Plasma DNA Sequencing: A Tool for Noninvasive Prenatal Diagnosis and Research into Circulating Nucleic Acids

ZHENG, Wenli

A Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Doctor of Philosophy

in

Chemical Pathology

The Chinese University of Hong Kong

September 2010

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

PLASMA DNA SEQUENCING:

A TOOL FOR NONINVASIVE PRENATAL DIAGNOSIS AND RESEARCH INTO CIRCULATING NUCLEIC ACIDS

Submitted by ZHENG Wenli for thesis submitted for the degree of Doctor of Philosophy in Chemical Pathology at The Chinese University of Hong Kong in July 2010

Noninvasive prenatal detection of fetal chromosomal aneuploidies is a much sought-after goal in fetomaternal medicine. The discovery of fetal DNA in the plasma of pregnant women has offered new opportunities for this purpose. However, the fact that fetal DNA amounts to just a minor fraction of all DNA in maternal plasma makes it challenging for locus-specific DNA assays to detect the small increase in sequences derived from a trisomic chromosome. On the other hand, although the clinical applications of plasma DNA for prenatal diagnosis are expanding rapidly, the biological properties of circulating DNA in plasma remain unclear. Recently, next-generation sequencing technologies have transformed the landscape of biomedical research through the ultra-high-throughput sequence information generated in a single run. Massively parallel sequencing allows us to study plasma DNA at an unprecedented resolution and also precisely detect fetal chromosomal aneuploidies in a locus-independent way.

In the first part of this thesis, two chromosome Y specific genes (*SRY* and *TSPY*) were chosen as the molecular targets to investigate the characteristics of fetal-specific DNA fragments in maternal plasma. By employing the touch down

ligation-mediated PCR coupled with cloning and sequencing, the end property and the fragment species of fetal DNA were studied.

Our group has demonstrated the use of massively parallel sequencing to quantify maternal plasma DNA sequences for the noninvasive prenatal detection of fetal trisomy 21. In the second part of this thesis, the clinical utility of this new sequencing approach was extended to the prenatal detection of fetal trisomy 18 and 13. A region-selection method was developed to minimize the effects of GC content on the diagnostic sensitivity and precision for the prenatal diagnosis of trisomy 13. To facilitate the next-generation sequencing-based maternal plasma DNA analysis for clinical implementation, two measures, i.e., lowering the starting volume of maternal plasma and barcoding multiple maternal plasma samples, were investigated.

The third part of this thesis focuses on the massively parallel paired-end sequencing of plasma DNA. By analyzing millions of sequenced DNA fragments, the biological properties of maternal plasma DNA were elucidated, such as the size distribution of fetal-derived and maternally-contributed DNA molecules and the potential effect of epigenetic modification on DNA fragmentation. Moreover, the plasma DNA from hematopoietic stem cell transplant patients was characterized by paired-end sequencing approach. These sequencing data not only confirmed the predominant hematopoietic origin of cell-free DNA but also revealed the size difference between hematologically-derived and other tissue-derived DNA molecules in plasma.

Taken together, the results presented in this thesis have demonstrated the clinical utility of massively parallel sequencing of maternal plasma DNA and have also provided us a better understanding of the biology of circulating DNA molecules.

摘要

懷孕母體的血漿中存在胎兒遊離的ＤＮＡ，這一發現爲無創性產前診斷開闢了新的途徑。然而來自母體的背景ＤＮＡ干擾了對胎兒的有效檢測，這個問題在檢測三體綜合症胎兒引起的微量增加時尤爲嚴重。與此同時，儘管各種以母體血漿　ＤＮＡ爲基礎的胎兒診斷方法層出不窮，其生物學特性至今仍所知甚少。近年來，新一代高通量測序儀由於具備高準確性和高通量的優點，被廣泛應用到分子生物學和醫學研究中。利用這一高效測序平臺，本論文對血漿ＤＮＡ的基本特性和臨床應用進行了研究。

論文的第一部分應用連接反應介導的聚合酶鏈式反應 (ligation-mediated PCR)和傳統克隆測序相結合的方法，對胎兒特異ＤＮＡ的末端特性和片段種類進行了研究。

初步試驗已經證實對母體血漿ＤＮＡ進行高通量測序能夠實現對胎兒21三體綜合症的精確診斷。論文的第二部分將這一方法擴展至胎兒１８三體綜合症和１３三體綜合症的診斷，並且針對測序平臺的固有偏差，開發了一個序列選擇的方法。此外，本論文還研究了血漿起始用量和多重測序對胎兒21三體綜合症診斷結果的影響，以期簡化實驗步驟和優化實驗方法，從而促進母體血漿ＤＮＡ測序在實際臨床中的應用。

雙端測序技術是另一種測序模式，除了能夠提供序列的染色體分佈資訊之外,還能夠準確推斷出血漿ＤＮＡ的長度信息。論文的第三部分利用這一技術研究了血漿ＤＮＡ的生物學特性，包括母體血漿中胎兒遊離ＤＮＡ和母親遊離ＤＮＡ的長度分佈和ＤＮＡ表觀遺傳特性對血漿ＤＮＡ的影響。我們還對接受造血幹細胞移植的病人的血漿ＤＮＡ進行雙端測序分析，不僅證實了造血細胞是血漿遊離ＤＮＡ的主要來源，而且發現了不同細胞來源的血漿ＤＮＡ在分子長度上的顯著差異。

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Prof Dennis Lo, for introducing me to the exciting field of scientific research. He has demonstrated to me the passion, the wisdom, and the curiosity that a scientist should possess. I thank Prof Lo for his invaluable guidance and inspiring encouragement throughout my PhD training.

Special thanks go to Prof Rossa Chiu for her invaluable advice and patient guidance when I encountered difficulties in my study. Many thanks go to Prof Allen Chan for his guidance in data analyses and warm discussion as well as invaluable statistical suggestions. Prof Hao Sun's extensive bioinformatics knowledge and his help in the sequencing data mining has been motivating.

I wish to thank Prof Nancy Tsui, Dr Virginia Lau and Dr Fiona Lun for providing me with detailed guidance and helpful comments on laboratory work. Many thanks go to Dr Lisa Chan and Ms Yoyo Jin for technical assistance in sample preparation and sequencing processing, Mr Peiyong Jiang and Mr Zhang Chen for computer program compiling, Prof Emily Hung for kind help in blood collection of hematopoietic stem cell transplant recipients, and all the other team members for giving me support and laughter in the laboratory.

I would like to thank Prof Tze Kin Lau and Prof Tak Yeung Leung at the Prince of Wales Hospital, Hong Kong, and Prof Kypros Nicolaides and Dr Ranjit Akolekar at the King's College Hospital, London, UK, for sample recruitment. Tremendous thanks should go to the many pregnant mothers for their donation of blood for our research work.

I wish to thank my dear friends, Karen, Angel, Dana, Kiwi, Chuan and Bo, for their companion, support and encouragement during the past three years.

Lastly, I am grateful to my parents and my sisters, for their love and care.

PUBLICATIONS

Publications arising from this thesis:

1. Chiu, R.W., Chan, K.C., Gao, Y., Lau, V.Y., **Zheng, W.**, Leung, T.Y., Foo, C.H., Xie, B., Tsui, N.B., Lun, F.M., Zee, B.C., Lau, T.K., Cantor, C.R. & Lo, Y.M. (2008). Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A,* Vol. 105, No. 51, pp. 20458-63.

Published conference abstracts:

2. **Zheng, W.**, Lun, F.M., Chan, K.C., Sun H., Chiu, R.W. & Lo, Y.M. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidies by massively parallel sequencing of maternal plasma. Illumina User Group Meeting, Sabah, Malaysia, 6-9 April 2009

3. **Zheng, W.**, Lun, F.M., Chan, K.C., Sun H., Leung, T. Y., Lau, T. K., Chan, Y.S., Jin Y., Lau T.K., Chiu, R.W. & Lo, Y.M. Multiplexed massively parallel sequencing of maternal plasma DNA for noninvasive prenatal diagnosis of trisomy 21, The 6th International Conference on Circulating Nucleic Acids in Plasma and Serum (CNAPS-VI), Hong Kong, 9-11 November, 2009

CONTRIBUTORS

I declare that I am the main contributor to the works described in this thesis. Under Prof Dennis Lo's supervision, I was responsible for the design and development of the assays, experimental processing, data analyses and interpretation, and thesis writing unless otherwise stated below. In Chapter 4, the touch down ligation-mediated PCR assays were originally designed by Dr Virginia Lau. In Chapter 5, plasma DNA was kindly provided by Dr Fiona Lun and the library preparation and sequencing process were performed by Prof Yuan Gao at the Virginia Commonwealth University. The sample preparation and sequencing processing for multiplexed sequencing in Chapter 6 were done together with Dr Fiona Lun, Dr Lisa Chan and Ms Yoyo Jin. Dr Macy Heung, Dr Rebecca Chan, Dr Virginia Lau, Dr Fiona Lun and Ms Yoyo Jin assisted in the Basic Local Alignment Search Tool (BLAST) analysis in Chapter 7. Prof Hao Sun has kindly performed the analysis on the methylation effect on fragment size in Chapter 8. In Chapter 9, the fluorescence *in situ* hybridization and DNA short tandem repeat analyses for peripheral blood chimerism were kindly performed by Prof Emily Hung. All computer programs involved in this thesis were compiled by Prof Hao Sun, Mr Peiyong Jiang and Mr Zhang Chen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALL	acute lymphoblastic
AP	adapter primer
APS	adenosine phosphosulphate
ATP	adenosine triphosphate
beta-globin	beta-hemoglobin
BLAST	Basic Local Alignment Search Tool
bp	base pairs
CEA	carcinoembryonic antigen
CGB	chorionic gonadotropin, beta polypeptide
CGIs	CpG islands
CNA	circulating nucleic acids
CRH	corticotropin releasing hormone
CV	coefficient of variation
CVS	chorionic villus sampling
ddNTP	dideoxynucleotide triphosphate
DEPC	diethylpyrocarbonate
DMR	differentially methylated region
dNTP	deoxynucleotides
EAR	epigenetic allelic ratio
EB	elution buffer
EDTA	ethylenediaminotetraacetic acid
EGG	epigenetic-genetic
ELAND	Efficient Large-Scale Alignment of Nucleotide Databases
FAM	6-carboxyfluorescein
FISH	fluorescence in situ hybridization
FRET	fluorescence resonance energy transfer
FU	fluorescence unit
GA	Genome Analyzer
Gb	giga bases
GE	genome equivalent
GR	genomic representation
GSP	Sequences of gene-specific primers
HBV	hepatitis B virus
HCC	hepatocellular carcinoma
hCG	human chorionic gonadotropin
HIV	human immunodeficiency virus
HLCS	holocarboxylase synthetase
HLM	endogenous retroviral Pol-like sequence
hnRNP-B1	heterogeneous nuclear ribonucleoprotein B1
IPTG	isopropyl-beta-D-thiogalactopyranoside
ISSS	in silico size selection
IUGR	intrauterine growth restriction
kb	kilobases
LM-PCR	ligation-mediated PCR
Mb	megabases

