, we first calculated the %chrY in the cfDNA of an adult male sample to determine what is the expected %chrY in an undiluted male sample. In other words, the %chrY in the cfDNA of a male adult individual would represent the expected proportion of chrY molecules in a hypothetical maternal sample in which 100% of the molecules in the cfDNA pool were of male fetal origin. The calculated value (%chrY$_{\text{male ref}}$ = 0.17%) was used as a scaling factor to calculate the ffChrY, that is, the proportion of fetal genome in the maternal sample. Results show that the resultant ffChrY values are found very close to the absolute values provided by ffPanorama (**Figure 49B** and **Supplementary Table 11**).

Because we have previously shown that a proportion of the reads mapping to chrY should be filtered out to better discriminate between males and females, we filtered the reference male sample chrY resulted in a scaling factor of 0.12 (%chrY$_{\text{filt male ref}}$ = 0.12%). When a reference value of 0.12 was used to generate FF values, the determined ffChrY was considerably higher for all samples. The ffChrY values were not only higher compared to ffPanorama values but were also slightly higher than what it is reported in the literature[115] (**Supplementary Figure 3**).

Ideally, the transformation of %chrY$_{\text{filt}}$ values into ffChrY values in pregnant women samples should be performed taking an averaged scaling factor obtained from a pool of reference adult males, which was not possible in the context of this work. In addition, the approximation based on the removal of those regions identified in female-bearing pregnancies might not be adequate in this context. Therefore, we decided to use the unfiltered scaling factor of 0.17 for the reference male sample as it provided the best fit when comparing our ffChrY with the external ffPanorama.

Because determining the FF based on chrY was valid only for male pregnancies, we next aimed at finding a method for calculating FF that could be applied to both male and female fetuses. In this analysis, only samples with maternal information available (that is, matched gDNA) were used, which represent a total of 16 samples (10 male fetuses and 6 female fetuses).

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

4.10.3 SNP-based approach

One of the most well-established chrY-independent approaches that works equally well for both male and female fetuses is the use of SNPs to distinguish fetal from maternal DNA. Because it evaluates the presence of paternally-inherited alleles in the cfDNA, similar to the presence of chrY, it allows a direct assessment of the proportion of the cffDNA in the maternal circulation. Nevertheless, the SNP strategy comes at the cost of having to perform an additional assay, since the maternal gDNA needs to be sequenced. With the variant call information of maternal gDNA and cfDNA, fetal-specific alleles can be identified at those loci in which cfDNA (contains maternal and paternal alleles) and maternal gDNA genotypes differ. The ratio of the fetal-specific to maternal variants has been shown to be proportional to the FF[135,136].

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

In summary, we have shown that it is possible to derive accurate FF calculation methods by [redacted] The methods described here covered those already described (chrY-dependent and SNP-dependent) and include a novel method [redacted] that are valid for both male and female-bearing pregnancies. A summary of all the measured correlations with the different FF calculation methods described here can be found in **Table 21**.

[redacted]

4.10.6 FF and z-score correlation in aneuploid chromosomes

It is well established that the degree of deviation (here represented by the z-score value) shown by a given aneuploid chromosome from the reference euploid chromosomes is directly related to the FF[211,212,130]. This is, the more cffDNA molecules in cfDNA, the higher the deviation will be for an aneuploid chromosome with respect to the distribution of euploid chromosomes. To test the validity of this premise in our samples, we compared the z-score values obtained for our 4 aneuploid samples

(AM14_S2 [T21], NIPT1_S4 [T13], NIPT1_S5 [T21] and NIPT3_S3 [T21]) with their FF (**Figure 57**). For these samples, the FF measurements provided by the Panorama test were not available, since patients presenting abnormal results in the screening test would directly undergo invasive testing rather than NIPT. Therefore, the z-score values were compared to our own FF measurements, [redacted], which, as we have previously shown, are well correlated with ffPanorama values and work with male and female-bearing pregnancies.

[redacted]

[redacted]

[redacted]

Sample corresponding to T13 shows the lowest FF not only when compared to the other trisomic samples, but also when compared to the rest of samples in the dataset.

This observation is consistent with previous findings that trisomies of chr13 and chr18 generally show a lower FF than trisomies of the chr21[125,213] and it is believed to be caused by a decrease in the release of cffDNA in blood by placentas affected by this kind of aneuploidy. Because 1) FF likely depends on the presence of specific aneuploidies and 2) it has been shown that NIPT tests fail to detect chromosomal aneuploidies in which the FF is <4%[214], the aneuploidy detection sensitivity and specificity will likely have to be tailored to each chromosome.

4.11 Summary of fetal sex, aneuploidy and FF estimations

A summary of several parameters calculated regarding fetal sex determination, aneuploidy detection and FF estimation can be found in **Table 22**.

[redacted]

4.11 Summary of fetal sex, aneuploidy and FF estimation

4.12 Costs

We first aimed at calculating the cost of each set of steps involved in producing a single sample as described in this work. The goal was to compare the overall costs of targeting Alu elements for the detection of chromosomal aneuploidies to other commercial NIPT tests. [redacted]

[redacted]

In a second step, we aimed at identifying the steps that could be potentially optimized in order to bring costs down. A total of four main optimization points were initially identified:

1. [redacted]

2. The second point relates to the multiplexing level: this is, how many samples are hybridized together and co-sequenced. [redacted] We calculated the costs of co-sequencing 16 samples, which are summarized in **Table 23B** (see Mid output 2x150 bp columns). We have conducted preliminary downsampling experiments that indicate that it is possible to specifically detect all the aneuploidies in this study at 1/2 the coverage that is achieved at an 8-sample multiplexing level. Other parameters, such as fetal sex determination of FF estimation have not been evaluated yet. This coverage would correspond to a multiplexing level of 16 samples per run.

3. The third point is related to sample preparation: although 2 ml of plasma were used for sample preparation in this work, we have successfully prepared samples from as low as 1 mL of plasma. Using only 1 mL of plasma would bring costs related to extraction and sample preparation down to around 25%.

4. The last point is related to the number of sequencing cycles used. In the experiments described here, samples were sequenced using a 2x150 bp protocol. This means that each dsDNA molecule is read 150 bp from each end. Because most of the sequenced molecules have an average size of ~165 bp (see **Section 4.4**), this means that for a majority of molecules the forward and reverse reads largely overlap. Thus, it would likely be possible to generate the same amount of unique information by using a cheaper 2x75 bp protocol, reading a total of 150 bp per dsDNA fragment (Table **23B**, see Mid output 2x75 bp columns).

In summary, this preliminary cost an optimization analysis indicates that it is likely possible to produce and sequence cfDNA libraries at costs around 100€ per sample, starting from 1 mL of plasma and using a multiplex level of 16 samples per run and a 2x75 bp sequencing protocol.

5. Discussion

5.1 The access to clinical samples

The works described in this Thesis report have been mostly carried out within the Research and Development department of the company qGenomics. For the time being, qGenomics has very limited access to patients. Therefore, in order to have consistent access to specific types of samples that can be used for research purposes, multiple research agreements had to be made over the years with different institutions. This has led to some very fruitful collaborations with public as well as private hospitals. Relevant to this work, a close collaboration was established with Hospital Universitari Dexeus (HUD - Grupo Quirónsalud) to have access to samples from pregnant women. [redacted]. From 24 samples analyzed in this work, only 4 correspond to pregnant women carrying a trisomic fetus (either T13 or T21) as confirmed in the laboratory by karyotyping and QF-PCR. The access to trisomic samples has been limited by at least two situations: first, the fact that these aneuploidies have a low prevalence in the population and second, although the possibility to participate was offered equally to all pregnant women, women falling within the high-risk group for fetal aneuploidies were more prone to refuse to participate in the study than those who were informed to have a low risk.

In order to better validate our experimental approach, we plan to enroll a second, larger series of samples. To this end, we have applied to, and successfully obtained approval from, the ethics committee at HUD for an amendment of the project that allows the recruitment of more samples. The amendment will allow the recruitment of up to 100 controls and 50 potential aneuploid cases. In order to increase the enrollment rate of high-risk pregnancies we are considering expanding the information provided to the mothers in a high-risk situation about the benefits of participating in research projects.

5.2 Special handling and quality control requirements for the manipulation of plasma and cfDNA samples

Typically, the bulk of NGS procedures that are performed in a clinical genetics' laboratory involve the extraction of gDNA, mostly from whole-blood samples. The manipulation of these samples follows very standard procedures used at qGenomics and in laboratories worldwide that require little caution for their handling and manipulation, with the basic goal of preserving gDNA's integrity. [redacted]. However, the manipulation of samples for cfDNA extraction requires much more care in their collection, transport and extraction in order to preserve cfDNA's integrity and relative concentration. For instance, more expensive special blood collection tubes (Streck BCTs) that preserve cell integrity must be used. Additionally, these BCTs cannot be stored cold, as low temperatures facilitate hemolysis, and large whole-blood volumes (typically ~8 mL) are required to even extract a few nanograms of cfDNA. Moreover, plasma samples must be obtained from whole-blood samples as soon as possible after sample collection to, once more, avoid hemolysis. All these limitations pose many challenges that had to be initially overcome in a laboratory that had no previous experience in the extraction and manipulation of cfDNA samples.

At a time when, other than the widespread use of BCTs, current cfDNA processing workflows lack standardization, the present project has tried to establish quality controls at different points of the processing workflow mainly focused on i) assessing the presence of hemolysis in plasma and to ii) assure an effective cfDNA isolation.

Regarding the presence of hemolysis in plasma, we have not found any references in the literature that clearly establish methods to quantify it in the context of the NIPT routine. However, according to our experience in the laboratory, plasma samples showing a reddish coloration by visual inspection are usually discarded for downstream analyses by NIPT companies[236]. Other than the mere visual plasma color inspection, in this work we have [redacted]

[redacted] to evaluate the presence of hemolysis [redacted] This allowed us to determine that none of the plasma samples collected exhibited evident hemolysis signs as shown by the plasma coloration, which was in agreement with [redacted]. The generation of a distribution of reference [redacted] given our specific sample collection and transport conditions will allow us, in the future, to obtain a specific hemolysis value threshold to either accept or reject a specific sample for further processing and analysis. In addition, the extraction of cfDNA showed that no correlation was observed between pre-processing time and [redacted] of extracted cfDNA, suggesting that if low hemolysis levels (undetectable by our absorbance method) were present, they did not increase under the established transport and processing conditions to detectable levels.

In this work we also used fluorometric, capillary electrophoresis and real-time qPCR methods to confirm that an effective isolation of cfDNA had been achieved and to exclude the presence of detectable amount of gDNA.

The fact that the concentrations of the extracted cfDNA measured through fluorometric methods showed comparable concentrations to the reported values in published data[27] and the identification of chromosome-Y-derived sequences (indicative of the presence of cffDNA) in the maternal plasma of a male-bearing pregnancy suggested that cfDNA was being isolated effectively. In addition, the size distribution of the extracted fragments, assessed through electrophoretic methods, was found within the expected nucleosomal range for cfDNA[20], not only confirming the efficient extraction of cfDNA, but it also excluding the possibility of gDNA contamination.

A more sensitive quality control might be applied in the future, based on the quantification of short (mainly originating from cfDNA) and long DNA (mainly originating from cellular gDNA) fragments using real-time qPCR that can provide an accurate quantification of the proportions of cfDNA and gDNA in the extracted samples and that can be applied to both male and female fetuses.

5.3 Challenges and consequences of working with low input DNA

The fact that the levels of cfDNA in blood, although variable among individuals, are typically very low (range 1–30 ng/mL approximately)[27] has important implications for library preparation for NGS. In this work, the total cfDNA used in library preparation ranges from less than 1 to 22 ng, which corresponds to ~290 to 6,380 haploid genomes. [redacted]

Such low input amounts of DNA result in a pool of adapter-ligated molecules with low molecular complexity, which can exacerbate sequence representation bias and duplicate generation during PCR amplification and sequencing. Sequence representation bias is a common PCR artifact and is mainly due to variable amplification efficiencies among sequences containing different GC content, broad DNA library size range or polymerase errors[15,216].

In the context of this work, several critical points in the standard library preparation workflow were identified and modified in order to minimize the aforementioned effects. These modifications included the optimization of adapter molarity and the reduction of the number of PCR cycles to a minimum for both gDNA and cfDNA libraries.