MGB	minor-groove-binding
miRNAs	microRNAs
MSRE	methylation-sensitive restriction endonuclease
mtDNA	mitochondrial DNA
NASS	nucleic acid size selection
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
NT	nuchal translucency
PCR	polymerase chain reaction
PE	paired-end
PLAC4	placenta-specific 4
PPi	pyrophosphate
psi	pounds per square inch
QPCR	real-time quantitative polymerase chain reaction
RASSF1A	ras association (RalGDS/AF-6) domain family member 1A
RCD	relative chromosome dosage
RhD	Rhesus D
RMD	relative mutation dosage
ROC	receiver operating characteristic
SAA	Severe aplastic anaemia;
SBS	sequencing-by-synthesis
SD	standard deviation
SEM	transmission electron microscopy
SERPINB2	serpin peptidase inhibitor, clade B (ovalbumin), membrane 2
SERPINB5	serpin peptidase inhibitor, clade B (ovalbumin), membrane 5
SMRT	single-molecule real-time
SOC	super optimal broth, catabolite repression
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SR	single-read
SRY	Sex-determining Region Y
sstDNA	single-stranded template DNA
STM	scanning tunneling microscopy
T13	trisomy 13
T18	trisomy 18
T21	trisomy 21
T4 PNK	T4 polynucleotide kinase
TAE	Tris-acetate-EDTA
TAMRA	6-carboxytetramethylrhodamine
TE	Tris-EDTA
tSMS	true single molecule sequencing
TSPY	Testis Specific Protein, Y-linked
UNG	uracil N-glycosylase
UT	universal-template
WAS	Wiskott-Aldrich Syndrome
ZFY	Y-chromosome-specific zinc finger protein
ZMW	zero-mode waveguide
βhCG	human chorionic gonadotropin beta-subunit

SECTION I : BACKGROUND

CHAPTER 1: PRENATAL DIAGNOSIS

1.1 Current prenatal diagnosis

1.1.1 The need for noninvasive prenatal diagnosis

Prenatal diagnosis is now an essential part of modern obstetrics. In recent years, there is an increasing demand for prenatal diagnosis, especially for the pregnant women with advancing maternal age in view of the increased incidence of major fetal aneuploidy disorders (Heffner 2004). Conventional prenatal diagnosis of fetal genetic diseases requires obtaining fetal materials for analysis through procedures such as amniocentesis and chorionic villous sampling (CVS). These invasive methods constitute a low but finite risk to the fetus (Mennuti *et al.* 2003). The inherent risk of fetal loss therefore is a major deterrent to couples when they decide whether to opt for invasive prenatal diagnosis. As a result, it has been a long-sought goal for investigators to develop noninvasive approaches of prenatal diagnosis.

1.1.2 Noninvasive alternatives

To stratify pregnant women according to their risk of carrying a fetus affected by chromosomal aneuploidy, several biochemical markers in maternal serum and ultrasonography have been developed (Nicolaides *et al.* 1992; Saller *et al.* 2008). Nonetheless, these approaches are essentially detecting phenotypic features rather than the core pathology. Because of this indirect nature, such methods have a number of limitations, such as a strictly defined gestational age window in which a specific test can be used and the suboptimal sensitivity and specificity profiles (Saller *et al.* 2008). A combination of ultrasound and biochemical screening has shown improvement in risk estimation, but there are a number of false-positive results with

consequently unnecessary invasive procedures (Kirkegaard *et al.* 2008; Saller *et al.* 2008; Stenhouse *et al.* 2004).

The presence of fetal cells in the maternal circulation allows one to noninvasively obtain fetal genetic materials and consequently target the core molecular pathology for prenatal diagnosis (Bianchi 1997; Bianchi *et al.* 1990). However, it is difficult to translate this phenomenon into clinical practice in view of the hindrances as follow. (i) The fetal cell is scarce in maternal circulation. Bianchi *et al.* reported the average concentration of circulating fetal cells in maternal blood was ~1 nucleated fetal-cell DNA equivalent/mL for pregnancies involving karyotypically normal fetuses (Bianchi *et al.* 1997). (ii) Due to the low concentrations of circulating fetal cells, a relatively large volume of blood is needed to enrich such cells for analysis (Bianchi *et al.* 1997). The labor-intensive and costly procedures for fetal cell isolation and enrichment make this technology less feasible for routine clinical application (Hahn *et al.* 2002). (iii) The prolonged persistence of residual fetal cells from previous pregnancies in the maternal circulation (Bianchi *et al.* 1996) can interfere with the current prenatal diagnosis and perhaps causes false-positive results. Therefore, alternative noninvasive sources of fetal genetic materials are needed.

1.2 Circulating fetal DNA for noninvasive prenatal diagnosis

Cell-free fetal DNA in maternal plasma is a noninvasive source of fetal genetic material. Since its discovery, numerous diagnostic applications have emerged. Meanwhile, investigators have made efforts to understand the biological characteristics and mechanisms of this extracellular fetal DNA species.

1.2.1 Historical overview

In 1948, Mandel and Métais reported for the first time the existence of extracellular nucleic acids in human blood (Mandel *et al.* 1948). Later studies revealed the elevation of cell-free nucleic acid levels in the sera of patients suffering from systemic lupus erythematosus (Tan *et al.* 1966), rheumatoid arthritis (Leon *et al.* 1977a) and cancer (Leon *et al.* 1977b). In 1989, Stroun and colleagues reported the finding of the DNA sequences with neoplastic characteristics in the plasma of cancer patients (Stroun *et al.* 1989). The presence of tumor-derived DNA in plasma/serum was further conclusively demonstrated by Sorenson *et al.* and Vasioukhin *et al.* (Sorenson *et al.* 1994; Vasioukhin *et al.* 1994), who successfully detected cancer-derived oncogene mutations in the circulation of cancer patients. Inspired by the pseudo-malignant nature of the placenta, Lo *et al.* searched for the fetal specific DNA in maternal circulation and successfully demonstrated the existence of cell-free fetal DNA in maternal plasma/serum by detecting Y chromosome-specific DNA sequence in pregnant women carrying male fetuses (Lo *et al.* 1997). This discovery has since stimulated many later promising developments in noninvasive prenatal diagnosis by maternal plasma/serum analysis.

1.2.2 Biological characteristics

1.2.2.1 Quantitative aspects of fetal DNA in maternal plasma

The existence of fetal DNA in the plasma and serum of pregnant women was first reported by Lo *et al.* using conventional PCR (Lo *et al.* 1997). Later on, real-time quantitative PCR (QPCR) was used for quantifying fetal-derived sex-determining region Y (*SRY*), a Y chromosome-specific gene. Fetal DNA in maternal plasma were found at surprisingly high concentrations, reaching a mean of 25.4 and 292.2 genome equivalents (GE) per milliliter in early (11-17 weeks) and late (37-43 weeks) pregnancies, respectively (Lo *et al.* 1998b). The fractional concentrations of

circulating fetal DNA amount to 3.4% and 6.2% of the total DNA concentrations in maternal plasma in early and late pregnancies, respectively (Lo *et al.* 1998b). The ratio of fetal to maternal DNA in plasma is almost 1000-fold greater than the amount of DNA derived from circulating fetal nucleated cells (Ariga *et al.* 2001; Lo *et al.* 1998b), suggesting that fetal DNA can be readily detected in maternal plasma/serum and may be a valuable source of material for noninvasive prenatal diagnosis.

The sequential follow-up study of women who conceived via *in vitro* fertilization showed that fetal DNA could be detected in maternal serum at the 7^{th} week of gestation (Lo *et al.* 1998b). In other studies, the detection of fetal DNA in maternal plasma could be achieved as early as the 5^{th} week of gestation (Guibert *et al.* 2003; Honda *et al.* 2002). Early detection of fetal DNA indicates that the release of fetal DNA into maternal circulation is a very early phenomenon during pregnancy, thus making it possible for early prenatal diagnosis by maternal plasma DNA analysis. The fetal DNA concentration increases as pregnancy progresses (Ariga *et al.* 2001; Birch *et al.* 2005; Galbiati *et al.* 2005; Lo *et al.* 1998b). The serial analyses of fetal DNA concentrations in maternal plasma during the late third trimester revealed a sharp increase of fetal DNA in maternal plasma (Chan *et al.* 2003b; Lo *et al.* 1998b), suggesting that circulating fetal DNA concentrations were undergoing a time of flux when delivery is imminent.

Additionally, when high-precision measurement techniques are used, such as digital PCR, the fetal DNA concentrations can be quantified even more accurately. Recently, by the use of microfluidics digital PCR assays, Lun *et al.* have demonstrated that the median fractional concentration of fetal DNA in maternal plasma is ≥ 2 times higher for all 3 trimesters of pregnancy than previously thought (Lun *et al.* 2008a).

1.2.2.2 Post-partum clearance of circulating fetal DNA

In contrast to circulating fetal cells which can persist in maternal blood for years (Bianchi *et al.* 1996), cell-free fetal DNA in maternal plasma DNA exhibits rapid clearance following delivery, with a mean half-life of about 16 minutes (Lo *et al.* 1999c). Most studies agree that cell-free fetal DNA is unlikely to persist in maternal plasma following delivery (Johnson-Hopson *et al.* 2002; Smid *et al.* 2003), except one report from Invernizzi *et al.*, who showed the presence of Y chromosome sequences in 22% of plasma samples from non-pregnant women who previously had carried male fetuses (Invernizzi *et al.* 2002). It has been pointed out that in the study of Invernizzi *et al.*, plasma samples were harvested by only one centrifugation step of 3,000 *g*, which would not result in acellular plasma optimally (Chiu *et al.* 2001). Thus, the positive detection of fetal signals probably arose from the contamination of residual fetal cells persisting from the previous pregnancies in the plasma of non-pregnant women (Bianchi *et al.* 1996). In general, the rapid kinetics of fetal DNA suggests that plasma DNA analysis may be less susceptible to false-positive results and can provide nearly real-time monitoring for the current pregnancy.

Potential mechanisms for circulating DNA clearance include plasma nucleases and hepatic and renal clearance (Lo 2001). Plasma nucleases were found to account for only part of the clearance of plasma fetal DNA (Lo *et al.* 1999c). The renal system is proposed to be one of the mechanisms for the clearance of plasma DNA. An early line of evidence is from Tsumita *et al.*, who reported that > 90% of injected calf thymus DNA was removed from the circulation of mice within 30 min and that the major organ of uptake was the kidney (Tsumita *et al.* 1963). Recently, tumor- and fetal-derived DNA has been found to be present in the urine of cancer patients and pregnant women, respectively (Botezatu *et al.* 2000; Majer *et al.* 2007; Shi *et al.*

2003), indicating that plasma DNA can pass through the glomerular barrier and be excreted in the urine. The liver is another suggested organ for the removal of circulating DNA (Emlen *et al.* 1984). Nelson *et al.* reported that a marked delay in disappearance of fetal DNA in a mother, who was suffering from acute fatty liver disease, with the fetal DNA sequences being detectable from maternal plasma for at least 11 days after delivery (Nelson *et al.* 2001). The impaired fetal DNA clearance in this case is supportive of the notion that hepatic metabolism is another potential mechanism of circulating fetal DNA clearance. The impaired fetal DNA clearance from maternal plasma has also been observed from pregnancies complicated by preeclampsia (Lau *et al.* 2002), which is probably associated with the organ damage of the liver and kidney in the preeclamptic women (Roberts *et al.* 1993).

1.2.2.3 Origin of circulating fetal DNA

The origin and release mechanisms of circulating nucleic acids remain unclear. A sex-mismatched bone marrow transplant model has been employed to demonstrate that the majority of circulating DNA molecules in plasma and serum are hematopoietic in origin (Lui *et al.* 2002). The nonhematopoietic tissues, including the heart, the liver, and the kidneys, account for only a minority of the free circulating DNA (Lui *et al.* 2003). Cell death is a likely mechanism of DNA release into the circulation. It has been reported that tumor cell death is associated with the release of tumor-derived circulating DNA (Fournie *et al.* 1995; Giacona *et al.* 1998), suggesting that at least a portion of the DNA in serum and plasma originates from apoptotic and/or necrotic cells. The size patterns of DNA fragments in the plasma of cancer patients characterized by gel electrophoresis favored the hypothesis that apoptotic and necrotic cells were a major source for plasma DNA in cancer patients (Jahr *et al.* 2001). Moreover, plasma DNA concentrations have been found to

correlate with the levels of circulating nucleosomes, which are the characteristic by-products of apoptosis (Holdenrieder *et al.* 2005). On the other hand, active cellular release of newly synthesized DNA has been suggested to be a possible source of circulating DNA (Anker *et al.* 1975; Anker *et al.* 1976; Stroun *et al.* 2001b; Stroun *et al.* 2000). Cultured cell lines can release newly synthesized DNA spontaneously *in vitro* (Anker *et al.* 1976). Another *in vitro* experiment showed that the ladder pattern of extracellular DNA on electrophoresis, which was regarded as a hallmark of cell apoptosis, could also be observed from the active release of DNA from cells (Stroun *et al.* 2001b). However, whether active DNA release is involved *in vivo* is still being debated.