5.3.1 Adapter optimization to maximize library complexity and minimize adapter dimer formation

Ideally, ligation of 100% of the molecules in the sample is desirable, as it theoretically assures a maximal library complexity for a given DNA input. [redacted] A common strategy to increase ligation efficiency is to increase the amount of adapter. The downside of this strategy is that unligated adapter molecules can self-ligate during the ligation reaction to form what are known as adapter dimers. Because adapter dimers are the smallest molecules within the library that can be formed, they are preferentially amplified and sequenced in the subsequent steps, greatly decreasing the number of informative sequencing reads. Therefore, the optimal amount of adapter to use has to be optimized for a given amount of input DNA so that the number of ligated DNA molecules is maximized and the number of unligated adapter molecules is minimized.

As described in the **Results** chapter, the total number of PCR cycles performed after library ligation has an effect on the final adapter dimer to DNA library ratio, and therefore, the amount of adapter should be ideally optimized taking into account not only the DNA input, but also the number of cycles that will be applied to DNA libraries. In turn, the total number of cycles that will be applied to DNA libraries is dependent on the number of ligated molecules available. [redacted]. Although this represents a conservative approach, the fact that no adapter dimers were observed during any of the amplification reactions allowed us i) to be less strict during the bead-based size-selection purification, allowing the recovery of a greater number of cfDNA fragments, especially smaller fragments, which are thought to be preferentially of fetal origin[19,68,145] and ii) to use a single adapter-DNA ratio independent of the number of PCR cycles applied.

[redacted]

This result highlights the necessity of experimentally identifying the optimal amount of adapter to ligate, using manufacturer's instructions as a general starting point. This requirement is less likely to be relevant when high amounts of input DNA are used, although in any case it can lead to a significant reduction in the consumption levels of adapters and significant long-term savings.

5.3.2 PCR cycle optimization to minimize sequence bias, duplicate levels and polymerase errors

One of the central assumptions of the analysis of sequencing data is that each read constitutes an independent observation, that is, that each read comes from an individual biological molecule in the original sample. However, PCR-based library amplification introduces the possibility that, given a sufficiently low diversity and enough PCR cycles, multiple sequences in the final library come from the same original DNA fragment. PCR amplification biases (due to different GC content or broad DNA library size range) and an excess of PCR amplification rounds can ultimately lead to the generation of PCR duplicates, which are sequence reads that result from sequencing the exact same DNA fragment more than once. If not removed, PCR duplicates can lead to erroneous conclusions in certain analyses. For instance, in any scenario where depth of coverage is an important factor as in our work, PCR duplicates can erroneously inflate the coverage in certain regions.

Because very little amounts of amplified library are loaded onto the sequencer, when amplifying ligated DNA libraries, PCR cycles must be kept to the minimum necessary to obtain enough amplified product that can be accurately quantified with the methods available in the laboratory. For instance, a library with an average fragment size of ~300 bp can be amplified to a final concentration of 1 ng/µL and this concentration (roughly 5 nM) is still 2,500-fold higher than the sequencer loading concentration (1.8 pM). In the context of this work, in which the input amount of cfDNA is determined by its amount in plasma, the amount of PCR cycles required was specifically adjusted for each sample using real-time qPCR. [redacted]. In future studies, qPCR cycle determination may be unnecessary as ideally, the number of required PCR cycles should be established as a function of the initial amount of sample input. It should be noted that because cfDNA and gDNA libraries have different fragment sizes, the relation between input and number of cycles may vary depending whether the sample is cfDNA or gDNA.

Initial analysis of sequencing data quality metrics revealed that library optimization for low DNA inputs was successful in terms of sequencing diversity, as the average percentage of duplication in our cfDNA (except those samples for which there was technical error) and gDNA samples was lower than expected [redacted]. [redacted]

5.4 Targeted sequencing [redacted]: overcoming the challenges and taking advantage of the opportunities

[redacted]

[redacted]

5.4.1 The potential challenges

The use of [redacted] during hybridization come with a number of potential challenges, both experimental and bioinformatic, that were addressed at the beginning of the project.

[redacted]

[redacted]

[redacted]

5.4 Targeted sequencing [redacted]: overcoming the challenges and taking advantage of the opportunities

[redacted]

[redacted]

[redacted]

A second potential challenge is related to a well-known technical limitation of hybridization-based target enrichment methods such as the one used here and in routine procedures at qGenomics. This limitation, known as daisy chaining, takes place when two or more individual library molecules hybridize through their complementary adapter ends, in a structure that resembles a daisy chain. If this happens, when a target library molecule is specifically captured by a bait, it will also take with it the daisy-chained library molecules. These events result in the unwanted capture and sequencing of genomic regions that are of no interest for the experimenter (off-target regions), with the consequent waste of sequencing resources. This situation is typically limited by the use of blocking oligonucleotides (blockers) that hybridize the adapter portion of the library molecules, very effectively reducing the hybridization between library molecules through their adapters.

[redacted]

5.4.2 The opportunities

The reason we have taken this rather unorthodox approach is that [redacted] offers some very unique potential advantages over more traditional target-enrichment approaches [redacted].

The first advantage relates to the cost-efficiency of the methodology, a key point to keep in mind from a company's point of view. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

The overall results suggest that the presence and coverage [redacted] in a final [redacted] sequencing experiment is dependent on its presence/absence in plasma, its genetic distance to the probe (in turn dependent on [redacted]) and its degree of homology with [redacted]. The design of experiments that contribute to evaluate the effects of each of the above described potential factors as well as other well-known factors as DNA sequence composition or GC percentage (higher in younger families)[165] will be useful to deepen our knowledge as to whether a given [redacted] might or might not be captured. In addition, this could allow the design of additional probes to boost the capture of genomic regions of especial interest, for instance, [redacted].

5.6 General considerations on data pre-processing and normalization

From a bioinformatic standpoint, the data pre-processing steps described here resemble those in which no prior knowledge of the regions of interest is available. This is the case of chromatin immunoprecipitation sequencing (ChIP-seq) assays, in which the analysis pipelines require to initially identify what genomic regions bind a given protein. Thus, binding peaks must be identified as genomic regions where reads accumulate over a pattern of background signal. Similarly, the experiments described here required the initial identification of the genomic regions that were captured. [redacted]

5. Discussion

[redacted]

[redacted]

[redacted]

[redacted]

Although this strategy may seem convoluted [redacted], the fact is that it was key to the detection of chromosomal aneuploidies. Other normalization strategies that were assessed (data not shown) provided a lower sensitivity for the detection of T21 samples and provided a false negative result for T13. This is the case of a modification of the Transcript per kilobase Million (TPM) normalization method, based on dividing the chromosomal read density by the total sample read counts followed by chromosome-specific z-score calculation. We hypothesize the application of a modified version of the TPM method didn't work as well as our method because the TPM method was developed to analyze transcription units[221], much shorter than chromosomes. [redacted]

5.7 Normalized cfDNA coverage patterns are associated with key genome architecture features

[redacted]

Further analyses revealed that this strong correlation can be attributed to the fact that [redacted]. These results are in agreement with previous works showing that accessible genomic regions tend to be overrepresented in cfDNA as their cleavage by endonucleases is favored in comparison to more inaccessible regions[18,199,210]. The flipside of this observation is that some genomic regions of interest may be relatively absent from cfDNA and thus, when performing this kind of experiments with cfDNA it is crucial to assess whether the regions of interest are well represented.

[redacted]

[redacted]

The results described here are very reminiscent of a large body of classic experiments that have aimed at studying chromatin accessibility in many model organisms under different experimental conditions, both *in vivo* and *in vitro*. For instance, MNase has been largely used in experiments directed at both addressing nucleosome positions and nucleosome accessibility in the context of transcription, recombination, repair and replication[222]. Because MNase preferentially cuts chromatin in the accessible DNA linker spaces between nucleosomes, it produces a characteristic ladder pattern much like the one observed in cfDNA. In a sense, analyzing cfDNA is akin to performing a chromatin accessibility assay with human specimens *in vivo* in which the experimenter has no control whatsoever over the experimental conditions. Future experiments will be directed at expanding the limited body of work published to date regarding nucleosome position and cfDNA fragment size[18,199,210], a piece of information that, together with the identification of highly accessible regions in cfDNA, may help better identify what cell type(s) are majorly dying within an individual. We hypothesize such methods could serve in the near future to develop screening strategies that will allow the real-time monitoring of anomalies in the cell-death patterns associated to pathological processes.

5.8 Detection of fetal chromosomal aneuploidies through relative differences in chromosomal coverage: a proof-of-principle prototype

During the development of any new device or methodology that aims at providing a new or improved service, several phases are typically followed. First, an uncovered need or new functionality are identified. Design of a strategy to successfully cover the need follows, with the small-scale production of working prototypes that prove the feasibility of the approach. Larger-scale testing subsequently proves (or disproves) that the new process or device can be scaled and meets the specified characteristics it was designed for.

Although the methodologies described here can be applied to several fields, in this project we aimed at showing the feasibility of [redacted] through its application to the NIPT field. To this end, we have designed a cost-effective methodology that would allow to sensitively and specifically detect chromosomal aneuploidies at early stages of fetal development. The work presented here describes the first-generation prototype that, as we will discuss in the following sections, demonstrates the feasibility of the project. As mentioned at the beginning of this discussion, a new phase will follow in which we will assess whether the conclusions reached here hold true when a larger series of samples is used.

[redacted]

5.8.1 Sensitivity in the detection of aneuploidies

Although we acknowledge that the number of true positive aneuploid samples used in this work is very limited (n = 4 aneuploid cases), the generated results clearly indicate that our strategy is able to sensitively detect full chromosome aneuploidies, as it correctly identified all 4 cases with trisomies of chr13 and chr21. Moreover, the same strategy that is able to detect extra chromosomal copies was used to very sensitively detect the presence of a single chrX copy in all 14 male fetuses. This result strongly suggest that our methodology may be also suitable for the detection of chromosomal deletions.

5.8.2 Specificity in the detection of aneuploidies

Commercial NIPT tests clearly state the scope of their analysis, which means that they only provide results for the chromosomes of interest. For instance, many NIPT tests offer the possibility to assess the potential presence of aneuploidies only in chr13, 18 and 21. Therefore, if only these three chromosomes are taken into account, our test is 100 specific in the identification of triploidies, as no false positives were detected among the 24 cases (20 true negative and 4 true positive cases). However, if the scope of the test widens to assess potential aneuploidies in any chromosomes, then two of the samples (AM6_S43 and AM3_S43) would present one or more chromosomes beyond the range of normality defined in this work, which includes all z-score values between ±2.58, and more specifically z-scores >2.58, indicative of the presence of a potential trisomy. Hence, using this threshold, only 1% of true negative control samples would have z-score value higher than 2.58. One possibility to avoid false positive cases would be to adjust the threshold values for each chromosome, raising the threshold in those chromosomes that are more prone to variation among control samples. We believe this type of analysis will only be possible when a higher number of true negative and positive samples are added to the analysis. Having access to more samples will also facilitate the analysis of the different samples according to their FF, as we and others[211,212,130] have shown that the deviation that a triploid chromosome departs from the average normal chromosomes depends on the FF.

Regardless of the thresholds used, we must assume all the extreme z-scores in sample AM6_S43 cannot be a faithful reflection of the chromosomal copy number status of the fetus, as these alterations, if true, would be incompatible with life[223]. Therefore, a potential explanation is that this sample shows more dispersion than the rest of samples due to differences related to maternal factors, for instance maternal age: this sample corresponds to a woman aged 46, making her the eldest woman in the sample set. Because it has been shown that aging causes profound epigenetic changes in the genome, and cfDNA is a reflection of chromatin's structure, it's not far-fetched to think that the high dispersion we see in this sample may be related to the mother's age[224].

[redacted]

A second sample, AM3_S43, shows a positive z-score value above 2.58 for chr9, and therefore, it is compatible with the presence of a trisomy in chr9. Although very rare, fetuses with a trisomy in chr9 (T9) have been found to develop to term[225,226]. In this case, we know the baby developed normally and no T9 was identified by NIPT. Interestingly however, we also know this sample carried a vanishing twin, in which chromosomal abnormalities are common, and are, in fact, important contributors of false positive results in NIPT counting methods (15%)[124]. Thus, the results reported for this sample are either compatible with a false positive result in our test or a true positive T9 for the vanishing twin.

Importantly, besides the 4 true positive samples we received that had increased risk from the first trimester combined screening and a positive karyotype result, we also received another sample that had increased risk from the first trimester screening but had a negative CVS result. This sample (NIPT1_S8) tested negative (normal) with our test, thus confirming the false positive of the first trimester combined screening. Had this sample been tested by NIPT, a confirmatory invasive technique would have been likely avoided.