With regard to the cellular origin of circulating fetal DNA in maternal plasma, many lines of evidence indicate that the placenta is the predominant origin. Firstly, investigators have demonstrated the presence of circulating nucleic acid carrying placental specific signatures, such as the placental DNA with confined placental mosaicism (Masuzaki *et al.* 2004), the DNA with placental epigenetic signatures (Chim *et al.* 2005) and the mRNA of placental origin (Ng *et al.* 2003b). Secondly, circulating fetal DNA is detectable as early as the 5th week of gestation (Guibert *et al.* 2003), at a time after the placenta has been formed but before the fetal circulatory system has been established. Also, cell-free fetal DNA could be detected in maternal plasma from pregnancies with a placenta but without an embryo, in support of the trophoblastic origin of circulating fetal DNA (Alberry *et al.* 2007).

With regard to the release mechanism of circulating fetal DNA, the most accepted hypothesis now is release following placental apoptosis. Apoptosis is a common event in the placenta (Huppertz *et al.* 2001). Fetal DNA concentration in maternal plasma increases as pregnancy advances (Ariga *et al.* 2001; Lo *et al.* 1998b), while

the apoptotic rate increases with the gestational age (Smith *et al.* 1997). Recently, Tjoa *et al.* have demonstrated that oxidative stress on placental tissue can induce apoptosis, perhaps causing a subsequent increase of DNA release (Tjoa *et al.* 2006). Preeclampsia is associated with an increase in placental apoptosis (Ishihara *et al.* 2002) and such abnormality may partially explain the previous observation that pregnant women with preeclampsia tend to have elevated levels of fetal DNA in their plasma (Leung *et al.* 2001; Lo *et al.* 1999b).

1.2.2.4 Structural characteristics of circulating DNA

Circulating cell-free DNA was shown to be double-stranded and of low molecular weight (Anker *et al.* 1975; Stroun *et al.* 1987). The structure analysis of plasma DNA from healthy individuals indicated that the 5' and 3' ends of circulating DNA fragments were rich in C and G, respectively, with the 5' end protruding (Suzuki *et al.* 2008). The molecular size of circulating DNA in healthy individuals and cancer patients has been studied. In healthy individuals, circulating plasma DNA primarily consists of short DNA fragments, with fragments > 500 bp observed to a much lesser extent (Suzuki *et al.* 2008), whereas in patients suffering from neoplastic diseases, the presence of long circulating DNA strands (i.e. increased DNA integrity) has been observed (Jiang *et al.* 2006; Wang *et al.* 2003). By gel electrophoresis and electron microscopy, Giacona *et al.* have shown that circulating DNA exists in lengths which are multiples of ~180 bp, with stronger ladder patterns observed in pancreatic cancer patients (Giacona *et al.* 1998). Additionally, filtration experiments have demonstrated that circulating RNA seems to be associated with particles, whereas DNA is not (Ng *et al.* 2002). This might be attributable to the arrangement of DNA in the form of nucleosomes or apoptotic bodies, which protects them from proteolytic digestion in blood (Bischoff *et al.* 2005).

Yet, few studies have been performed to characterize the fetal DNA molecules in maternal plasma. One valuable information is provided by Chan *et al.*, who reported that the fetal DNA in maternal plasma mainly consisted of short fragments and was significantly shorter than the background maternal DNA (Chan *et al.* 2004).

1.2.2.5 Genomic representation of plasma DNA

Sequence distribution of plasma DNA is of particular interest. Based on a relatively small-scale analysis, Puszyk *et al.* suggested that different DNA sequences were not equally abundant in plasma (Puszyk *et al.* 2009). In a particular study from Stroun *et al.*, it was found that the proportion of *Alu* repeat sequences relative to beta-globin (*HBB*) gene sequences was significantly greater in serum DNA than in lymphocyte DNA (Stroun *et al.* 2001a). To obtain a relatively comprehensive profiling, a number of groups have used conventional cloning and DNA sequencing techniques to study circulating plasma DNA (Suzuki *et al.* 2008; van der Vaart *et al.* 2008). A total of 556 clones of plasma DNA from healthy individuals showed that the number of clones derived from each chromosome was correlated with the chromosomal size (Suzuki *et al.* 2008).

Recently, the complete analyses of genomic representation in circulating DNA from pregnant women and healthy individuals have been achievable by next-generation sequencing (NGS) platforms (Beck *et al.* 2009; Chiu *et al.* 2008; Fan *et al.* 2008). Using the Illumina sequencing platform, Chiu *et al.* (Chiu *et al.* 2008) and Fan *et al.* (Fan *et al.* 2008) have independently demonstrated the feasibility of direct sequencing of maternal plasma DNA. Notably, the plasma DNA sequences from both pregnant women and adult males were shown to evenly distribute across the human genome (Chiu *et al.* 2008). However, both Chiu *et al.* and Fan *et al.* observed a strong bias in the representation of sequences with extreme GC contents for

maternal plasma samples, probably produced by the intrinsic bias of Illumina sequencing platforms (Chiu *et al.* 2008; Fan *et al.* 2008). Additionally, using the 454 sequencing platform, Beck *et al.* found that most classes of sequences that they analyzed (e.g., genes and RNA/DNA coding sequences) did not appear to differ between serum DNA and genomic DNA in apparently healthy individuals, except a slight underrepresentation of chromosome 19 and an overrepresention of *Alu* elements (Beck *et al.* 2009).

1.2.3 Diagnostic applications

1.2.3.1 Qualitative applications

Fetal gender determination

Since the first demonstration of the presence of fetal-derived Y chromosome specific sequences in maternal blood by Lo *et al.* (Lo *et al.* 1997), sensitive and specific assays through cell-free fetal DNA for fetal gender determination have been developed by the use of QPCR techniques (Birch *et al.* 2005; Cremonesi *et al.* 2004; Lo *et al.* 1998b). Such tests are close to 100% accurate (Scheffer *et al.* 2009), thus providing an effective means for fetal sex determination, which is of particular importance in the prenatal diagnosis of X-linked diseases such as hemophilia and in the decision of prenatal therapeutic intervention in congenital adrenal hyperplasia (Costa *et al.* 2002; Rijnders *et al.* 2001; Santacroce *et al.* 2006).

Fetal rhesus blood group genotyping

Antenatal determination of fetal Rhesus D (RhD) status is clinically useful in the management of sensitized RhD-negative pregnant women whose partner is heterozygous for the *RHD* gene (Lo *et al.* 1998a). These mothers, when carrying a

RhD-positive fetus, might have a risk of RhD sensitization (Lo *et al.* 1998a). By cell-free fetal DNA analysis, a number of studies showed close to 100% accuracies in fetal RhD status determination (Gautier *et al.* 2005; Van der Schoot *et al.* 2006). In view of its accuracy and accessibility, this approach of antenatal RhD genotyping has become the first noninvasive maternal plasma DNA-based procedure that is adopted in routine clinical screening (Finning *et al.* 2004; Gautier *et al.* 2005).

Diagnosis of monogenic diseases

Since paternally inherited fetal alleles that are not shared by the maternal genome are distinguishable as fetal-specific in maternal plasma, the detection of the presence of paternally inherited mutations in maternal plasma can be readily applied to the noninvasive diagnosis of paternally inherited monogenic diseases. Such a strategy has been applied to several autosomal dominant diseases, including achondroplasia (Saito *et al.* 2000), myotonic dystrophy (Amicucci *et al.* 2000) and Huntington disease (Gonzalez-Gonzalez *et al.* 2003), and certain autosomal recessive diseases, such as cystic fibrosis (Gonzalez-Gonzalez *et al.* 2002) and congenital adrenal hyperplasia (Rijnders *et al.* 2001). However, due to the high background maternal DNA, it is difficult to diagnose the fetal status from maternal plasma if the mother has an autosomal dominant mutation or if the mother and father are both carriers for the same autosomal recessive mutation (Ding *et al.* 2004). Instead, researches turned to noninvasively exclude the fetal inheritance of autosomal recessive diseases based on the absence of paternally-inherited mutations in maternal plasma. The feasibility of such a strategy has been demonstrated in autosomal recessive conditions such as congenital adrenal hyperplasia (Chiu *et al.* 2002a) and beta-thalassemia major (Chiu *et al.* 2002b; Ding *et al.* 2004). Invasive prenatal diagnosis thus can be avoided in 50% of these pregnancies.

1.2.3.2 Quantitative applications

Pregnancy-related complications

Quantitative aberration of circulating fetal DNA has been found in a number of pregnancy-associated complications, such as preterm labor (Leung *et al.* 1998) and preeclampsia (Lo *et al.* 1999b). Besides, elevation of fetal DNA in maternal plasma has been observed in other pathologic pregnancies, such as intrauterine growth restriction (IUGR) (Smid *et al.* 2001), premature separation of the placenta (Shimada *et al.* 2004), invasive placentation (Sekizawa *et al.* 2002) and ectopic pregnancies (Lazar *et al.* 2006).

In preeclamptic pregnant women, a five-fold increase of plasma fetal DNA has been reported (Lo *et al.* 1999b). This increase may be related to the impaired clearance of fetal DNA from maternal plasma in such subjects (Lau *et al.* 2002). Further studies have demonstrated that the increased fetal DNA levels are associated with the severity of the pathological conditions (Swinkels *et al.* 2002; Zhong *et al.* 2001). Remarkably, the rise of fetal DNA concentration in maternal plasma appears prior to the onset of preeclampsia (Leung *et al.* 2001; Levine *et al.* 2004; Zhong *et al.* 2002), suggesting the possibility to predict the disease by closely monitoring the fetal DNA level in maternal plasma.

Detection of fetal chromosomal aneuploidies

With respect to fetal aneuploidy detection, a moderate elevation of fetal DNA levels was reported in trisomy 21 (Lo *et al.* 1999a; Zhong *et al.* 2000) and trisomy 13 (Wataganara *et al.* 2003), but not in trisomy 18 (Wataganara *et al.* 2003; Zhong *et al.* 2000) pregnancies. In particular, there was a modest two-fold elevation of fetal DNA

levels for trisomy 21 pregnancies (Lo *et al.* 1999a). However, a considerable degree of overlap in the fetal DNA levels between euploid and trisomy pregnancies (Lo *et al.* 1999a) makes such quantitative analysis less promising for the noninvasive prenatal detection of trisomy 21. Moreover, most groups have used loci on the Y chromosome as markers for measuring the circulating fetal DNA levels, thus limiting this approach to male pregnancies only.

1.2.3.3 Universal fetal specific markers

In order to expand the feasibility of maternal plasma analysis for noninvasive prenatal diagnosis, investigators began to search for universal fetal specific markers, which are independent of fetal polymorphism and gender.