[redacted]

When the first cfDNA extractions were performed in the laboratory, we asked for the easiest way to check if we had performed a successful isolation. At the moment we reasoned that if we were able to specifically detect Y-derived sequences in the cfDNA of male-bearing pregnancies, we must have successfully isolated cfDNA. These experiments, performed by qPCR, were successful and lead to the development of the rest of the sex determination methods described here based on the use of NGS read count densities. [redacted].

With regards to chrY, we first established the necessity to filter out those chrY intervals that do not discriminate between males and females. This initially unexpected finding cannot be explained by mapping ambiguities, since during data preprocessing multimappers, chimeric reads and reads with a low mapping quality (MAPQ score below 20) had been removed from the analysis. A literature search indicated that the presence of reads mapping to chrY in female-bearing pregnancies is a recurrent problem and is mostly attributed to technical artifacts such as male DNA contamination during library preparation[65,227]. We hypothesize that, besides the above-mentioned potential contamination, at least two other plausible situations may explain the observed results. [redacted]. Second, from a biological standpoint, it is well known that a continuous bidirectional trafficking of a variety of both cells and cell-free substances (e.g., DNA) exists between the mother and the fetus during pregnancy. Trophoblasts, nucleated erythroblasts, leukocytes, hematopoietic progenitor cells, and in rare cases, mesenchymal stem cells of fetal origin have been identified in maternal blood during pregnancy[228]. Interestingly, the detection of microchimerism in maternal circulation and tissues, and also the detection of Y-derived sequences in women with male offspring at different time intervals after delivery ranging from 1 to 60 years strongly suggests that cells of fetal origin can persist in the mother for decades after pregnancy[229]. These previously published findings would be compatible with the detection of Y-derived sequences in female-bearing pregnancies that have had a previous male pregnancy. However, we currently don’t have information on previous pregnancies that may help us determine whether we are detecting this type of chimerism in any of our samples. Regardless of the origin of reads mapping to chrY in female pregnancies, regions observed in both male and female fetuses were filtered out during data analysis, which contributed to increase the chrY male-bearing to female-bearing pregnancies ratio from 1.92-fold to 60.90-fold.

[redacted]

[redacted]

[redacted]

The sex determination methodology described here is very similar to the methodology used in the detection of potential autosomal aneuploidies in control samples. With this approach we not only wanted to be able to accurately define fetal sex, but also be able to accurately predict the presence of sex chromosome aneuploidies. [redacted]

[redacted]

[redacted]

[redacted]

Finally, although the methodology has been designed to be able to identify aneuploidies of the sex chromosomes, we have only been able to test normal samples, therefore showing that the test is able to discriminate when chrY is absent or present and [redacted]. Based on the results presented here, we hypothesize that this method should be able to identify sex chromosome aneuploidies. For instance, a X0 female (Turner) should be readily detected by our test as both male-compatible (one chrX copy) and female-compatible (absence of Y). A XXY male (Klinefelter) should readily be detected as both female-compatible (two chrX copies) and male-compatible (presence of Y). Again, having access to these and similar samples with sex chromosome aneuploidies will be required to test the accuracy of detection.

[redacted]

Being able to accurately measure FF is critical in NIPT tests, as FF is one of the major factors behind a no call (failed) result or a false negative result if not measured or not measured properly. As has been introduced at different points throughout this work, several ways have been described to measure FF. Some, like counting chrY-derived reads, are very accurate and relatively inexpensive but only work in male-bearing pregnancies (not universal). Some others, like the SNP-based methods, are also very accurate and can be used with both male and female-bearing pregnancies (universal), but they generally require the analysis of both the mother's own DNA and cfDNA, which increases the production costs. At a midpoint, there have been attempts to develop count-based methods that, unlike the chrY count-based method, can be used with both males and females[146]. However, the computational methods behind these analyses are complex and require the use of large datasets for their proper training[141,146].

In this work we have attempted to develop a novel approach to measure FF that overcomes some of the limitations of the current FF measure methods. [redacted] And we also wanted this method to be universal and as computationally simple as possible. We acknowledge that the path we have taken towards the successful achievement of our goal is not a simple one, but it is a faithful reflection of the complexity of the problem we wanted to tackle and the many hurdles we had to overcome during the process. This development path took us to explore the different FF calculation possibilities that [redacted]

[redacted]

Counting the fraction of chrY-derived reads in male-bearing pregnancies was the first and simplest strategy tested to measure FF (ffChrY) and showed the feasibility of our approach. Comparison of the ffChrY to ffPanorama (our external reference, available only towards the end of the project) revealed that our FF estimates were very close to those of the reference in terms of absolute numbers. [redacted]

[redacted]. Our results not only indicated that this is actually feasible, but that the FF calculated [redacted] is as good as using ffChrY itself or ffPanorama.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Because at the time this data analysis was being performed the ffPanorama data had not been made available to us, we were thus forced to come up with a FF calculation method that would work with males and females [redacted]. This situation led to the development of our SNP-based approach for FF calculation.

SNP methods for FF estimation are the gold standard because they provide high accuracy, but this accuracy generally comes at the expense of having to sequence maternal gDNA, which may increase the production costs. This is the case of the Panorama test, whose FF estimations have been used throughout this work as an external reference and has been shown to be able to measure FFs as low as 1.4% in aneuploid samples[137]. [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

In summary, we have implemented three possible approaches for calculating the FF that will be further validated in the second phase of the study in which more samples will be collected. From the methodologies tested, the one based in the proportion of reads mapping to chrY has shown the highest correlation with the reference ffPanorama values, which is expected, as chrY is exclusively derived from the fetus. The SNP-based approach has also proved to be strongly correlated to the FF in both males and females [redacted]

[redacted]

5.10.4 Costs and turn-around times

As it has been described throughout this work, the technologies behind commercially available NIPT tests and their associated costs are varied. However, this is not reflected in the resultant market prices, which have reached certain homogeneity, probably due to the competitiveness of the market. In fact, the price of any given test is mostly dependent on the conditions that are detected: the cost of detecting chromosomal aneuploidies, fetal sex determination and FF estimation is around 450-550 € and is increased up to 725-750 € if microdeletions are included.

One of the major goals of this project was to produce a NIPT test whose costs could be competitive in today's crowded market. Our experimental strategy, which makes use of a simple yet effective targeted sequencing approach, could easily deliver results around the 100 € cost tag (reagents only) in a laboratory with a mid-throughput sequencing platform like qGenomics. We believe there is ample room for improving cost-efficiency by calculating the minimum amount of coverage required through coverage simulations. Future cost calculations will have to consider the costs related to human resources.

For a given sample, obtaining results in a timeframe as the one offered by most commercial NIPT tests (5 to 10 calendar days except for the Harmony test, offering results from 3 to 5 days as microarrays are faster compared to NGS) is feasible taking only into account the time that is needed for each step in our workflow. However, it should be evaluated how this workflow fits in the daily company routine, very much dependent on equipment and human resources.

5.11 NIPT ethical concerns

The high sensitivity and specificity figures reported for NIPT has led to their implementation in the National Health Systems of several countries. However, it has also led some providers to believe that NIPS tests are diagnostic or virtually diagnostic. In The Netherlands, NIPT is offered as an alternative to invasive testing to women with a high-risk outcome from first trimester combined screening (risk ≥1:200) or women with an increased risk for T21, T18 or T13 because of medical history since 2014 (TRIDENT-1 study)[235].

Regardless of the performance of current NIPT technologies, there is an underlying biological reason for which cfDNA should not be regarded as a diagnostic tool, and that is the fact that it originates from the placenta cytotrophoblast cells, meaning there are a certain number of fetal conditions (see the **Introduction** chapter, **Section 1.9**) that can lead to the extraction of incorrect conclusions. It should be noted that CVS, which is considered a diagnostic test, analyzes the placenta too, although it does so by analyzing two different layers from the placenta (cytotrophoblast and mesenchyme), which increases the opportunity to detect conditions like placental or fetal mosaicism. As a consequence, there is general agreement that the use of NIPT should be restricted to a screening tool.

In addition, the fact that the sensitivities, specificities and FPRs of NIPT are superior to first trimester combined screening for the detection of T13, T18 and T21[235-238], has led some countries to offer NIPT as an alternative to first-combined screening: in the Netherlands, since April 2017, NIPT is also offered within the TRIDENT-2 study to low-risk pregnant women, who are given a choice between first trimester combined screening and NIPT. The current NIPT-based prenatal test includes detection of T21, T18 and T13[239]. However, NT thickness, evaluated during first trimester screening, has been demonstrated to be a marker of a range of fetal disorders that cfDNA is not able to detect, as for example cardiac anomalies, microdeletion syndromes, and some single-gene disorders[240-246].

Therefore, NIPT should be used in combination with first trimester combined screening for a more accurate risk assessment of chromosomal aneuploidies with a special emphasis in reducing the number of women undergoing unnecessary invasive procedures, but it should not replace neither screening nor invasive confirmation.

With noninvasive genetic screening on the rise, companies are racing to add chromosomal abnormalities even when research is still limited (for instance, sexual chromosome aneuploidies), the detection accuracy is not well established, or even if the effects derived from these chromosomal alterations are sometimes uncertain. An example of this is the detection of subchromosomal alterations like microdeletions, and in fact, professional guidelines do not recommend their detection in routine screening[242]. Other than the limited number of studies and the potential technical challenges of current methodologies[247], the fact is that because microdeletions affect fewer genes, the physical or mental effects are not always predictable. For example, in the case of the microduplication causing the DiGeorge syndrome, the effect can range from severe heart defects and learning difficulties to asymptomatic individuals.

5.12 Future directions

5.12.1 NIPT

This Thesis has mainly focused on the development of a set experimental tools for the manipulation and analysis of the information contained in cfDNA, first through [redacted] and its application in the field of NIPT. Basic questions regarding the applicability of the methods have been answered by means of a reduced number of samples obtained in the context of a pilot study in collaboration with HUD. A larger number of samples is being collected to, on one hand, assess a higher number of chromosomal aneuploidies and to build a larger population of control samples, and, on the other hand, to try to address several questions that remained unanswered in the context of this Thesis.

One important question that needs to be assessed with further detail is the resolution our technique is capable of. For this work we have focused on full chromosome aneuploidies but the fact that [redacted] suggests that we should be able to detect subchromosomal alterations such as microduplications or microdeletions at high resolution. [redacted]

Besides chromosome count, we have identified three main factors that may have a role in determining the final [redacted] and that may explain a considerable proportion of the variability observed [redacted] in our preliminary analyses. The first factor is the inherent chromatin accessibility differences among genomic regions. As it has been repeatedly discussed throughout this work, the extent to which a certain genomic region can be released into plasma depends on its accessibility, which also determines how much is protected from endonucleases and therefore its integrity. The second factor is the GC content of our captured regions: regions with extremely high or low GC content are subject to a negative PCR bias and may end up being poorly covered compared to more intermediate GC regions. [redacted]. Therefore, if we want our [redacted] measures to more faithfully reflect the copy number status of any given genomic location, we have to at least address the correction of our data for GC content and the genetic distance to the probe. These corrections will have a larger impact as we try to increase the resolution of the method [redacted] and should allow the detection of subchromosomal aberrations.

[redacted]

A recent interest expressed by our HUD collaborators involved the detection of signatures in cfDNA that could explain various pregnancy complications, as for instance preeclampsia and spontaneous abortion in apparently healthy pregnancies.

5.12.2 Cancer

[redacted]

[redacted]

5.12.3 [redacted]

Most targeted NGS approaches are directed at finding both single nucleotide variants (SNVs) as well as copy number variants (CNVs). However, the detection of the latter is usually more difficult when only small portions of the genome are analyzed, as is the case of targeted sequencing like WES. Because we have shown that [redacted] is able to detect very small differences in coverage in the context of NIPT, we are currently assessing the use of this capture panel for the identification of germline CNVs, which in theory should be much easier to detect due to their high frequency in blood. To this end, samples previously analyzed in the laboratory by comparative genomic hybridization (CGH) - array will be used as controls for the NGS analysis.

5.12.4 Intellectual property protection

A measure of the success of the research carried out in a company is the creation of new products that can be commercialized, with or without intellectual property protection (IPP) measures. Because of the obvious commercial applications of the methodology described here, qGenomics is taking action in two directions: first, a freedom to operate (FTO) analysis was commissioned to a law firm specializing in IPP. Due to the complexity of the patent landscape surrounding the NIPT field we are currently analyzing the FTO report. This analysis should provide clear information on whether the technology we have developed infringes any patents and therefore, in what countries could be potentially commercialized and under what conditions. Second, we will commission a second analysis aimed at assessing whether the works described here can be patented in part or in their totality. Due to the potential patentability of the methods described here, critical parts of this Thesis are being embargoed.

6. Conclusions

The work involved in the development of this Thesis can be divided in three main parts. The first part is fully experimental and has consisted in the development and optimization of existing tools to allow the production of good quality cfDNA sequencing libraries, from where we have reached the following conclusions:

- The established circuit, conditions and quality controls of sample collection, storage, shipment, plasma separation and extraction allow the obtention of high quality cfDNA samples that can be used for downstream analyses in a timeframe that is feasible in the context of the clinical setting.

- Optimization of each of the steps involved in library preparation has been key to successfully avoid adapter dimer and duplicate formation as revealed by subsequent data processing and analysis and, as a result, to obtain good quality sequencing libraries from small amounts of partially degraded DNA.

- [redacted]

The second part of this Thesis involved sequencing data processing and analysis of the [redacted] as well as several relevant biological factors affecting the amount of information available in cfDNA, from where we concluded that:

- [redacted]

- At least three factors govern the amount of coverage of a specific [redacted] have been detected: its GC content, [redacted], and its vicinity to accessible regions such as TSSs.

Finally, the third part of this Thesis focused on the feasibility of using ~~Alu sequencing~~ in the NIPT context. As a proof-of-concept, we focused on the detection of chromosomal aneuploidies, fetal sex determination and fetal fraction calculation, the main parameters calculated in a NIPT test. From this part, a series of conclusions were extracted:

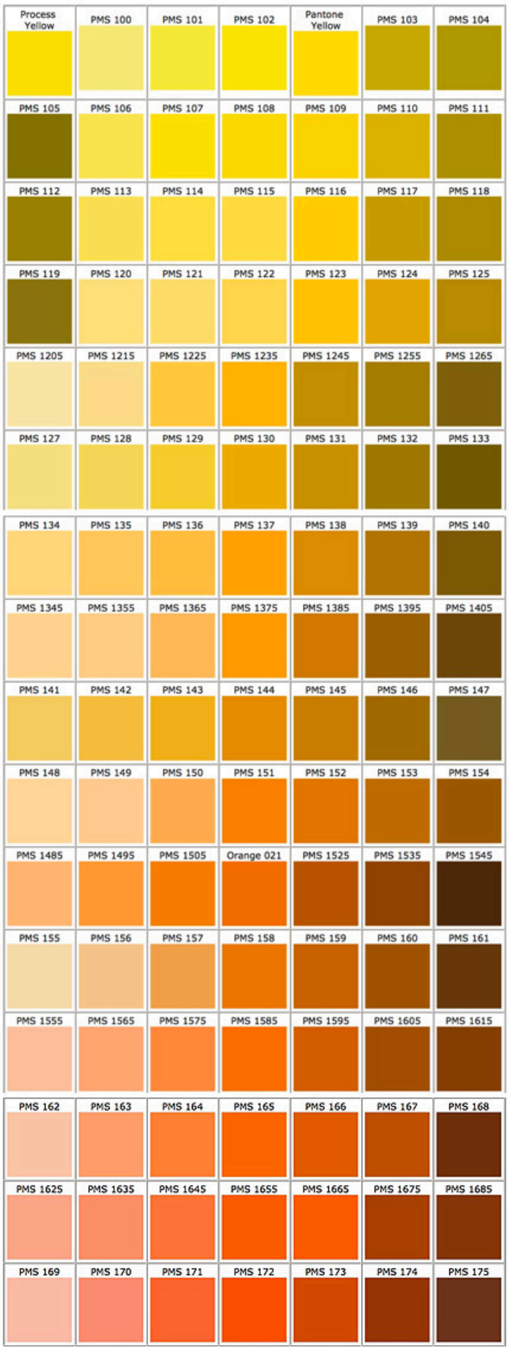
- [redacted]

- [redacted]

- [redacted]

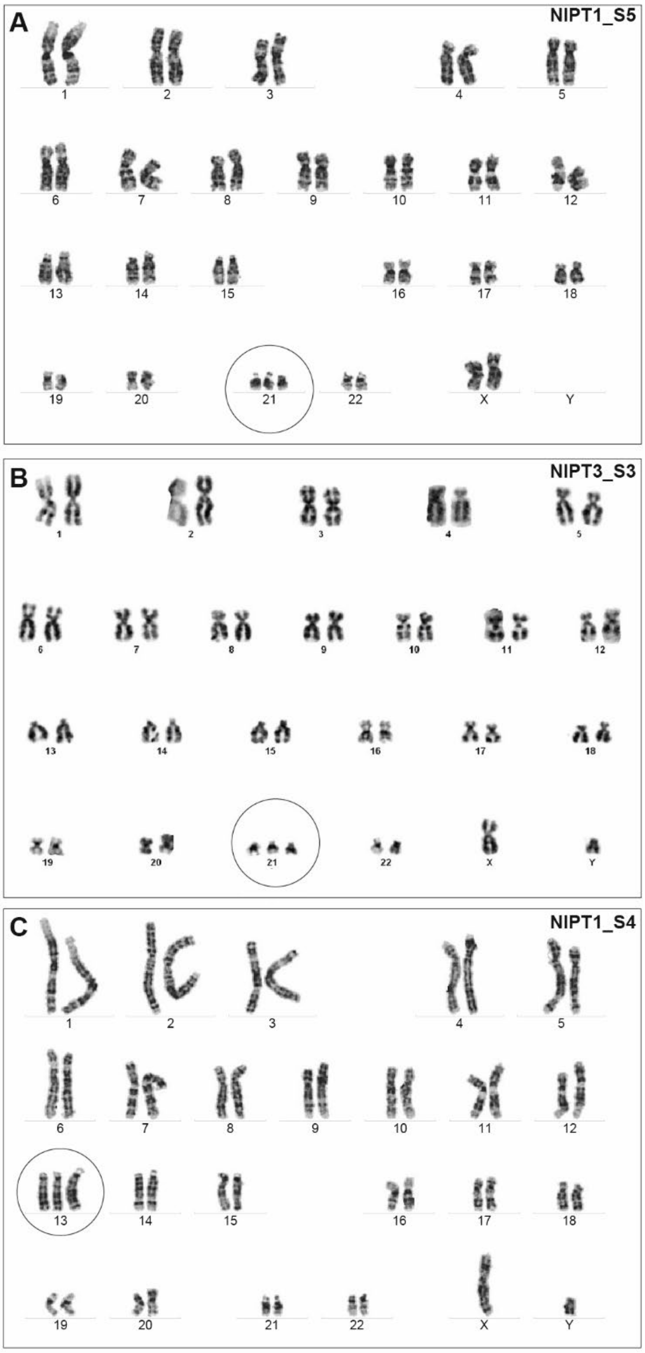
- The analysis of sequencing costs shows that the optimization of several steps in the protocol may contribute to a more cost-effective test that can compete with current market prices.

7. Supplementary Figures

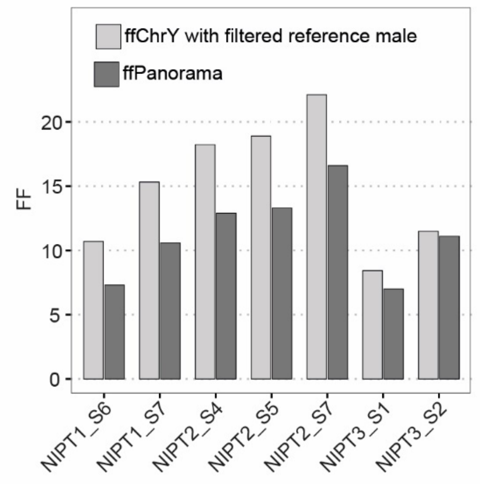


Supplementary Figure 1. **Pantone® color codes used for visual plasma inspection**.

7. Supplementary Figures



Supplementary Figure 2. **Karyotype results for three out of four aneuploid samples**. **A**. NIPT1_S5 (T21). **B**. NIPT3_S3 (T21). **C**. NIPT1_S4 (T13).



Supplementary Figure 3. **ChrY-based FF estimation**. Comparison of ffChrY-based and ffPanorama-based FF values after filtering male reference sample.

[redacted]

[redacted]

7. Supplementary Figures

[redacted]

[redacted]

8. Supplementary Tables

Supplementary Table 1. **Patient information**. **A**. Batch 1 samples. **B**. Batch 2 samples. Gestational age is expressed as weak.day. Abbreviations: body mass index, BMI; vanishing twin, VT; nuchal translucency, NT; not available, NA.; increased, Inc.

A

#	Sample id	Type	Maternal age	BMI	Gestational age	Pregnancy type	Smoker	Sport last 24h	Informed fetal sex	Echographic findings
1	AM14_S1	Control	42	NA	14.1	Simple	NA	NA	Male	NA
2	AM14_S2	T21	39	NA	14	NA	NA	NA	Female	NA
3	AM3_S43	Control	35	NA	17.5	Simple VT	NA	NA	Male VT	NA
4	AM6_S43	Control	46	NA	12	Simple	NA	NA	Female	NA
5	AM7_S1	Control	40	NA	14.2	Simple VT	NA	NA	Female VT	NA
6	NIPT1_S1	Control	43	NA	13	Simple	NA	NA	Female	NA
7	NIPT1_S2	Control	41	NA	12	Simple	NA	NA	Male	NA
8	NIPT1_S3	Control	37	NA	14.6	Simple	NA	NA	Male	NA

B

#	Sample id	Type	Maternal age	BMI	Gestational age	Pregnancy type	Smoker	Sport last 24h	Informed fetal sex	Echographic findings
9	NIPT1_S4	T13	43	20.70	12.4	Simple	no	no	Male	Inc. NT
10	NIPT1_S5	T21	38	22.04	13.0	Simple	no	no	Female	Inc. NT
11	NIPT1_S6	Control	31	32.05	14.1	Simple	no	no	Male	None
12	NIPT1_S7	Control	36	23.78	13.1	Simple	no	no	Male	None
13	NIPT1_S8	Control	40	19.71	12.2	Simple	no	no	Male	Inc. NT
14	NIPT2_S1	Control	34	19.72	11.5	Simple	no	no	Female	None
15	NIPT2_S2	Control	41	22.06	12.4	Simple	no	yes	Female	None
16	NIPT2_S3	Control	38	25.46	13.0	Simple	no	no	Female	None
17	NIPT2_S4	Control	41	24.54	13.3	Simple	no	no	Male	None
18	NIPT2_S5	Control	33	21.36	12.4	Simple	yes	yes	Male	None
19	NIPT2_S6	Control	38	22.83	13.4	Simple	no	no	Female	None
20	NIPT2_S7	Control	38	27.18	13.3	Simple	no	no	Male	None
21	NIPT2_S8	Control	37	21.77	15.1	Simple	no	no	Female	None
22	NIPT3_S1	Control	35	22.65	13.5	Simple	no	no	Male	None
23	NIPT3_S2	Control	30	25.81	16.6	Simple	no	no	Male	None
24	NIPT3_S3	T21	35	17.99	13.0	Simple	no	no	Male	Inc. NT

8. Supplementary Tables

Supplementary Table 2. **Processing parameters for plasma and cfDNA**. **A**. Batch 1 samples. **B**. Batch 2 samples. Days between blood collection and plasma separation (P_{sep} – P_{obt}), plasma volume (Plasma vol), color (PMS code, see **Supplementary Figure 1**), [redacted], and amount of cfDNA extracted per mL of plasma are shown. For two samples, cfDNA concentration was too low (out-of-range, OOR) to be detected through fluorometric methods. Abbreviations: non-available (NA).