Epigenetic modifications

Poon *et al.* demonstrated the first use of a fetal epigenetic marker for maternal plasma detection by targeting a differentially methylated region (DMR) in the human *IGF-H19* locus (Poon *et al.* 2002). Chim *et al.* next identified the differential methylation of the promoter region of *SERPINB5* (serpin peptidase inhibitor, clade B (ovalbumin), membrane 5) gene between maternal blood cells (hypermethylated) and placental tissues (hypomethylated) (Chim *et al.* 2005). As the *SERPINB5* gene is located on chromosome 18, it becomes a potential marker for fetal trisomy 18 detection. Tong *et al.* thus developed an epigenetic allelic ratio (EAR) analysis of the *SERPINB5* gene in maternal plasma for noninvasively identifying fetal trisomy 18 (Tong *et al.* 2006). However, the detection of this marker requires bisulfite conversion which is associated with DNA degradation. Thus, the ideal marker would be hypermethylated in the fetus and hypomethylated in maternal blood cells so that a methylation-sensitive restriction endonuclease (MSRE) could be used to digest away

the maternal sequences (Tong *et al.* 2009). The promoter of the *RASSF1A* (Ras association (RalGDS/AF-6) domain family member 1A) gene on chromosome 3 is one such marker, offering a potentially universal fetal specific marker for quantitative analysis irrespective of the polymorphism and gender of the fetus (Chan *et al.* 2006; Chiu *et al.* 2007). To extend the same approach to the prenatal detection of trisomy 21, extensive searches for chromosome 21 loci bearing differential methylation between placental tissues and maternal blood cells have been conducted (Chim *et al.* 2008a; Old *et al.* 2007; Papageorgiou *et al.* 2009). As a result, another fetal epigenetic marker with hypermethylation pattern, i.e., the putative promoter region of the *HLCS* (holocarboxylase synthetase) gene on chromosome 21, was discovered, and this marker was applied to the triosmy 21 detection by using an epigenetic-genetic (EGG) chromosome-dosage approach (Tong *et al.* 2009).

Circulating fetal RNA

The presence of fetal RNA in maternal plasma was first reported in 2000 (Poon *et al.* 2000). This discovery was surprising, in view of the known lability of RNA. Subsequent studies demonstrated that the unexpected stability of plasma RNA molecules might be related to their association with particulate matter (Ng *et al.* 2002), which might protect them against plasma RNase digestion (Tsui *et al.* 2002). The detection of circulating fetal RNA species in maternal plasma as early as the 4th week of gestation and their rapid clearance after delivery are desirable features which might facilitate the potential clinical use of such markers (Chiu *et al.* 2006; Ng *et al.* 2003b).

Similar to fetal epigenetic markers, plasma RNA markers can in principle be used irrespective of fetal sex and even polymorphisms. In this regard, quantitative aberrations of selected placental mRNA species could be used to monitor the various

pregnancy-associated disorders, such as *CRH* (corticotropin releasing hormone) mRNA in preeclampsia (Ng *et al.* 2003a) and *CGB* (chorionic gonadotropin, beta polypeptide) mRNA in gestational trophoblastic disease (Masuzaki *et al.* 2005). Additionally, Ng *et al.* found that the maternal serum *βhCG* (human chorionic gonadotropin beta-subunit) mRNA concentration was elevated in aneuploid pregnancies (Ng *et al.* 2004). To search for additional fetal mRNA markers, a microarray-based method was used to systematically identify the gene transcripts that were expressed in the placenta but absent in the maternal blood cells (Tsui *et al.* 2004).

A strategy termed the RNA-SNP approach for fetal chromosome dosage assessment was developed by Lo and co-workers to detect fetal trisomy 21 noninvasively (Lo *et al.* 2007b). *PLAC4* (placenta-specific 4) mRNA is fetal derived and is transcribed from chromosome 21. In heterozygous fetuses, the allelic ratio of the SNP on *PLAC4* is theoretically 1:1 in euploid cases but 2:1 or 1:2 in trisomic ones. The sensitivity (90%) and specificity (96.5%) of this single marker test for the noninvasive prenatal detection of trisomy 21 are comparable to many of the currently used multi-modality screening tests. Recently, the same strategy has been extended to the prenatal detection of trisomy 18 using *SERPINB2* mRNA in maternal plasma (Tsui *et al.* 2009). The main disadvantage of the RNA-SNP approach is that it can only be used if the fetus is heterozygous for the tested polymorphism.

Placental-derived microRNA

MicroRNAs (miRNAs) are a class of small single-stranded non-coding RNA species that regulate gene expression at the posttranscriptional level by degrading or blocking translation of mRNA targets (Ambros 2004). MiRNAs play important regulatory roles in a variety of cellular functions and in some diseases, including

cancer (Calin *et al.* 2004; Umbach *et al.* 2008). Circulating miRNAs were detectable in both serum and plasma samples (Mitchell *et al.* 2008) and the levels of particular miRNAs are associated with diverse pathological conditions, thus acting as potential biomarkers for those diseases (Adachi *et al.* 2010; Dijckmeester *et al.* 2009; Liu *et al.* 2010; Vasilescu *et al.* 2009).

MiRNAs of placental origin have been detected and characterized in maternal plasma by Chim *et al.* (Chim *et al.* 2008b). Apart from the stability and physical nature of circulating miRNAs in maternal plasma, the authors also demonstrated the correlation between circulating placental miRNAs and the stage of pregnancy. In another report, Gilad *et al.* showed that placental miRNA levels could be used to distinguish pregnant from nonpregnant women (Gilad *et al.* 2008). A recent work conducted by Luo *et al.* suggested that miRNAs were exported from human placental syncytiotrophoblasts into maternal circulation, where they could target maternal tissues (Luo *et al.* 2009). Yet, more studies are required to elucidate the biological significance of these placental miRNA markers.

1.2.3.4 Fetal DNA enrichment

Fetal cell-free DNA constitutes only a minor fraction of the total circulating DNA in maternal plasma (Lo *et al.* 1998b). Low fetal DNA concentration in maternal plasma has led to false-negative results (Chan *et al.* 2006) and could render quantitative analysis using maternal plasma DNA samples less precise (Lo *et al.* 2007a). Hence, researchers have been investigating methods for circulating fetal DNA enrichment. The finding that fetal-derived DNA molecules are generally shorter than maternal ones (Chan *et al.* 2004) has enabled fetal DNA enrichment by size selection using gel electrophoresis (Li *et al.* 2004b). The size-fractionated cell-free DNA in maternal plasma consequently improved the diagnostic performances of the maternal

plasma-based analyses for achondroplasia (Li, Holzgreve *et al.* 2004) and β-thalassemia (Li, Di Naro *et al.* 2005). However, the gel electrophoresis technique is prone to DNA contamination. A microsystem recently developed by Hahn *et al.* (Hahn *et al.* 2009), which carries out all size separation processing in an integrated environment, may overcome this concern and standardize such processing for fetal DNA enrichment. Instead of physical means, one can facilitate fetal DNA detection by designing short PCR amplicons. For example, Sikora *et al.* developed a novel universal-template (UT) QPCR to detect the short PCR amplicons and obtained almost 1.6-fold more cell-free fetal DNA than the conventional QPCR assay with longer amplicons (Sikora *et al.* 2009).

One can also achieve relative fetal DNA enrichment by suppressing the amount of background maternal DNA. Dhallan and co-workers claimed that formaldehyde could server as a cell stabilizing agent to minimize DNA release from maternal blood cells (Dhallan *et al.* 2004). However, this suppression technique remains questionable as it has not been universally reproducible (Benachi *et al.* 2005; Chinnapapagari *et al.* 2005; Chung *et al.* 2005).

1.2.3.5 Recent development of single molecule counting technologies

The approaches described previously enable the direct detection of fetal chromosomal aneuploidies, but they are only applicable to fetuses with certain genotypes. For example, RNA–SNP and EAR tests are only informative for heterozygous fetuses; thus, multiple markers are required for genetically diverse populations (Chiu *et al.* 2009a). To overcome the requirement for heterozygosity, methods that directly and precisely measure the relative dose of target chromosomes are needed to extract fetal genetic information from maternal plasma DNA analysis despite the low fractional concentration of circulating fetal DNA in maternal plasma.

The emerging single molecule counting techniques, such as digital PCR and massively parallel sequencing, are therefore adopted for maternal plasma DNA analysis.

Digital PCR

In digital PCR, template DNA molecules are amplified in individual wells under limiting-dilution conditions (Pohl *et al.* 2004). By directly counting the positive wells/compartments, the absolute quantification of the original template DNA can be achieved without the use of calibration standards. Additionally, by compartmentalizing individual template DNA molecules, digital PCR enables each template to be analyzed separately without cross-interference. Compared with conventional QPCR, digital quantification is more accurate and precise (Pohl *et al.* 2004).

Relative chromosome dosage analysis by digital PCR

If affected by trisomy, the fetus would release an extra copy of the trisomic chromosome into maternal plasma, thus leading to an overrepresentation of that chromosome compared with other chromosomes. Due to the overwhelming maternal background, high analytical precision is required for the quantification of such overrepresentation from maternal plasma. In this regard, digital PCR seems to be a promising tool in view of its high precision. Lo *et al.* thus developed a digital PCR-based approach for chromosome dose analysis, termed the digital relative chromosome dosage (RCD) approach, in which digital PCR was performed for an amplicon located on chromosome 21 and another amplicon on a reference chromosome (Lo *et al.* 2007a). The authors demonstrated the conceptual application of digital PCR in prenatal diagnosis of trisomy 21 by mixing the placental tissue

DNA and maternal blood cell DNA samples obtained from euploid and trisomy 21 pregnancies. Theoretically, the imbalanced ratio (> 1) of the total number of chromosome 21 amplicons (inclusive of maternal and fetal contributions) to the number of reference chromosome amplicons would reflect the overrepresentation of chromosome 21 and thus indicate a trisomy 21 fetus. However, the degree of overrepresentation relies on the fetal DNA concentration and is smaller at lower fetal DNA concentrations. Therefore, higher numbers of digital PCR analyses are required to achieve an adequate statistical power to determine whether the fetus is affected by trisomy 21 (Fan *et al.* 2007; Lo *et al.* 2007a). In Lo *et al.*'s work, it was shown that the trisomy 21 fetuses could be accurately detected or excluded in 97% of cases by performing 7680 PCR analyses when the sample contained 25% fetal DNA (Lo *et al.* 2007a). Fan *et al.* also demonstrated the use of digital PCR to detect the overrepresentation of chromosome 21 in trisomy 21 pregnancies using DNA mixtures from cell lines (Fan *et al.* 2007).

Quantification of fetal DNA in maternal plasma by digtal PCR

Previous quantification of fetal DNA levels in maternal plasma relied heavily on the conventional QPCR (Galbiati *et al.* 2005; Lo *et al.* 1998b). Lun *et al.* employed the digital PCR to reassess the fetal DNA concentration in maternal plasma and found ≧ 2 times higher fractional concentrations of fetal DNA in maternal plasma than previously thought (Lun *et al.* 2008a). Such a finding not only helps establish a normative range for future maternal plasma analysis, but also demonstrates the superior precison of digital PCR.

Relative mutation dosage analysis by digital PCR

As discussed above, due to the high background maternal DNA in maternal plasma, the prenatal diagnosis of monogenic diseases from maternal plasma is confined to the detection of the presence or absence of paternally inherited mutations. Recently, a digital PCR-based relative mutation dosage (RMD) approach developed by Lun *et al.* allows one to determine if the fetus has inherited the maternal mutant allele depite the high background maternal DNA (Lun *et al.* 2008b). Digital RMD determines if the mutant and wildtype alleles in maternal plasma are in allelic balance or imbalance (Lun *et al.* 2008b). Allelic balance is expected when the fetal genotype is identical to that of the mother (i.e., heterozygous); while allelic imbalance occurs if the fetus is homozygous for the normal allele (resulting in the overrepresentation of the wildtype allele in maternal plasma) or for the mutant allele (leading to an overrepresentation of the mutant allele in maternal plasma). This application is most clinically relevant for pregnant women who are heterozygous for a given mutation (Lun *et al.* 2008b; Zimmermann *et al.* 2008). Digital RMD is similar to digital RCD (Fan *et al.* 2007) in that the allelic ratio between the mutant and wildtype alleles is determined by counting both the maternal and fetal contributions. Thus, the fractional fetal DNA concentration directly influences the expected extent of allelic imbalance when the fetus is homozygous for either allele. A maternal plasma sample with a lower fetal DNA concentration consequently requires many more digital PCR analyses with larger plasma volume as starting material (Lun *et al.* 2008b). On the basis of the size difference between maternal and fetal DNA molecules (Chan *et al.* 2004), Lun *et al.* developed a digital nucleic acid size selection (NASS) strategy that enriches the fetal DNA without additional plasma sampling or experimental time (Lun *et al.* 2008b). During NASS analysis, only wells showing the presence of short DNA molecules were counted for RMD assessment. As a result, the combination of digital NASS and RMD enabled fetal genotyping to be achieved in cases in which RMD alone would be insufficient (Lun *et al.* 2008b). With this method, noninvasive

prenatal diagnosis of autosomal recessive monogenic diseases with paternally or maternally contributed can be achieved.