A

#	Sample id	P_{sep} - P_{opt}	Plasma vol (mL)	PMS	[redacted]	cfDNA (ng / mL)
1	AM14_S1	1	NA	122	[redacted]	OOR
2	AM14_S2	0	NA	128	[redacted]	OOR
3	AM3_S43	2	NA	123	[redacted]	2.112
4	AM6_S43	2	NA	1215	[redacted]	3.840
5	AM7_S1	2	NA	116	[redacted]	2.560
6	NIPT1_S1	1	NA	122	[redacted]	8.133
7	NIPT1_S2	2	NA	120	[redacted]	3.060
8	NIPT1_S3	1	NA	121	[redacted]	4.205

B

#	Sample id	P_{sep} - P_{opt}	Plasma vol (mL)	PMS	[redacted]	cfDNA (ng / mL)
9	NIPT1_S4	0	4.35	128	[redacted]	6.75
10	NIPT1_S5	1	4.6	121	[redacted]	3.51
11	NIPT1_S6	2	3.6	134	[redacted]	7.14
12	NIPT1_S7	2	4	121	[redacted]	5.17
13	NIPT1_S8	1	4	122	[redacted]	8.25
14	NIPT2_S1	1	4	134	[redacted]	2.98
15	NIPT2_S2	1	3.6	1215	[redacted]	2.14
16	NIPT2_S3	1	3.6	121	[redacted]	2.38
17	NIPT2_S4	1	4	122	[redacted]	2.13
18	NIPT2_S5	0	3.8	109	[redacted]	1.99
19	NIPT2_S6	1	4.8	127	[redacted]	10.66
20	NIPT2_S7	2	5	116	[redacted]	2.48
21	NIPT2_S8	3	4	1215	[redacted]	2.74
22	NIPT3_S1	2	4.5	116	[redacted]	6.96
23	NIPT3_S2	1	4.5	123	[redacted]	11.10
24	NIPT3_S3	1	4	1215	[redacted]	8.03

Supplementary Table 3. **Genomic coordinates and length of β-globin and DYZ1 amplicons generated by *in silico* PCR**. A total of 2 β-globin amplicons of 102 bp and 16 DYZ1 amplicons of 85 and 2462 bp were obtained.

β-globin amplicons

>chr11:5234329-5234430	102 bp
>chr11:5226917-5227018	102 bp

DYZ1 amplicons

>chrY:27008214+27008298	85 bp
>chrY:27012994+27013078	85 bp
>chrY:27017751+27017835	85 bp
>chrY:25399693+25399777	85 bp
>chrY:25392538+25392622	85 bp
>chrY:25387758+25387842	85 bp
>chrY:25382980+25383064	85 bp
>chrY:26932643-26932727	85 bp
>chrY:26937400-26937484	85 bp
>chrY:26942156-26942240	85 bp
>chrY:25301303-25301387	85 bp
>chrY:25306082-25306166	85 bp
>chrY:26930266-26932727	2462 bp
>chrY:26935023-26937484	2462 bp
>chrY:26939779-26942240	2462 bp
>chrY:25298926-25301387	2462 bp

8. Supplementary Tables

[redacted]

[redacted]

8. Supplementary Tables

[redacted]

[redacted]

8. Supplementary Tables

[redacted]

[redacted]

8. Supplementary Tables

[redacted]

Supplementary Table 8. Percentage of reads mapping to unfiltered chrY (%chrY) and filtered chrY (%chrY$_{filt}$) for each sample.

Sample	Fetal sex	%chrY	%chrY$_{filt}$
NIPT2_S7	male	6.28e-02	2.66e-02
AM14_S1	male	6.20e-02	1.58e-02
NIPT3_S3	male	5.70e-02	2.50e-02
NIPT1_S3	male	5.81e-02	1.69e-02
NIPT1_S7	male	5.64e-02	1.84e-02
NIPT1_S8	male	5.66e-02	1.69e-02
NIPT2_S4	male	5.48e-02	2.20e-02
NIPT2_S5	male	5.48e-02	2.28e-02
NIPT1_S2	male	5.04e-02	1.28e-02
NIPT1_S6	male	4.80e-02	1.29e-02
AM3_S43	male	4.38e-02	1.59e-02
NIPT1_S4	male	4.30e-02	9.83e-03
NIPT3_S2	male	4.09e-02	1.38e-02
NIPT3_S1	male	3.65e-02	1.02e-02
NIPT1_S1	female	3.55e-02	4.33e-04
NIPT1_S5	female	3.43e-02	3.10e-04
NIPT2_S3	female	3.05e-02	3.52e-04
NIPT2_S1	female	2.84e-02	3.55e-04
NIPT2_S2	female	2.76e-02	2.27e-04
NIPT2_S6	female	2.69e-02	2.50e-04
AM7_S1	female	2.37e-02	2.86e-04
AM14_S2	female	2.13e-02	4.91e-04
AM6_S43	female	2.01e-02	2.91e-05
NIPT2_S8	female	2.08e-02	7.77e-05

[redacted]

[redacted]

[redacted]

[redacted]

Supplementary Table 11. Summary of FF estimations based on different approaches. **A**. Female fetuses. **B**. Male fetuses. Only those samples for which paired gDNA was available were included in the estimations. The reference ffPanorama (available for all except aneuploid fetuses), the ffChrY (applicable only to male fetuses) and the [redacted]

A

Sample	Reported fetal sex	ffPanorama	ffChrY	[redacted]
NIPT2_S6	female	6.3	NA	[redacted]
NIPT2_S3	female	10.2	NA	[redacted]
NIPT2_S2	female	13.8	NA	[redacted]
NIPT2_S8	female	14.5	NA	[redacted]
NIPT2_S1	female	18	NA	[redacted]
NIPT1_S5	female	NA	NA	[redacted]

B

Sample	Reported fetal sex	ffPanorama	ffChrY	[redacted]
NIPT3_S2	male	11.1	8.06	[redacted]
NIPT3_S1	male	7	5.92	[redacted]
NIPT1_S6	male	7.3	7.5	[redacted]
NIPT1_S7	male	10.6	10.75	[redacted]
NIPT2_S4	male	12.9	12.79	[redacted]
NIPT2_S5	male	13.3	13.26	[redacted]
NIPT2_S7	male	16.6	15.52	[redacted]
NIPT1_S4	male	NA	5.73	[redacted]
NIPT1_S8	male	NA	9.86	[redacted]
NIPT3_S3	male	NA	14.57	[redacted]

[redacted]

[redacted]

Supplementary document 2. **Informed consent**.

[redacted]

[redacted]

9. Supplementary Documents

[redacted]

10. References

1. Lee S, Huang H, Zelen M. Early detection of disease and scheduling of screening examinations. *Stat Methods Med Res* 2004; **13**: 443–456.

2. Institute for Quality and Efficiency in Health Care (IQWiG). *What happens during a biopsy?* Available from: https://www.ncbi.nlm.nih.gov/books/NBK65083.

3. Agència de Salut Pública de Catalunya. *Protocol de cribratge d'anomalies congènites a Catalunya*. Available from: http://salutpublica.gencat.cat/ca/ambits/promocio_salut/Embaras-part-i-puerperi/Protocol-de-seguiment-de-lembaras/protocol-de-cribratge-prenatal/.

4. Akolekar, R., Beta, J., Picciarelli, G., Ogilvie, C. & D'Antonio, F. Procedure-related risk of miscarriage following amniocentesis and chorionic villus sampling: a systematic review and meta-analysis. *Ultrasound Obst Gyn* 2015; **45,** 16–26.

5. Maxim DL, Niebo R, Utell MJ. Screening tests: a review with examples. *Inhal Toxicol* 2014; **26**: 811–828.

6. Sanger F, Nicklen S, Coulson A. DNA sequencing with chain-terminating inhibitors. *Proc National Acad Sci* 1977; **74**: 5463–5467.

7. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 1953; **171**: 737–738.

8. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860.

9. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931.

10. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014; **30**: 418–426.

11. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biology* 2018; **122**: e59.

12. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc National Acad Sci* 2009; **106**: 19096–19101.

13. Kumar K, Cowley M, Davis R. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* 2019. doi:10.1055/s-0039-1688446.

14. Goodwin S, McPherson JD, McCombie RW. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**: 333.

15. Head SR, Komori KH, LaMere SA, Whisenant T, Nieuwerburgh F, Salomon DR *et al.* Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* 2014; **56**. doi:10.2144/000114133.

16. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. Chromatin Architecture of the Human Genome Gene-Rich Domains Are Enriched in Open Chromatin Fibers. *Cell* 2004; **118**: 555–566.

17. Nagata S, Nagase H, Kawane K, Mukae N, Fukuyama H. Degradation of chromosomal DNA during apoptosis. *Cell Death Differ* 2003; **10**: 4401161.

18. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* 2016; **164**: 57–68.

19. Grunt M, Hillebrand T, Schwarzenbach H. Clinical relevance of size selection of circulating DNA. *Transl Cancer Res* 2018; **7**: S171–S184.

20. Thierry A, Messaoudi ES, Gahan P, Anker P, Stroun M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metast Rev* 2016; **35**: 347–376.

21. Gavrieli Y, Sherman Y, Ben-Sasson S. Identification of programmed cell death in situ via specific labeling of nuclear DNA fragmentation. *J Cell Biology* 1992; **119**: 493–501.

22. Rodriguez J, Tsukiyama T. ATR-like kinase Mec1 facilitates both chromatin accessibility at DNA replication forks and replication fork progression during replication stress. *Gene Dev* 2013; **27**: 74–86.

23. Choi J, Reich CF, Pisetsky DS. The role of macrophages in the in vitro generation of extracellular DNA from apoptotic and necrotic cells. *Immunology* 2005; **115**: 55–62.

24. Holdenrieder S, Stieber P, Bodenmüller H, Fertig G, Fürst H, Schmeller N *et al.* Nucleosomes in Serum as a Marker for Cell Death. *Clin Chem Lab Med* 2001; **39**: 596–605.

25. Gahan PB, Stroun M. The virtosome—a novel cytosolic informative entity and intercellular messenger. *Cell Biochem Funct* 2010; **28**: 529–538.

26. Brinkmann V, Reichard U, Goosmann C, Fauler B, Uhlemann Y, Weiss DS *et al.* Neutrophil Extracellular Traps Kill Bacteria. *Science* 2004; **303**: 1532–1535.

27. Rykova EY, Morozkin ES, Ponomaryova AA, Loseva EM, Zaporozhchenko IA, Cherdyntseva NV *et al.* Cell-free and cell-bound circulating nucleic acid complexes: mechanisms of generation, concentration and content. *Expert Opin Biol Th* 2012; **12**: S141–S153.

28. Mittra I, Nair N, Mishra P. Nucleic acids in circulation: Are they harmful to the host? *J Biosciences* 2012; **37**: 301–312.

29. Jong OG, Balkom BW, Schiffelers RM, Bouten CV, Verhaar MC. Extracellular Vesicles: Potential Roles in Regenerative Medicine. *Front Immunol* 2014; **5**: 608.

30. Tetta C, Ghigo E, Silengo L, Deregibus M, Camussi G. Extracellular vesicles as an emerging mechanism of cell-to-cell communication. *Endocrine* 2013; **44**: 11–19.

31. Chelobanov B, Laktionov P, Vlasov V. Proteins involved in binding and cellular uptake of nucleic acids. *Biochem Mosc* 2006; **71**: 583–596.

32. Vlassov VV, Laktionov PP, Rykova EY. Extracellular nucleic acids. *Bioessays* 2007; **29**: 654–667.

33. Thakur B, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B *et al.* Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Res* 2014; **24**: 766.

34. Khier S, Lohan L. Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Futur Sci Oa* 2018; **4**: FSO295.

35. Gosse C, Pecq LJ, Defrance P, Paoletti C. Initial degradation of deoxyribonucleic acid after injection in mammals. *Cancer Res* 1965; **25**: 877–83.

36. Chused T, Steinberg A, Talal N. The clearance and localization of nucleic acids by New Zealand and normal mice. *Clin Exp Immunol* 1972; **12**: 465–76.

37. Emlen W, Mannik M. Kinetics and mechanisms for removal of circulating single-stranded DNA in mice. *J Exp Medicine* 1978; **147**: 684–699.

38. Haller N, Helmig S, Taenny P, Petry J, Schmidt S, Simon P. Circulating, cell-free DNA as a marker for exercise load in intermittent sports. *Plos One* 2018; **13**: e0191915.

39. Leon SA, Shapiro B, Yaros MJ, D.iaroff. Free DNA in the Serum of Cancer Patients and the Effect of Therapy. *Cancer Research* 1977; **37**: 646–650.

40. Phimister EG, Phillippe M. Cell-free Fetal DNA — A Trigger for Parturition. *New Engl J Medicine* 2014; **370**: 2534–2536.

41. Duvvuri B, Lood C. Cell-Free DNA as a Biomarker in Autoimmune Rheumatic Diseases. *Front Immunol* 2019; **10**: 502.

42. Lui YY, Chik K-W, Chiu RW, Ho C-Y, Lam CW, Lo YM. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem* 2002; **48**: 421–427.

43. Lam JW, Gai W, Sun K, Wong RS, Chan RW, Jiang P *et al.* DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia. *Clin Chem* 2017; **63**: 1614–1623.

44. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheim J, Vaknin-Dembinsky A *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc National Acad Sci* 2016; **113**: E1826–E1834.

45. Bischoff FZ, Lewis DE, Simpson J. Cell-free fetal DNA in maternal blood: kinetics, source and structure. *Hum Reprod Update* 2005; **11**: 59–67.

46. Tjoa M, Cindrova-Davies T, Spasic-Boskovic O, Bianchi DW, Burton GJ. Trophoblastic Oxidative Stress and the Release of Cell-Free Feto-Placental DNA. *Am J Pathology* 2006; **169**: 400–404.

47. Alberry M, Maddocks D, Jones M, Hadi AM, Abdel-Fattah S, Avent N *et al.* Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. *Prenatal Diag* 2007; **27**: 415–418.

48. Sekizawa A, Jimbo M, Saito H, Iwasaki M, Sugito Y, Yukimoto Y *et al.* Increased cell-free fetal DNA in plasma of two women with invasive placenta. *Clin Chem* 2002; **48**: 353–4.

49. Jimbo M, Sekizawa A, Sugito Y, Matsuoka R, Ichizuka K, Saito H *et al.* Placenta Increta: Postpartum Monitoring of Plasma Cell-free Fetal DNA. *Clin Chem* 2003; **49**: 1540–1541.

50. Taglauer ES, Wilkins-Haug L, Bianchi DW. Review: Cell-free fetal DNA in the maternal circulation as an indication of placental health and disease. *Placenta* 2014; **35**: S64–S68.

51. Huppertz B, Kingdom JC. Apoptosis in the Trophoblast—Role of Apoptosis in Placental Morphogenesis. *J Soc Gynecol Invest* 2004; **11**: 353–362.

52. Mandel P, Métais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C R Seances Soc Biol Fil* 1948; **142**: 241–243.

53. Tan E, Schur P, Carr R, Kunkel H. Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. *J Clin Invest* 1966; **45**: 1732–1740.

54. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients. *Oncology* 1989; **5**: 318–322.

55. Vasioukhin V, Anker P, Maurice P, Lyautey J, Lederrey C, Stroun M. Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Brit J Haematol* 1994; **86**: 774–779.

56. Sorenson G, Pribish D, Valone F, Memoli V, Bzik D, Yao S. Soluble Normal and Mutated DNA Sequences from Single-Copy Genes in Human Blood. *Cancer Epidemiology, Biomarkers & Prevention* 1994; **3**: 67–71.

57. Nawroz H, Koch W, Anker P, Stroun M, Sidransky D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nat Med* 1996; **2**: 1035-1037.

58. Nunes SP, Moreira-Barbosa C, Salta S, de Sousa S, Pousa I, Oliveira J *et al.* Cell-Free DNA Methylation of Selected Genes Allows for Early Detection of the Major Cancers in Women. *Cancers* 2018; **10**: 357.

59. Corcoran RB, Chabner BA. Application of Cell-free DNA Analysis to Cancer Treatment. *New Engl J Med* 2018; **379**: 1754–1765.

60. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW *et al.* Presence of fetal DNA in maternal plasma and serum. *Lancet* 1997; **350**: 485–487.

61. Costa J-M, Benachi A, Gautier E. New Strategy for Prenatal Diagnosis of X-Linked Disorders. *New Engl J Medicine* 2002; **346**: 1502–1502.

62. Lo YM, Hjelm MN, Fidler C, Sargent IL, Murphy MF, Chamberlain PF *et al.* Prenatal Diagnosis of Fetal RhD Status by Molecular Analysis of Maternal Plasma. *New Engl J Medicine* 1998; **339**: 1734–1738.

63. Lo YM, Lun FM, Chan AK, Tsui NB, Chong KC, Lau TK *et al.* Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc National Acad Sci* 2007; **104**: 13116–13121.

64. Fan CH, Quake SR. Detection of Aneuploidy with Digital Polymerase Chain Reaction. *Anal Chem* 2007; **79**: 7576–7579.

65. Chiu RW, Chan AK, Gao Y, Lau VY, Zheng W, Leung TY *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc National Acad Sci* 2008; **105**: 20458–20463.

66. Chiu RW, Akolekar R, Zheng YW, Leung TY, Sun H, Chan AK *et al.* Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *Bmj* 2011; **342**: c7401.

67. Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, Ehrich M *et al.* DNA sequencing of maternal plasma to detect Down syndrome: An international clinical validation study. *Genet Med* 2011; **13**: 913.

68. Lo YM, Chan AK, Sun H, Chen EZ, Jiang P, Lun FM *et al.* Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Sci Transl Med* 2010; **2**: 61ra91.

69. Lam K-WG, Jiang P, Liao G, Chan KC, Leung TY, Chiu R *et al.* Noninvasive Prenatal Diagnosis of Monogenic Diseases by Targeted Massively Parallel Sequencing of Maternal Plasma: Application to β-Thalassemia. *Clin Chem* 2012; **58**: 1467–1475.

70. New MI, Tong YK, Yuen T, Jiang P, Pina C, Chan AK *et al.* Noninvasive Prenatal Diagnosis of Congenital Adrenal Hyperplasia Using Cell-Free Fetal DNA in Maternal Plasma. *J Clin Endocrinol Metabolism* 2014; **99**: E1022–E1030.

71. Yu SC, Jiang P, Choy KW, Chan K, Won H-S, Leung WC *et al.* Noninvasive Prenatal Molecular Karyotyping from Maternal Plasma. *Plos One* 2013; **8**: e60968.

72. Grati F, Gomes D, Ferreira JB, Dupont C, Alesi V, Gouas L *et al.* Prevalence of recurrent pathogenic microdeletions and microduplications in over 9500 pregnancies. *Prenatal Diag* 2015; **35**: 801–809.

73. Lun F, Chiu R, Sun K, Leung TY, Jiang P, Chan KC *et al.* Noninvasive Prenatal Methylomic Analysis by Genomewide Bisulfite Sequencing of Maternal Plasma DNA. *Clin Chem* 2013; **59**: 1583–1594.

74. Tsui N, Jiang P, Wong Y, Leung TY, Chan KC, Chiu R *et al.* Maternal Plasma RNA Sequencing for Genome-Wide Transcriptomic Profiling and Identification of Pregnancy-Associated Transcripts. *Clin Chem* 2014; **60**: 954–962.

75. Wang E, Batey A, Struble C, Musci T, Song K, Oliphant A. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. *Prenatal Diag* 2013; **33**: 662–666.

76. Takoudes T, Hamar B. Performance of non-invasive prenatal testing when fetal cell-free DNA is absent. *Ultrasound Obst Gyn* 2015; **45**: 112–112.

77. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. *J Next Generation Sequencing Appl* 2014; **2014**. doi:10.4172/2469-9853.1000106.

78. Beck J, Oellerich M, Schulz U, Schauerte V, Reinhard L, Fuchs U *et al.* Donor-Derived Cell-Free DNA Is a Novel Universal Biomarker for Allograft Rejection in Solid Organ Transplantation. *Transplant P* 2015; **47**: 2400–2403.

79. Truszewska A, Foroncewicz B, Pączek L. The role and diagnostic value of cell-free DNA in systemic lupus erythematosus. *Clin Exp Rheumatol* 2017; **35** : 330–336.

80. Glebova KV, Veiko NN, Nikonov AA, Porokhovnik LN, Kostuyk SV. Cell-free DNA as a biomarker in stroke: Current status, problems and perspectives. *Crit Rev Cl Lab Sci* 2018;**55**: 55–70.

81. Assou S, Aït-Ahmed O, Messaoudi S, Thierry AR, Hamamah S. Non-invasive pre-implantation genetic diagnosis of X-linked disorders. *Med Hypotheses* 2014; **83**: 506–508.

82. Morris J, Wald N, Watt H. Fetal loss in Down syndrome pregnancies. *Prenatal Diag* 1999; **19**: 142–145.

83. Nicolaides KH. Screening for fetal aneuploidies at 11 to 13 weeks. *Prenatal Diag* 2011; **31**: 7–15.

84. Merkatz IR, Nitowsky HM, Macri JN, Johnson WE. An association between low maternal serum α-fetoprotein and fetal chromosomal abnormalities. *Am J Obstet Gynecol* 1984; **148**: 886–894.

85. Canick J, Knight G, Palomaki G, Haddow J, Cuckle H, Wald N. Low second trimester maternal serum unconjugated oestriol in pregnancies with Down's syndrome. *Bjog Int J Obstetrics Gynaecol* 1988; **95**: 330–333.

86. Macri JN, Kasturi RV, Krantz DA, Cook EJ, Moore ND, Young JA *et al.* Maternal serum Down syndrome screening: Free β-protein is a more effective marker than human chorionic gonadotropin. *Am J Obstet Gynecol* 1990; **163**: 1248–1253.

87. Lith VJ, Pratt J, Beekhuis J, Mantingh A. Second-trimester maternal serum immunoreactive inhibin as a marker for fetal Down's syndrome. *Prenatal Diag* 1992; **12**: 801–806.

88. Brambati B, Macintosh M, Teisner B, Maguiness S, Shrimanker K, Lanzani A *et al.* Low maternal serum levels of pregnancy associated plasma protein A (PAPP-A) in the first trimester in association with abnormal fetal karyotype. *Bjog Int J Obstetrics Gynaecol* 1993; **100**: 324–326.

89. Wald N, Hackshaw A, Walters J, Mackinson A, Rodeck C, Chitty L. First and Second Trimester Antenatal Screening for Down's Syndrome: The Results of the Serum, Urine and Ultrasound Screening Study (SURUSS). *J Med Screen* 2003; **10**: 56–104.

90. Wald NJ, Huttly WJ, Hackshaw AK. Antenatal screening for Down's syndrome with the quadruple test. *Lancet* 2003; **361**: 835–836.

91. Krantz D. First-Trimester down syndrome screening using dried blood biochemistry and nuchal translucency. *Obstetrics Gynecol* 2000; **96**: 207–213.

92. Schuchter K, Hafner E, Stangl G, Metzenbauer M, Höfinger D, Philipp K. The first trimester 'combined test' for the detection of Down syndrome pregnancies in 4939 unselected pregnancies. *Prenatal Diag* 2002; **22**: 211–215.

93. Wapner R, Thom E, Simpson J, Pergament E, Silver R, Filkins K *et al.* First-Trimester Screening for Trisomies 21 and 18. *New Engl J Medicine* 2003; **349**: 1405–1413.

94. Nicolaides K, Spencer K, Avgidou K, Faiola S, Falcon O. Multicenter study of first-trimester screening for trisomy 21 in 75 821 pregnancies: results and estimation of the potential impact of individual risk-orientated two-stage first-trimester screening. *Ultrasound Obst Gyn* 2005; **25**: 221–226.

95. Ekelund CK, Jørgensen F, Petersen O, Sundberg K, Tabor A, Group D. Impact of a new national screening policy for Down's syndrome in Denmark: population based cohort study. *Bmj Br Medical J* 2008; **337**: a2547.

96. Kagan K, Etchegaray A, Zhou Y, Wright D, Nicolaides K. Prospective validation of first-trimester combined screening for trisomy 21. *Ultrasound Obst Gyn* 2009; **34**: 14–18.

97. Leung T, Chan L, Law L, Sahota D, Fung T, Leung T *et al.* First trimester combined screening for Trisomy 21 in Hong Kong: outcome of the first 10,000 cases. *J Maternal-fetal Neonatal Medicine* 2009; **22**: 300–304.

98. Borrell A, Casals E, Fortuny A, Farre TM, Gonce A, Sanchez A *et al.* First-trimester screening for trisomy 21 combining biochemistry and ultrasound at individually optimal gestational ages. An interventional study. *Prenatal Diag* 2004; **24**: 541–545.

99. Kagan K, Wright D, Baker A, Sahota D, Nicolaides K. Screening for trisomy 21 by maternal age, fetal nuchal translucency thickness, free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A. *Ultrasound Obst Gyn* 2008; **31**: 618–624.

100. Kirkegaard I, Petersen O, Uldbjerg N, Tørring N. Improved performance of first-trimester combined screening for trisomy 21 with the double test taken before a gestational age of 10 weeks. *Prenatal Diag* 2008; **28**: 839–844.

101. Wright D, Spencer K, K. KK, Tørring N, Petersen O, Christou A *et al.* First-trimester combined screening for trisomy 21 at 7–14 weeks' gestation. *Ultrasound Obst Gyn* 2010; **36**: 404–411.

102. Matias A, Gomes C, Flack N, Montenegro N, Nicolaides K. Screening for chromosomal abnormalities at 10–14 weeks: the role of ductus venosus blood flow. *Ultrasound Obst Gyn* 1998; **12**: 380–384.

103. Cicero S, Curcio P, Papageorghiou A, Sonek J, Nicolaides K. Absence of nasal bone in fetuses with trisomy 21 at 11–14 weeks of gestation: an observational study. *Lancet* 2001; **358**: 1665–1667.