Massively parallel maternal plasma DNA sequencing

Although the digital PCR approach is conceptually feasible for the noninvasive prenatal diagnosis of fetal chromosomal aneuploidy, the low fetal DNA fraction in maternal plasma requires the performance of thousands of PCRs to generate a ratio with statistical confidence (Lo *et al.* 2007a). Differing from the digital PCR in which the quantitative comparison is carried out between specific loci, massively parallel sequencing of maternal plasma DNA is performed in a locus-independent manner. In the latter scenario, multiple loci alongside a chromosome would contribute to the quantitative analysis, thereby building a locus-independent single molecule counting method. The rationale of massively parallel maternal plasma DNA sequencing for fetal trisomy 21 detection is shown in Figure 1.1. Briefly, when a woman is pregnant with a trisomy 21 fetus, an overrepresentation of the fractional concentration of chromosome 21 sequences in her plasma is expected. If a random representative portion of DNA fragments from a maternal plasma sample is sequenced, the frequency distribution of the chromosomal origin of the sequenced DNA fragments should reflect the genomic representation of the original maternal plasma sample. In a trisomy 21 pregnancy, an increased proportion of chromosome 21 sequences in relation to the total sequenced reads could be observed when compared with the euploid pregnancies (Chiu *et al.* 2008). Millions of sequence reads obtained per sample would enable a highly precise estimation of the proportion of chromosome 21 sequences; hence, its overrepresentation in triosmy 21 pregnancies can be robustly detected.

Both Chiu *et al.* (Chiu *et al.* 2008) and Fan *et al.* (Fan *et al.* 2008) demonstrated the use of massively parallel maternal plasma DNA sequencing for the noninvasive prenatal diagnosis of trisomy 21 on the Illumina Genome Analyzer (GA) platform. Later, Chiu *et al.* adopted another NGS platform, i.e., the SOLiD system from Applied Biosystems, to achieve the noninvasive detection of trisomy 21 using the same analytical approach (Chiu *et al.* 2009b). These studies have opened a new avenue for assessing fetal aneuploidy precisely and provided a foundation for NGS-based analysis of cell-free DNA in both pathologic and physiological states (Lo *et al.* 2009). However, it is worth noting that the genomic representations of chromosomes 18 and 13, which are relevant for trisomy 18 and trisomy 13, respectively, cannot be measured as precisely as chromosome 21 (Chiu *et al.* 2008; Chiu *et al.* 2009b; Fan *et al.* 2008). Hence, further studies are required to either optimize the current analytical method or develop alternative NGS-based approaches.

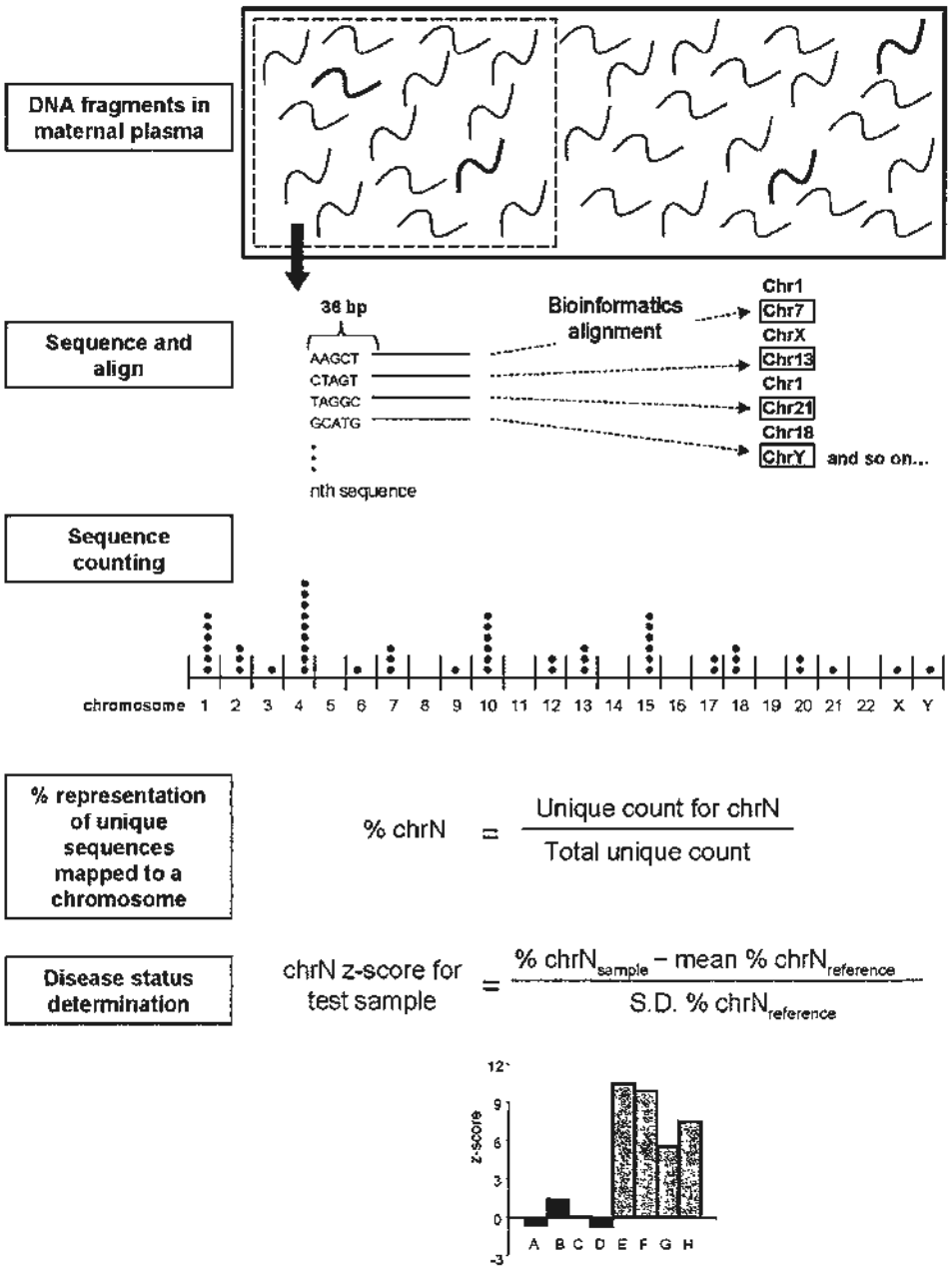


Figure 1.1 Illustration of massively parallel maternal plasma DNA sequencing for fetal chromosomal aneuploidy detection.

Fetal DNA (red lines) coexists with a high background of maternal DNA (black lines) in maternal plasma. A representative profile of maternal plasma DNA is obtained after sequencing. Plasma DNA molecule with 36 bp sequenced for one end can be sorted to its chromosomal origin by alignment against human reference genome. Chromosomal representation is expressed as a percentage of the total unique sequence reads. *Z*-scores of a potentially trisomic chromosome are expected to be higher for pregnancies with a trisomic fetus (green bars) than for those with a euploid fetus (blue bars).

Figure extracted from Chiu *et al.*, 2008. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105:** 20458-20463.

CHAPTER 2: NEXT-GENERATION SEQUENCING TECHNOLOGIES

The advent of DNA sequencing has significantly accelerated biological research and discovery in the past few decades. From the earlier Sanger dideoxy sequencing to the current NGS, along with the upcoming third-generation sequencing, the fast, cheap and accurate sequencing technologies have become indispensable tools in today's biological research.

2.1 Sanger dideoxy sequencing

Sanger dideoxy sequencing or chain-terminator sequencing (Sanger *et al.* 1975; Sanger *et al.* 1977) with the subsequent modifications to it (Cohen *et al.* 1988; Huang *et al.* 1992; Madabhushi 1998; Prober *et al.* 1987; Smith *et al.* 1986), is regarded as the "first-generation" sequencing technique. Sanger sequencing has remained the most commonly used DNA sequencing technique over the past three decades.

Figure 2.1 illustrates the principle of Sanger dideoxy sequencing. The key of this technique is the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. These ddNTPs lack 3' hydroxyl group required for the formation of a phosphodiester bond between two nucleotides and thus prevent a DNA strand from further elongation (Sanger *et al.* 1977). The DNA sample is divided into four separate sequencing reactions, containing DNA templates, DNA primers, DNA polymerases, and four deoxynucleotides (dNTPs), with either the primer or one of the dNTPs radioactively labeled. In each reaction, a particular ddNTP is added to

obtain a collection of single-stranded fragments which end in the ddNTP and differ in varying length. Gel electrophoresis is then used to separate the fragments with each of the four reactions run in one of four individual lanes. By knowing the relative positions of the different bands among the four lanes on the gel, the base sequence can be deduced.

The chain-termination method had become the method of choice for DNA sequencing, owing to its relative ease and reliability. However, the separation of the DNA strands by electrophoresis is a time-consuming step and the sequencing processing is performed manually, both limiting its speed of nucleotide sequence generation. Besides, it requires the use of radioisotopes, which are hazardous, costly and unstable (Smith *et al.* 1986). Hence, a lot of effort had been put to improve the earlier version of Sanger dideoxy sequencing on the basis of the same principle as the chain-termination method (Cohen *et al.* 1988; Huang *et al.* 1992; Madabhushi 1998; Prober *et al.* 1987; Smith *et al.* 1986). Dye-terminator sequencing, in which the fluorescently labeled ddNTPs or primers are utilized and laser-based detection of fluorescence is used to read off the nucleotide sequence (Prober *et al.* 1987; Smith *et al.* 1986), has made DNA sequencing faster, more accurate and more automated.

With the continuous protocol and instrument refinements, such as the development of capillary electrophoresis and the modification of DNA polymerase enzyme and dyes, sequencing efficiency has increased with reduction in error rate and increase in read length (Metzker 2005). As a consequence, the automated dye-terminator sequencing platform (mainly from Applied Biosystems) had been leading the way in sequencing technology in the following 20 years and enabled geneticists to accomplish sequencing projects of large genomes, among which the best known one is the Human Genome Sequencing achieved independently by the International

Human Genome Sequencing Consortium and Celera Genomics (Consortium 2004; Lander *et al.* 2001; Venter *et al.* 2001).