104. Huggon I, DeFigueiredo D, Allan L. Tricuspid regurgitation in the diagnosis of chromosomal anomalies in the fetus at 11–14 weeks of gestation. *Heart* 2003; **89**: 1071–1073.

105. Hyett J, Noble P, Snijders R, Montenegro N, Nicolaides K. Fetal heart rate in trisomy 21 and other chromosomal abnormalities at 10–14 weeks of gestation. *Ultrasound Obst Gyn* 1996; **7**: 239–244.

106. Liao A, Snijders R, Geerts L, Spencer K, Nicolaides K. Fetal heart rate in chromosomally abnormal fetuses. *Ultrasound Obst Gyn* 2000; **16**: 610–613.

107. Papageorghiou AT, Avgidou K, Spencer K, Nix B, Nicolaides KH. Sonographic screening for trisomy 13 at 11 to 13+6 weeks of gestation. *Am J Obstet Gynecol* 2006; **194**: 397–401.

108. Santorum M, Wright D, Syngelaki A, Karagioti N, Nicolaides K. Accuracy of first-trimester combined test in screening for trisomies 21, 18 and 13. *Ultrasound Obstetrics Amp Gynecol* 2017; **49**: 714–720.

109. Kagan KO, Wright D, Valencia C, Maiz N, Nicolaides KH. Screening for trisomies 21, 18 and 13 by maternal age, fetal nuchal translucency, fetal heart rate, free β-hCG and pregnancy-associated plasma protein-A. *Hum Reprod* 2008; **23**: 1968–1975.

110. Gil M, Accurti V, Santacruz B, Plana M, Nicolaides K. Analysis of cell-free DNA in maternal blood in screening for aneuploidies: updated meta-analysis. *Ultrasound Obst Gyn* 2017; **50**: 302–314.

111. Hayden E. Prenatal-screening companies expand scope of DNA tests. *Nat News* 2014; **507**: 19.

112. Sedrak M, Hashad D, Adel H, Azzam A, Elbeltagy N. Use of Free Fetal DNA in Prenatal Noninvasive Detection of Fetal RhD Status and Fetal Gender by Molecular Analysis of Maternal Plasma. *Genet Test Mol Bioma* 2011; **15**: 627–631.

113. Kotsopoulou I, Tsoplou P, Mavrommatis K, Kroupis C. Non-invasive prenatal testing (NIPT): limitations on the way to become diagnosis. *Diagnosis* 2015; **2**: 141–158.

114. Lench N, Barrett A, Fielding S, McKay F, Hill M, Jenkins L *et al.* The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made. *Prenatal Diag* 2013; **33**: 555–562.

115. Ashoor G, Syngelaki A, Poon L, Rezende J, Nicolaides K. Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks' gestation: relation to maternal and fetal characteristics. *Ultrasound Obst Gyn* 2013; **41**: 26–32.

116. Greeley ET, Kessler KA, Vohra N. Clinical Applications of Noninvasive Prenatal Testing. *J Fetal Medicine* 2015; **2**: 11–17.

117. Wong A, Lo YM. Noninvasive fetal genomic, methylomic, and transcriptomic analyses using maternal plasma and clinical implications. *Trends Mol Med* 2015; **21**: 98–108.

118. Leonard S. Current Concepts in Noninvasive Prenatal Screening (NIPS). *J Fetal Medicine* 2017; **4**: 125–130.

119. Vermeesch J, Voet T, Devriendt K. Prenatal and pre-implantation genetic diagnosis. *Nat Rev Genet* 2016; **10**: 643–656.

120. Ehrich M, Deciu C, Zwiefelhofer T, Tynan JA, Cagasan L, Tim R *et al.* Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. *Am J Obstet Gynecol* 2011; **204**: 205.e1-11.

121. Liao G, Lun F, Zheng Y, Chan KC, Leung TY, Lau TK *et al.* Targeted Massively Parallel Sequencing of Maternal Plasma DNA Permits Efficient and Unbiased Detection of Fetal Alleles. *Clin Chem* 2011; **57**: 92–101.

122. Liao GJ, Chan AK, Jiang P, Sun H, Leung TY, Chiu RW *et al.* Noninvasive Prenatal Diagnosis of Fetal Trisomy 21 by Allelic Ratio Analysis Using Targeted Massively Parallel Sequencing of Maternal Plasma DNA. *Plos One* 2012; **7**: e38154.

123. Fan CH, Quake SR. Sensitivity of Noninvasive Prenatal Detection of Fetal Aneuploidy from Maternal Plasma Using Shotgun Sequencing Is Limited Only by Counting Statistics. *Plos One* 2010; **5**: e10439.

124. Curnow KJ, Wilkins-Haug L, Ryan A, Kırkızlar E, Stosic M, Hall MP *et al.* Detection of triploid, molar, and vanishing twin pregnancies by a single-nucleotide polymorphism–based noninvasive prenatal test. *Am J Obstet Gynecol* 2015; **212**: 79.e1-9.

125. Pergament E, Cuckle H, Zimmermann B, Banjevic M, Sigurjonsson S, Ryan A *et al.* Single-Nucleotide Polymorphism–Based Noninvasive Prenatal Screening in a High-Risk and Low-Risk Cohort. *Obstetrics Gynecol* 2014; **124**: 210–218.

126. Fan CH, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc National Acad Sci* 2008; **105**: 16266–16271.

127. Cho E-H. Whole genome sequencing based noninvasive prenatal test. *J Genetic Medicine* 2015; **12**: 61–65.

128. Sirinivasan A, Bianchi DW, Huang H, Sehnert AJ, Rava RP. Noninvasive Detection of Fetal Subchromosome Abnormalities via Deep Sequencing of Maternal Plasma. *Am J Hum Genetics* 2013; **92**: 167–176.

129. Boon EM, Faas BH. Benefits and limitations of whole genome versus targeted approaches for noninvasive prenatal testing for fetal aneuploidies. *Prenatal Diag* 2013; **33**: 563–568.

130. Sparks AB, Wang ET, Struble CA, Barrett W, Stokowski R, McBride C *et al.* Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. *Prenatal Diag* 2012; **32**: 3–9.

131. Koumbaris G, Kypri E, Tsangaras K, Achilleos A, Mina P, Neofytou M *et al.* Cell-Free DNA Analysis of Targeted Genomic Regions in Maternal Plasma for Non-Invasive Prenatal Testing of Trisomy 21, Trisomy 18, Trisomy 13, and Fetal Sex. *Clin Chem* 2016; **62**: 848–855.

132. Neofytou MC, Tsangaras K, Kypri E, Loizides C, Ioannides M, Achilleos A *et al.* Targeted capture enrichment assay for non-invasive prenatal testing of large and small size sub-chromosomal deletions and duplications. *Plos One* 2017; **12**: e0171319.

133. Juneau K, Bogard PE, Huang S, Mohseni M, Wang ET, Ryvkin P *et al.* Microarray-Based Cell-Free DNA Analysis Improves Noninvasive Prenatal Testing. *Fetal Diagn Ther* 2014; **36**: 282–286.

134. Sparks AB, Struble CA, Wang ET, Song K, Oliphant A. Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. *Am J Obstet Gynecol* 2012; **206**: 319.e1-9.

135. Zimmermann B, Hill M, Gemelos G, Demko Z, Banjevic M, Baner J *et al.* Noninvasive prenatal aneuploidy testing of chromosomes 13, 18, 21, X, and Y, using targeted sequencing of polymorphic loci. *Prenatal Diag* 2012; **32**: 1233–1241.

136. Nicolaides K, Syngelaki A, Gil M, Atanasova V, Markova D. Validation of targeted sequencing of single-nucleotide polymorphisms for non-invasive prenatal detection of aneuploidy of chromosomes 13, 18, 21, X, and Y. *Prenatal Diag* 2013; **33**: 575–579.

137. Ryan A, Hunkapiller N, Banjevic M, Vankayalapati N, Fong N, Jinnett KN *et al.* Validation of an Enhanced Version of a Single-Nucleotide Polymorphism-Based Noninvasive Prenatal Test for Detection of Fetal Aneuploidies. *Fetal Diagn Ther* 2016; **40**: 219–223.

138. Rabinowitz M, Savage M, Pettersen B, Sigurjonsson S, Hill M, Zimmermann B. Noninvasive Cell-Free DNA-Based Prenatal Detection of Microdeletions Using Single Nucleotide Polymorphism–Targeted Sequencing. *Obstetrics Gynecol* 2014; **123**: 167S.

139. Agarwal A, Sayres LC, Cho MK, Cook-Deegan R, Chandrasekharan S. Commercial landscape of noninvasive prenatal testing in the United States. *Prenatal Diag* 2013; **33**: 521–531.

140. Cirigliano V, Ordoñez E, Rueda L, Syngelaki A, Nicolaides K. Performance of the neoBona test: a new paired-end massively parallel shotgun sequencing approach for cell-free DNA-based aneuploidy screening. *Ultrasound Obst Gyn* 2017; **49**: 460–464.

141. Peng X, Jiang P. Bioinformatics Approaches for Fetal DNA Fraction Estimation in Noninvasive Prenatal Testing. *Int J Mol Sci* 2017; **18**: 453.

142. Lo YM, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM *et al.* Quantitative Analysis of Fetal DNA in Maternal Plasma and Serum: Implications for Noninvasive Prenatal Diagnosis. *Am J Hum Genetics* 1998; **62**: 768–775.

143. Sun K, Jiang P, Chan AK, Wong J, Cheng YK, Liang RH *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc National Acad Sci* 2015; **112**: E5503–E5512.

144. Poon LL, Leung TN, Lau TK, Chow KC, nnis Lo Y. Differential DNA methylation between fetus and mother as a strategy for detecting fetal DNA in maternal plasma. *Clin Chem* 2002; **48**: 35–41.

145. Chan KC, Zhang J, Hui A, Wong N, Lau TK, Leung TN *et al.* Size Distributions of Maternal and Fetal DNA in Maternal Plasma. *Clin Chem* 2004; **50**: 88–92.

146. Kim SK, Hannum G, Geis J, Tynan J, Hogg G, Zhao C *et al.* Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenatal Diag* 2015; **35**: 810–815.

147. Hartwig T, Ambye L, Sørensen S, Jørgensen F. Discordant non-invasive prenatal testing (NIPT) – a systematic review. *Prenatal Diag* 2017; **37**: 527–539.

148. Dharajiya NG, Grosu DS, Farkas DH, McCullough RM, Almasri E, Sun Y *et al.* Incidental Detection of Maternal Neoplasia in Noninvasive Prenatal Testing. *Clin Chem* 2018; **64**: 329–335.

149. Bianchi DW. Cherchez la femme: maternal incidental findings can explain discordant prenatal cell-free DNA sequencing results. *Genet Med* 2018; **20**: 910–917.

150. Bianchi DW, Chudova D, Sehnert AJ, Bhatt S, Murray K, Prosen TL *et al.* Noninvasive Prenatal Testing and Incidental Detection of Occult Maternal Malignancies. *Obstet Gynecol Surv* 2015; **70**: 744–746.

151. Canick JA, Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE. The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies. *Prenatal Diag* 2013; **33**: 667–674.

152. Bianchi DW, Chiu R. Sequencing of Circulating Cell-free DNA during Pregnancy. *New Engl J Med* 2018; **379**: 464–473.

153. Hui L. Noninvasive prenatal testing for aneuploidy using cell-free DNA – New implications for maternal health. *Obstetric Medicine* 2016; **9**: 148–152.

154. [redacted]

155. [redacted]

156. [redacted]

157. [redacted]

158. [redacted]

159. [redacted]

160. Versteeg, van Schaik B, van Batenburg MF, Roos M, Monajemi R, Caron H *et al.* The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Res* 2003; **13**: 1998–2004.

161. [redacted]

162. [redacted]

163. [redacted]

164. [redacted]

165. [redacted]

166. [redacted]

167. Gahlawat A, Lenhardt J, Witte T, Keitel D, Kaufhold A, Maass KK *et al.* Evaluation of Storage Tubes for Combined Analysis of Circulating Nucleic Acids in Liquid Biopsies. *Int J Mol Sci* 2019; **20**: 704.

168. Fernando M, Chen K, Norton S, Krzyzanowski G, Bourne D, Hunsley B *et al.* A new methodology to preserve the original proportion and integrity of cell-free fetal DNA in maternal plasma during sample processing and storage. *Prenatal Diag* 2010; **30**: 418–424.

[redacted]

170. Sorber L, Zwaenepoel K, Deschoolmeester V, Roeyen G, Lardon F, Rolfo C *et al.* A Comparison of Cell-Free DNA Isolation Kits Isolation and Quantification of Cell-Free DNA in Plasma. *J Mol Diagnostics* 2017; **19**: 162–168.

171. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res* 2018; **47**: D94-D99.

172. [redacted]

173. [redacted]

174. [redacted]

175. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton. J Mol Biol 2000; **302**: 205–217.

176. Roewer L, Amemann J, Spurr N, Grzeschik K-H, Epplen J. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum Genet* 1992; **89**: 389–394.

177. Rahman M, Bashamboo A, Prasad A, Pathak D, Ali S. Organizational Variation of DYZ1 Repeat Sequences on the Human Y Chromosome and Its Diagnostic Potentials. *Dna Cell Biol* 2004; **23**: 561–571.

178. Fazi A, Gobeski B, Foran D. Development of two highly sensitive forensic sex determination assays based on human DYZ1 and Alu repetitive DNA elements. *Electrophoresis* 2014; **35**: 3028–3035.

179. Wataganara T, Chen AY, LeShane ES, llivan L, Borgatta L, Bianchi DW *et al.* Cell-free fetal DNA levels in maternal plasma after elective first trimester termination of pregnancy. *Fertil Steril* 2004; **81**: 638–644.

180. Vasavda N, Ulug P, Kondaveeti S, Ramasamy K, Sugai T, Cheung G *et al.* Circulating DNA: a potential marker of sickle cell crisis. *Brit J Haematol* 2007; **139**: 331–336.

181. llumina, Inc. 2017. *bcl2fastq2 Conversion Software v2.20.*

182. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J* 2011; **17**: 10–12.

183. Kent W. The Human Genome Browser at UCSC. *Genome Res* 2002; **12**: 996–1006.

184. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010; **26**: 589–595.

185. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

186. Broad Institute. *Picard*. Available from: http://broadinstitute.github.io/picard/ [Accessed 11th July 2019]

187. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.

188. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.

189. Zhou X, Wang T. Using the Wash U Epigenome Browser to Examine Genome-Wide Sequencing Data. *Curr Protoc* 2012; **40**: 10.10.1-10.10.14.

190. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU Epigenome Browser update 2019. *Nucleic Acids Res* 2019; **47**: W158–W165.

191. [redacted]

192. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010; **38**: 576–589.

193. R Core Team (2014). R: A Language and environament for statistical computing. R Foundation for Satistical Computing. Vienna, Austria. Available from: http://www.R-project.org/.

194. International Human Genome Sequencing Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57.

195. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* 2017; **46**: D794-D801.

196. Bcbio-nextgen. Available from: https://github.com/bcbio/bcbio-nextgen

197. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 2016; **44**: e108–e108.

198. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009; **25**: 2283–2285.

199. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *Bmc Genomics* 2015; **16**: S1.

200. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie RW. A comparative analysis of exome capture. *Genome Biol* 2011; **12**: R97.

201. Mayer R, Brero A, von Hase J, Schroeder T, Cremer T, Dietzel S. Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *Bmc Cell Biol* 2005; **6**: 44.

202. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E *et al.* The accessible chromatin landscape of the human genome. *Nature* 2012; **489**: 75.

203. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm J-P, Marth C, Höfer T *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* 2012; **19**: 1185.

204. Sekizawa A, Kondo T, Iwasaki M, Watanabe A, Jimbo M, Saito H *et al.* Accuracy of fetal gender determination by analysis of DNA in maternal plasma. *Clin Chem* 2001; **47**: 1856–8.

205. Devaney SA, Palomaki GE, Scott JA, Bianchi DW. Noninvasive Fetal Sex Determination Using Cell-Free Fetal DNA: A Systematic Review and Meta-analysis. *Jama* 2011; **306**: 627–636.

206. Miura K, Higashijima A, Shimada T, Miura S, Yamasaki K, Abe S *et al.* Clinical application of fetal sex determination using cell-free fetal DNA in pregnant carriers of X-linked genetic disorders. *J Hum Genet* 2011; **56**: 296

207. Lyon MF. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 1961*;* **190**: 372–373.

208. Boumil RM, Lee JT. Forty years of decoding the silence in X-chromosome inactivation. *Human Molecular Genetics* 2001; **10**: 2225–2232.

209. Ahn J, Lee J. X chromosome: X inactivation. *Nature Education* 2008; **1**:24

210. Sun K, Jiang P, Cheng S, Cheng T, Wong J, Wong V *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res* 2019; **29**: 418–427.

211. Xu X-P, Gan H-Y, Li F-X, Tian Q, Zhang J, Liang R-L *et al.* A Method to Quantify Cell-Free Fetal DNA Fraction in Maternal Plasma Using Next Generation Sequencing: Its Application in Non-Invasive Prenatal Chromosomal Aneuploidy Detection. *Plos One* 2016; **11**: e0146997.

212. Balslev-Harder M, Richter SR, Kjærgaard S, Johansen P. Correlation between Z score, fetal fraction, and sequencing reads in non-invasive prenatal testing. *Prenatal Diag* 2017; **37**: 943–945.

213. Zhou Y, Zhu Z, Gao Y, Yuan Y, Guo Y, Zhou L *et al.* Effects of Maternal and Fetal Characteristics on Cell-Free Fetal DNA Fraction in Maternal Plasma. *Reprod Sci* 2015; **22**: 1429–1435.

214. Fiorentino F, Bono S, Pizzuti F, Mariano M, Polverari A, Duca S *et al.* The importance of determining the limit of detection of non-invasive prenatal testing methods. *Prenatal Diag* 2016; **36**: 304–311.

215. Madhavan D, Wallwiener M, Bents K, Zucknick M, Nees J, Schott S *et al.* Plasma DNA integrity as a biomarker for primary and metastatic breast cancer and potential marker for early diagnosis. *Breast Cancer Res Tr* 2014; **146**: 163–174.

216. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 2015; **43**: e143–e143.

217. [redacted]

218. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011; **12**: 363.

219. Lewis CA, Crayle J, Zhou S, Swanstrom R, Wolfenden R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proc National Acad Sci* 2016; **113**: 8194–8199.

220. Knierim E, Lucke B, Schwarz J, Schuelke M, Seelow D. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *Plos One* 2011; **6**: e28240.

221. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016; **17**: 13.

222. Farrants A-K. Chromatin Immunoprecipitation, Methods and Protocols. *Methods Mol Biology Clifton N J* 2017; **1689**: 77–82.

223. Subramaniyam S, Pulijaal VR, Mathew S. Double and multiple chromosomal aneuploidies in spontaneous abortions: A single institutional experience. *J Hum Reproductive Sci* 2014; **7**: 262–268.

224. Sen P, Shah PP, Nativio R, Berger SL. Epigenetic Mechanisms of Longevity and Aging. *Cell* 2016; **166**: 822–839.

225. Ma J, Cram DS, Zhang J, Shang L, Yang H, Pan H. Birth of a child with trisomy 9 mosaicism syndrome associated with paternal isodisomy 9: case of a positive noninvasive prenatal test result unconfirmed by invasive prenatal diagnosis. *Mol Cytogenet* 2015; **8**: 44.

226. Liang D, Lv W, Wang H, Xu L, Liu J, Li H *et al.* Non-invasive prenatal testing of fetal whole chromosome aneuploidy by massively parallel sequencing. *Prenatal Diag* 2013; **33**: 409–415.

227. Beek DM, Straver R, Weiss MM, Boon EM, Amsterdam K, Oudejans CB *et al.* Comparing methods for fetal fraction determination and quality control of NIPT samples. *Prenatal Diag* 2017; **37**: 769–773.

228. Huu S, Oster M, Uzan S, Chareyre F, Aractingi S, Khosrotehrani K. Maternal neoangiogenesis during pregnancy partly derives from fetal endothelial progenitor cells. *Proc National Acad Sci* 2007; **104**: 1871–1876.

229. Invernizzi P, Biondi M, Battezzati P, Perego F, Selmi C, Cecchini F *et al.* Presence of fetal DNA in maternal plasma decades after pregnancy. *Hum Genet* 2002; **110**: 587–591.

230. Futch T, Spinosa J, Bhatt S, Feo E, Rava RP, Sehnert AJ. Initial clinical laboratory experience in noninvasive prenatal testing for fetal aneuploidy from maternal plasma DNA samples. *Prenatal Diag* 2013; **33**: 569–574.

231. Sargent I. Maternal and fetal immune responses during pregnancy. *Exp Clin Immunogenet* 1993; **10**: 85–102.

232. McCracken SA, Gallery E, Morris JM. Pregnancy-Specific Down-Regulation of NF-κB Expression in T Cells in Humans Is Essential for the Maintenance of the Cytokine Profile Required for Pregnancy Success. *J Immunol* 2004; **172**: 4583–4591.

233. Sood R, Zehnder JL, Druzin ML, Brown PO. Gene expression patterns in human placenta. *Proc National Acad Sci* 2006; **103**: 5478–5483.

234. Cvitic S, Longtine MS, Hackl H, Wagner K, Nelson MD, Desoye G *et al.* The Human Placental Sexome Differs between Trophoblast Epithelium and Villous Vessel Endothelium. *Plos One* 2013; **8**: e79233.

235. Oepkes D, Page-Christiaens LC, Bax CJ, Bekker MN, Bilardo CM, Boon EM *et al.* Trial by Dutch laboratories for evaluation of non-invasive prenatal testing. Part I—clinical impact. *Prenatal Diag* 2016; **36**: 1083–1090.

236. Norton ME, Jacobsson B, Swamy GK, Laurent LC, Ranzini AC, Brar H *et al.* Cell-free DNA Analysis for Noninvasive Examination of Trisomy. *New Engl J Medicine* 2015; **372**: 1589–1597.

237. Bianchi DW, Rava RP, Sehnert AJ. DNA Sequencing versus Standard Prenatal Aneuploidy Screening. *New Engl J Medicine* 2014; **371**: 577–578.

238. Meck JM, Dugan E, Matyakhina L, Aviram A, Trunca C, Pineda-Alvarez D *et al.* Noninvasive prenatal screening for aneuploidy: positive predictive values based on cytogenetic findings. *Am J Obstet Gynecol* 2015; **213**: 214.e1-214.e5.

239. Kater-Kuipers A, Bunnik E, de Beaufort I, Galjaard R. Limits to the scope of non-invasive prenatal testing (NIPT): an analysis of the international ethical framework for prenatal screening and an interview study with Dutch professionals. *Bmc Pregnancy Childb* 2018; **18**: 409.

240. Miranda J, Miño F, Borobio V, Badenas C, Rodriguez-Revenga L, Pauta M *et al.* Can cell-free DNA testing be used in pregnancies with increased nuchal translucency? *Ultrasound Obst Gyn* 2019. doi:10.1002/uog.20397.

241. Souka A, Krampl E, Bakalis S, Heath V, Nicolaides K. Outcome of pregnancy in chromosomally normal fetuses with increased nuchal translucency in the first trimester. *Ultrasound Obst Gyn* 2001; **18**: 9–17.

242. Senat M, Keersmaecker DB, Audibert F, Montcharmont G, Frydman R, Ville Y. Pregnancy outcome in fetuses with increased nuchal translucency and normal karyotype. *Prenatal Diag* 2002; **22**: 345–349.

243. Bilardo C, Müller M, Pajkrt E, Clur S, van Zalen M, Bijlsma E. Increased nuchal translucency thickness and normal karyotype: time for parental reassurance. *Ultrasound Obst Gyn* 2007; **30**: 11–18.

244. Mula R, Goncé A, Bennásar M, Arigita M, Meler E, Nadal A *et al.* Increased nuchal translucency and normal karyotype: perinatal and pediatric outcomes at 2 years of age. *Ultrasound Obst Gyn* 2012; **39**: 34–41.

245. Pergament E, Alamillo C, Sak K, Fiddler M. Genetic assessment following increased nuchal translucency and normal karyotype. *Prenatal Diag* 2011; **31**: 307–310.

246. Gregg AR, Skotko BG, Benkendorf JL, Monaghan KG, Bajaj K, Best RG *et al.* Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American College of Medical Genetics and Genomics. *Genet Med* 2016; **18**: 1056.

247. Lo KK, Karampetsou E, Boustred C, McKay F, Mason S, Hill M *et al.* Limited Clinical Utility of Non-invasive Prenatal Testing for Subchromosomal Abnormalities. *Am J Hum Genetics* 2016; **98**: 34–44.

248. Hayden E. Prenatal-screening companies expand scope of DNA tests. *Nat News* 2014; **507**: 19.

249. Ehrlich M. DNA hypomethylation in cancer cells. *Epigenomics-uk* 2009; **1**: 239–259.