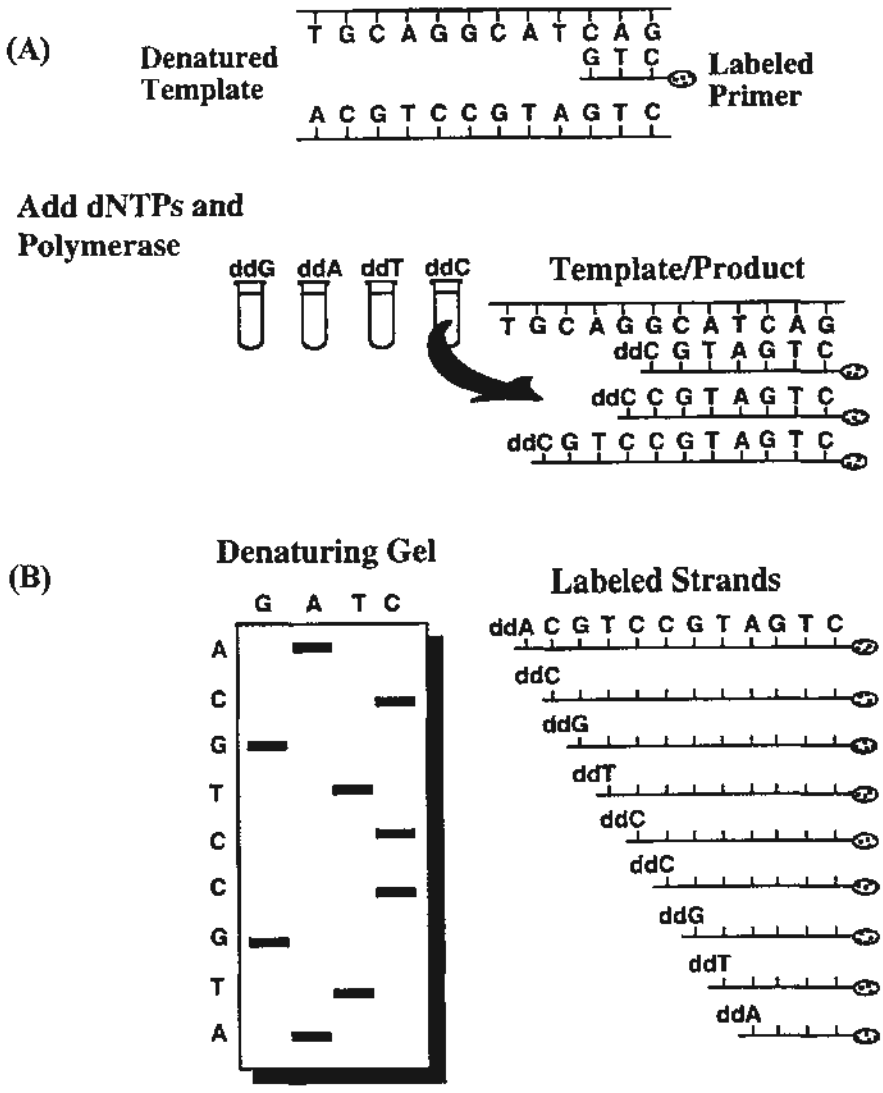


Figure 2.1 Sanger dideoxy chain termination sequencing.

(A) exemplifies how the polymerase reaction takes place in the ddCTP tube. The polymerase extends the labeled primer, randomly incorporating either a normal dCTP base or a ddCTP base. Once a ddCTP is inserted, the DNA strand extension terminates. The amount of ddNTP added is small enough (~1% of total dNTP) so that termination will occur only occasionally. In this way, all possible DNA fragments will be produced in varying length. The length of each terminated strand represents the relative distance from the modified base to the primer. ***(B)*** shows electrophoretic separation of the synthesized and labeled DNA fragments in each of the four reaction tubes above (ddG, ddA, ddT, and ddC) in individual lanes. The bands on the gel represent the respective fragments shown on the right. The original template (given on the left margin of the sequencing gel) is then deduced by reading gel from bottom to top.

Figure extracted from http://www.nwfsc.noaa.gov/publications/techmemos/tm17/

2.2 Next-generation sequencing

Despite the technical improvement, Sanger sequencing is limited by its relatively slow speed and relatively high cost as well as time-consuming operations, resulting from the necessity to separate elongated fragments by size before scanning and the need to produce amplified DNA fragments which is usually achieved by cloning into bacterial hosts (Morozova *et al.* 2008). These obstacles thus render it less applicable for large-scale sequencing projects, such as resequencing large numbers of human genomes. To meet the greater demand for sequence information acquisition with an increased speed and reduced costs, investigators have been developing entirely new strategies for DNA sequencing. The Roche/454 FLX Genome Sequencer, Illumina/Solexa GA and Applied Biosystems/SOLiD^{TM} System, representing the earliest NGS technologies/platforms, have emerged to partially supplant the automated Sanger sequencing in view of their ability to produce an enormous volume of data relatively cheaply.

2.2.1 454 sequencing technology

The 454 system is the first NGS platform available as a commercial product (Margulies *et al.* 2005; Schuster 2008). The 454 technology depends on an emulsion PCR followed by parallel pyrosequencing of the clonally amplified beads in a picotiter plate (Figure 2.2). Emulsion PCR is a highly efficient clonal amplification process performed in an oil-aqueous emulsion (Williams *et al.* 2006), which can circumvent the cloning requirement for preparing the templates for Sanger sequencing. In the emulsion PCR, a droplet acts as an individual amplification reactor which contains a primer-coated bead, a DNA fragment and other necessary components for PCR, producing 10^7 clonal copies of a unique DNA template per bead (Margulies *et al.* 2005). Once the emulsion is broken, beads not carrying any

amplified DNA are removed in an enrichment process, while the templates-containing bead is subsequently distributed to a picotiter plate (Margulies *et al.* 2005). The use of the picotiter plate allows hundreds of thousands of sequencing reactions to be performed in parallel, thereby massively increasing the sequencing throughput (Margulies *et al.* 2005).

The clonally-amplified templates on bead are analyzed using a pyrosequencing reaction. Pyrosequencing is a sequencing-by-synthesis (SBS) technique that measures the release of inorganic pyrophosphate (PPi) by chemiluminescence (Nyren *et al.* 1993; Ronaghi *et al.* 1996; Ronaghi *et al.* 1998). As the single-stranded DNA fragments on the beads have been amplified using general tags, a general primer is annealed and directs the elongation towards the bead. One exclusive dNTP is added per cycle. When the complementary nucleotide is incorporated, the release of PPi is detected as emitted photons through a series of enzymatic steps (Margulies *et al.* 2005). The sequence of DNA template is determined from a "pyrogram," which corresponds to the order of nucleotides that have been incorporated (Margulies *et al.* 2005). After each cycle, the excess nucleotide is degraded by apyrase, and the cyclic sequencing is repeated. As the chemiluminescent signal intensity is proportional to the number of incorporated nucleotides, the pyrosequencing approach is prone to errors for the estimation of the length of homopolymeric sequence stretches (Morozova *et al.* 2008).

Currently, the average read length per sample from 454 sequencing is over 400 bp in the latest instruments, the GS FLX Titanium (and the average read length is ~250 bp in the GS FLX system) (http://www.454.com/). Approximately 1.2 million wells will give unique sequence reads of 400 bp, on average generating less than 500 megabases (Mb) in one single run (Pettersson *et al.* 2009). The increased speed and

reduced costs for sequence data generation enable this new sequencing method to surpass traditional capillary sequencing in terms of sequencing large genomes. Whole-genome sequencing has been performed on bacterial genomes in single runs to demonstrate the efficiency of the 454 sequencing platform (Margulies *et al.* 2005). In 2008, using its technology, the company (454 Life Sciences, Roche) reported the DNA sequence of a diploid genome of a single individual, James D. Watson, which is the first genome sequenced by the NGS technology and therefore a pilot for the future challenges of personalized genome sequencing (Wheeler *et al.* 2008).

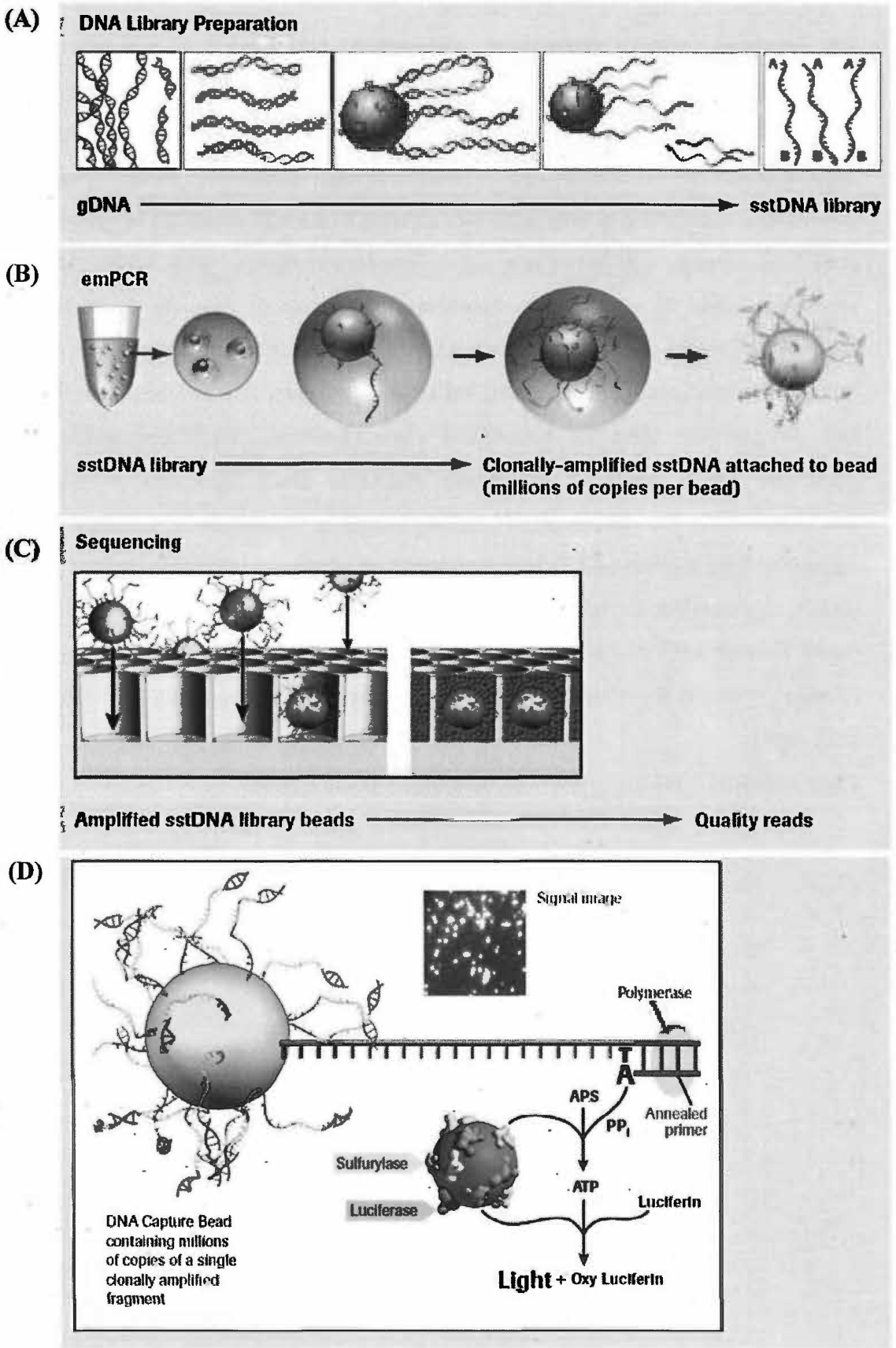


Figure 2.2 454 sequencing workflow.

(A) Template DNA is fragmented, end-repaired and ligated to adapters. These adapters provide priming sequences for both amplification and sequencing of the sample-library fragments. Adapter B contains a 5'-biotin tag that enables immobilization of the adapter-ligated fragments onto streptavidin coated beads. The non-biotinylated strand is then released and used as a single-stranded template DNA (sstDNA) library. ***(B)*** Emulsion PCR involves sstDNA with DNA-capturing sepharose beads in an emulsion containing thousands of droplets. Each bead is captured within its own micro-reactor where PCR amplification occurs. This results in bead-immobilized, clonally amplified DNA fragments. ***(C)*** The beads are loaded into the picotiter plate for subsequent pyrosequencing. ***(D)*** Incorporation of a nucleotide by a series of enzymatic reactions: Each incorporation event is accompanied by the release of PPi; ATP sulfurylase subsequently converts PPi to adenosine triphosphate (ATP) in the presence of adenosine phosphosulphate (APS); finally, the enzyme luciferase (together with D-luciferin and oxygen) can use the newly formed ATP to emit light. Another enzyme, apyrase, is used for degradation of unincorporated dNTPs as well as to stop the reaction by degrading ATP (Ronaghi *et al.* 1998).

Figure adapted from http://www.roche-applied-science.com/

2.2.2 Illumina sequencing technology

In the Illumina sequencing approach, single DNA molecules are attached to a flat surface for bridge amplification *in situ* (Adessi *et al.* 2000; Fedurco *et al.* 2006) and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides (Turcatti *et al.* 2008). The work flow of the Illumina sequencing approach is shown in Figure 2.3. The DNA templates are processed to form single-stranded, adapter-ligated DNA fragments and then added to the surface of a glass flow cell by the use of a microfluidic cluster station (Bentley *et al.* 2008). Each flow cell is divided into eight separate lanes, and the interior surfaces have covalently attached oligonucleotides complementary to the adaptor sequences that are ligated to the fragments during library construction (Bentley *et al.* 2008). Repeated cycles of isothermal bridge amplification generate more than 10 million colony clusters per lane, each containing approximately 1000 copies and with a diameter of about 1 μm (Pettersson *et al.* 2009).

DNA clusters are sequenced by synthetic extension using a set of four 3′-modified reversible terminators, each labeled with a different removable fluorophore (Bentley *et al.* 2008). The use of these nucleotides allows completion of the incorporation without risk of over-incorporation and also enables addition of all four nucleotides simultaneously rather than sequentially, minimizing risk of misincorporation (Bentley *et al.* 2008). After each sequencing cycle, the identity of the inserted base for each cluster is determined by laser-induced excitation of the fluorophores and imaging (Bentley *et al.* 2008). Subsequently, the fluorophores are cleaved off and terminator bases are activated, allowing another round of nucleotide incorporation.

Although more effective for sequencing homopolymeric stretches than pyrosequencing (Pettersson *et al.* 2009), an earlier version of the Illumina sequencing

system produces short sequence reads, typically of 36 bp, perhaps resulting from the incomplete incorporation of nucleotides and insufficient removal of reverse terminators or fluorophores (Pettersson *et al.* 2009). With the improved reagents, this approach can sequence 100 bp at each end of fragments (http://www.illumina.com/). Despite having shorter read length than the 454 system, the throughput of the Illumina system is much higher. For instance, more than 150 million raw sequence reads can be generated in each run with the current Illumina GA IIx system, taking approximately 3 days (http://www.illumina.com/). With the recently launched HiSeq 2000 system, some 1 billion raw sequence reads can be generated per run (http://www.illumina.com/).

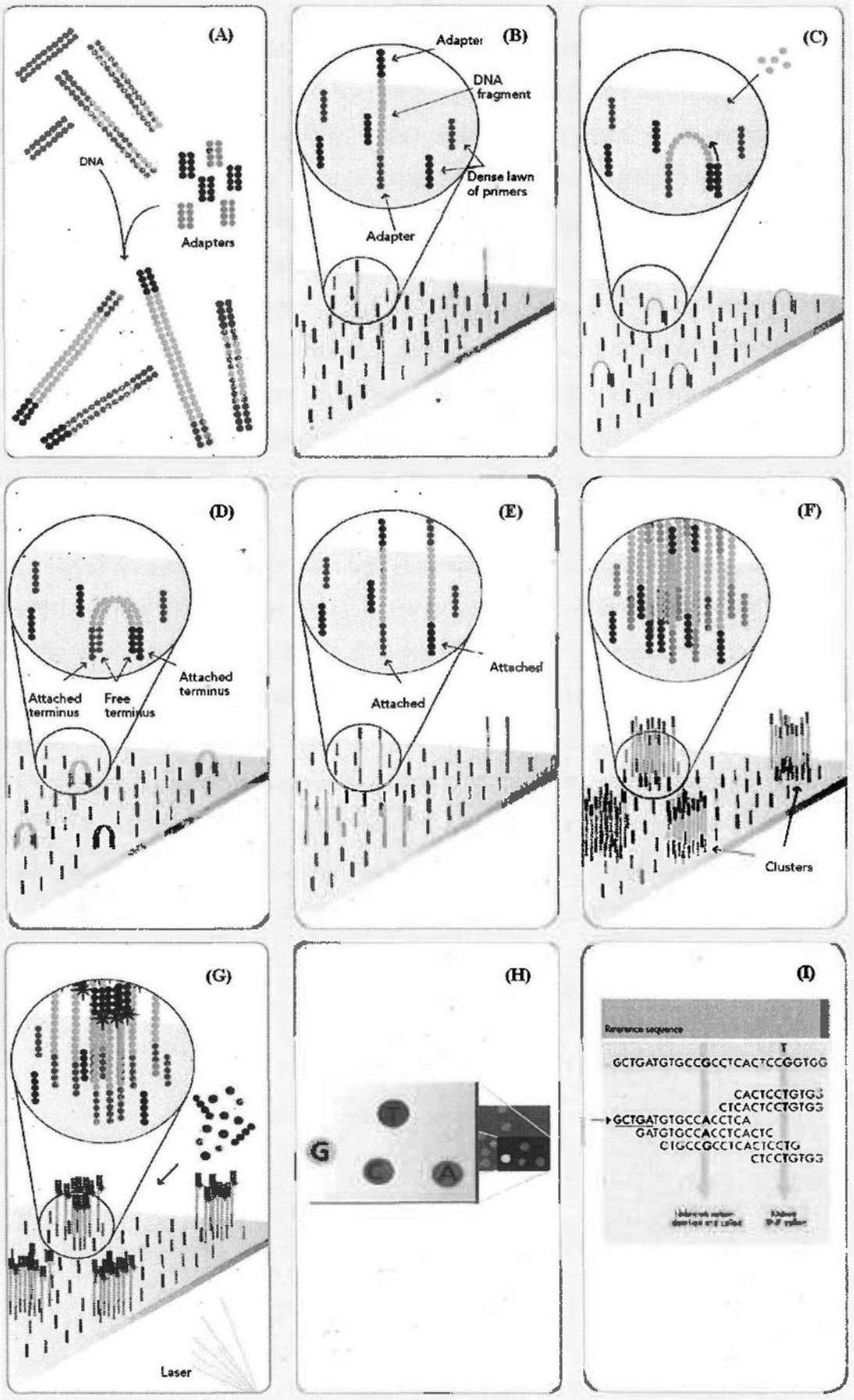


Figure 2.3 Illumina sequencing workflow.

(A) Sample preparation. Genomic DNA is fragmented, end-repaired and ligated with Illumina adapters to construct a DNA library. ***(B-F)*** Cluster generation by bridge amplification on the surface of the flow cell. The flow cell surface is coated with single-stranded oligonucleotides that correspond to the sequences of the adapter for library preparation. A single DNA molecule is captured and attached to the solid support of flow cell via the complementary adapter sequences. Each will anneal to a nearby primer on the surface to form a bridge. After elongation and denaturation, two strands will be free and fixed on the surface. Repeated cycles of amplification result in a colony cluster. Millions of such clusters will be produced across the flow cell surface. Then, DNA clusters are denatured and annealed with a sequencing primer for the subsequent sequencing. The whole process occurs in an Illumina cluster station, an automated flow cell processor. ***(G-I)*** SBS. A flow cell is then loaded into the GA, where automated cycles of extension and imaging occur. In each sequencing cycle, a set of four 3′-modified reversible terminators, each labeled with a different removable fluorophore, are added to the reaction. A single fluorescent nucleotide is extended for a cluster. The fluorescent emission excited by laser is captured and recorded by high-resolution imaging across the entire flow cell. This sequencing cycle is repeated, with one base at a time, generating a series of images each representing a single base extension for clusters. Base calls are performed with an algorithm that identifies the emission color over time and then the base called-nucleotide sequence is ready for downstream analysis.

Figure adapted from http://www.illumina.com/

2.2.3 SOLiD sequencing technology

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing platform employs sequencing-by-ligation strategy, which is initially developed by Shendure *et al.* along with the resequencing of an *Escherichia coli* genome (Shendure *et al.* 2005). The sample preparation for SOLiD sequencing is similar to the 454 technology in that DNA fragments are ligated to oligonucleotide adapters, attached to beads, and clonally amplified by emulsion PCR (McKernan *et al.* 2009). Beads with clonally amplified template are transferred onto a glass surface, where sequencing is started by annealing a sequencing primer complementary to the adapter at the adapter–template junction (Figure 2.4). In the first ligation-sequencing step, thermostable ligase and interrogation probes that are semi-degenerate 8-mer oligonucleotides labeled with four different fluorescent dyes are present. After annealing to the template sequences, the probe is ligated to the adjacent primer. Fluorescence signals are optically collected and then cleaved from the ligated probes. In the subsequent sequencing steps, interrogation probes are ligated to the 5' phosphate group of the preceding pentamer. Seven to ten cycles of ligation, referred to as a "round," are performed to extend the first primer. The synthesized strand is then stripped and a second round of sequencing is initiated by hybridization with an n-1 positioned universal primer, followed by another round of seven repeated ligation reactions. More rounds are performed, each time with a new primer with a successive offset (n - 2, n - 3, and so on).

Two base encoding utilizes four dyes to encode for 16 possible two base combinations. Through multiple rounds of ligation-based sequencing, each base is interrogated twice by two different dye-labeled probes. The identity of the nucleotide is determined by analyzing the color that results from two successive ligation

reactions (McKernan *et al.* 2009). This two-color query system, also known as color space in the SOLiD sequencing platform, greatly facilitates the discrimination of base-calling errors from true polymorphisms or indel events. For example, a sequencing error would be detected in only one particular ligation reaction, whereas a sequence polymorphism would be detected in both (Morozova *et al.* 2008).

SOLiD sequencing platform also generates short-read sequences, typically from 35 bp to 50 bp. Applied Biosystems claims that the recently released Applied Biosystems SOLiD™ 4 System can generate greater than 1.4 billion reads per run and this number will exceed 5 billion with the SOLiD™ 4hq System upgrade (http://www.appliedbiosystems.com/).

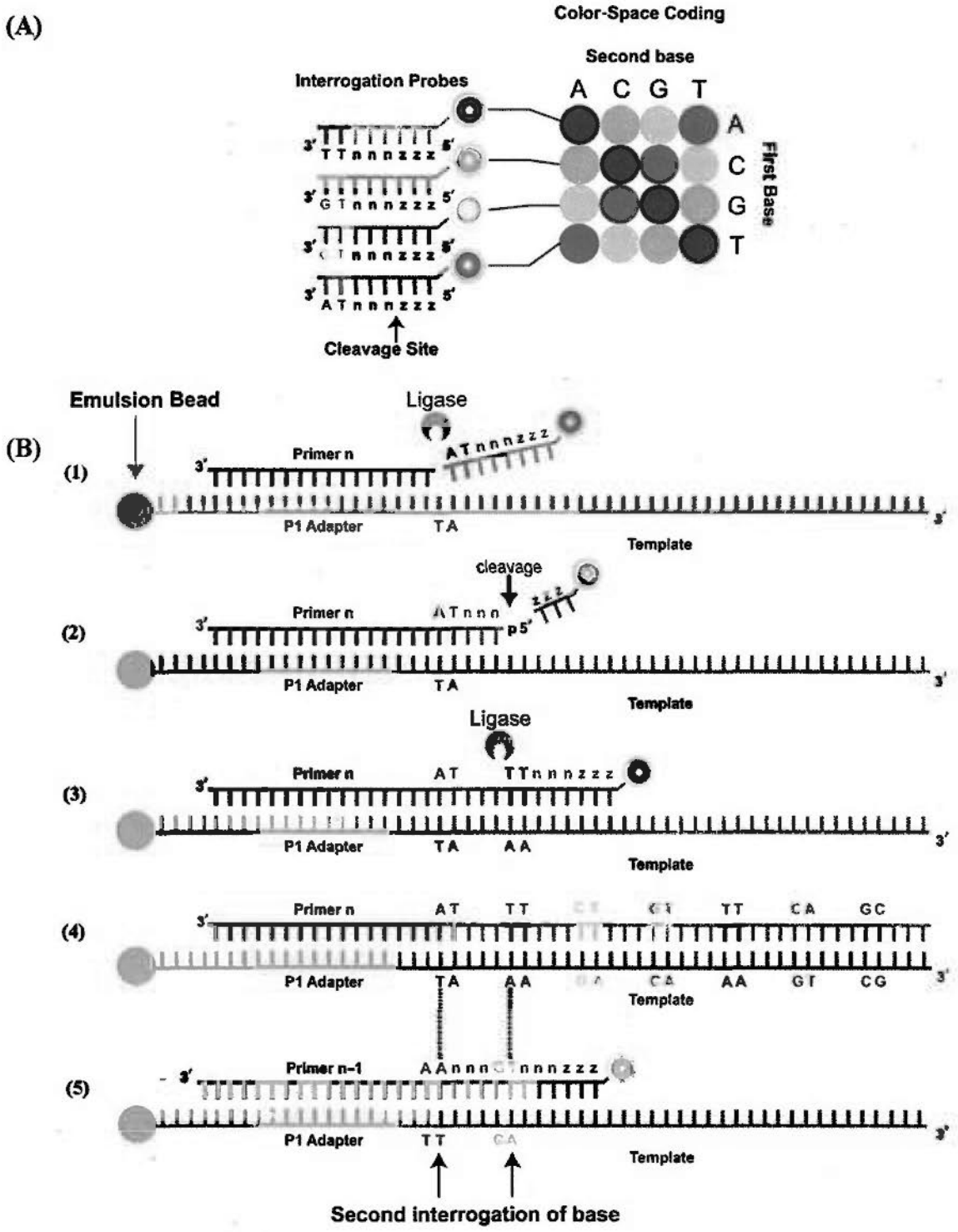


Figure 2.4 Sequencing by ligation in SOLiD sequencing system.

(A) SOLiD color-space coding. There are 16 possible combinations of dinucleotides, with four dinucleotides encoded by one dye. Each 8-mer oligonucleotide (probe), from 3'-to-5' direction, consists of 2 probe-specific bases and 6 degenerate bases (nnnzzz) with one of 4 fluorescent labels at the 5' end. *(B)* Sequencing-by-ligation

reaction. (1) Upon the annealing of a universal primer, a matched 8-mer oligonucleotides hybridizes to the DNA fragment sequence adjacent to the 5′ end of the universal primer. DNA ligase then seals the phosphate backbone. After the ligation step, the fluorescent readout can reflect the possible combination on the 8-mer oligonucleotide. (2) The fluorescent group is removed by chemical cleavage of the three 5' bases, leaving a 5-base ligated probe with 5' end phosphorylated for the next ligation. (3) The next 8-mer oligonucleotide is hybridized and ligated with the template, followed by fluorescent signal scanning. (4) By repeating the step 2 and 3 for six more times, the first round of sequencing-by-ligation is completed. (5) The extended primer is then stripped and a second round of sequencing is initiated by hybridization with an n-1 positioned universal primer, followed by another round of seven repeated ligation reactions.

Figure adapted from Voelkerding, K. V., S. A. Dames, *et al.* (2009). "Next-generation sequencing: from basic research to diagnostics." Clin Chem 55(4): 641-58.

2.3 Third-generation sequencing

Although in its infancy, the third-generation, single-molecule sequencing technology is pretty attractive due to its simplicity and no need of cloning or amplification in sample preparation. It is anticipated that these technologies will largely decrease sequencing time and reduce costs. HeliScope™ Single Molecule Sequencer (Helicos Biosciences) is the first commercial release of a single-molecule sequencing instrument (http://www.helicosbio.com/). Other third-generation sequencing technologies, such as the single-molecule real-time (SMRT) technology (Pacific Biosciences), fluorescence resonance energy transfer (FRET)-based approach (Visigen), nanopore sequencing technology and so on, are expected to reach market soon.

Helicos's true single molecule sequencing (tSMS) technology (Helicos Biosciences), which originates from the work of Braslavsky *et al.* (Braslavsky *et al.* 2003), relies on the SBS strategy using reversible terminators. In this system, a DNA polymerase adds labeled nucleotides to surface-immobilized primer-template duplexes in stepwise fashion, and the asynchronous growth of individual DNA molecules is monitored by fluorescence imaging (Harris *et al.* 2008). Its sequencing capacity was first demonstrated by resequencing a viral genome, the M13 genome (Harris *et al.* 2008). Later, the feasibility of this technology to sequence a human genome was presented (Pushkarev *et al.* 2009) as the first application of the third-generation sequencing technology for human genome sequencing.

Both Pacific Biosciences and VisiGen also adopt the SBS strategy but in a processive and real-time manner. In the approach of Pacific Biosciences, single DNA polymerase molecules are attached to the bottom surface of individual zero-mode waveguide (ZMW) detectors that can identify sequence information while

phospholinked nucleotides carrying their fluorescent labels on the terminal phosphate are being incorporated into the growing primer strand (Eid *et al.* 2008). Remarkably, a recent report has shown that the modified nucleotides in the DNA template, such as N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine, alter polymerase kinetics during SMRT sequencing with unique kinetic signatures, thus permitting discrimination between them in the same DNA sample (Flusberg *et al.* 2010). VisiGen's approach uses a similar nucleotide modification, but the base information is identified by detecting the FRET signals (http://www.visigenbio.com). In this method, the DNA polymerase is modified with a fluorescent donor molecule, while nucleotides are modified with fluorescent acceptor molecules. When a nucleotide is incorporated, the proximity of donor and acceptor fluorophores results in a FRET signal, which is specific for the nucleotide according to its particular fluorophore label (Pushkarev *et al.* 2009).

Nanopore DNA sequencing is a label-free, single-molecule approach that might make inexpensive, rapid DNA sequencing a possibility (Branton *et al.* 2008). Theoretically, when DNA is electronically driven through a nano-scale pore, a change in electrical signals, such as ionic current blockages, transverse tunneling currents, or capacitance, is expected to occur whereby the electrical characteristic for each nucleobase can be easily converted to DNA sequence (Lund *et al.* 2009; Xu *et al.* 2009). The "fifth" base, methylcytosine, with its unique current amplitude characteristic, can be distinguished from the four standard DNA bases (Clarke *et al.* 2009), creating great excitement in the study of epigenetics.

Other single molecule sequencing technologies include the direct sequencing by using transmission electron microscopy (SEM) and the electronic sequencing by scanning tunneling microscopy (STM) (Xu *et al.* 2009).

2.4 Applications of NGS technologies in clinical diagnosis

The emergence of NGS technologies over the recent years has accelerated the drive toward personalized medicine (Chin *et al.* 2008; Ginsburg *et al.* 2009). The NGS-based, in-depth investigations have discovered the underlying genetic causes of diseases (Voelkerding *et al.* 2009). It is also believed that these new DNA sequencing technologies might become powerful tools for timely disease detection, selection of best treatment options and monitoring of the disease course in many diseases (Chin *et al.* 2008).

In comparison with the finished-grade human reference genome (Consortium 2004) and Craig Venter's genome (Levy *et al.* 2007) sequenced by the automated Sanger sequencing platform, the cost of sequencing an entire human genome is dropping dramatically as DNA sequencing technology advances. To date, a number of personal genomes have been fully sequenced using the new sequencing platforms, including the genomes from James Watson (Wheeler *et al.* 2008), a Han Chinese man (Wang *et al.* 2008), a male Yoruba (Bentley *et al.* 2008; McKernan *et al.* 2009), two Korean individuals (Ahn *et al.* 2009; Kim *et al.* 2009), Stephen Quake (Pushkarev *et al.* 2009) and individuals from Southern Africa (Schuster *et al.* 2010). Large-scale whole-genome sequencing projects, such as the Personal Genome Project, Yanhuang Project and 1000 Genomes Project, have been proposed and are ongoing to further increase the freely available sequence data (Gupta 2008), pursuing the eventual goal of personalized genomics and medicine.

Personal genomics obtained by NGS technologies is also being applied to the study of diseases. The usefulness of whole-genome sequencing for genetic diagnosis through rapid identification of alleles that cause disease has been demonstrated. By sequencing the whole genome of the proband in a family with a recessive form of

Charcot-Marie-Tooth disease, Lupski *et al.* have successfully identified clinically relevant variants (Lupski *et al.* 2010). Also, Roach and colleagues have recently demonstrated the family-based complete genome sequencing is valuable for identifying the causative genes of Mendelian disorders, by analyzing the whole-genome sequences of a family of four where two siblings each had two recessive disorders known as Miller syndrome and primary ciliary dyskinesia (Roach *et al.* 2010). Other than whole-genome sequencing, selectively sequencing the complete coding regions (i.e., whole exome) would be an efficient strategy for the understanding of human diseases and the implementation of clinical utility, as protein coding genes constitute only approximately 1% of the human genome but harbor 85% of the mutations with large effects on disease-related traits (Choi *et al.* 2009; Ng *et al.* 2009). The efficacy of exome sequencing for the identification of candidate genes and mutations for diseases and the clinical diagnosis of patients has recently been demonstrated (Choi *et al.* 2009; Ng *et al.* 2009). Importantly, recent effort in sequencing cancer genomes has highlighted the significant impact of NGS on cancer research. For example, Mardis and colleagues presented the whole-genome sequencing of two acute myeloid leukemia cancer genomes along with the matched normal counterparts and identified somatic mutations that might be associated with the disease (Ley *et al.* 2008; Mardis *et al.* 2009).

Owing to the difficulty or impossibility of culturing most pathogenic species, the use of NGS platforms has had a tremendous impact on the study of pathogenic species in clinical samples. Operationally, nucleic acids from a human sample (such as feces, tooth or skin scrapings) is isolated and subjected to sequencing. The non-human sequence reads are assembled to reconstruct segments of the pathogenic genomes carried in the sample, which are further analyzed to identify the presence of known and potentially novel species (Mardis 2009; Voelkerding *et al.* 2009). Such strategy

has been applied to the management of human immunodeficiency virus (HIV) disease (Wang *et al.* 2007a; Wang *et al.* 2007b), the identification of a new arenavirus in transplantation patients (Palacios *et al.* 2008), the analysis of microflora present in the human oral cavity (Keijser *et al.* 2008), and the characterization of faecal microbial communities from lean and obese twins (Turnbaugh *et al.* 2009) and 124 European individuals (Qin *et al.* 2010).

Circulating nucleic acids (CNA) isolated from serum or plasma are increasingly recognized as biomarkers for pregnancy (Lo *et al.* 2007) and cancers (Fleischhacker *et al.* 2007). Recently developed NGS technologies enable the sequence profile of CNA at single molecule resolution (Chiu *et al.* 2008); hence, high-precision noninvasive assessments by the use of CNA become achievable for prenatal diagnosis and cancer screening and monitoring. Chiu *et al.* and Fan *et al.* have independently demonstrated the NGS-based, single molecule counting approach for the detection of fetal chromosomal aneuploidy by massively parallel maternal plasma DNA sequencing (Chiu *et al.* 2008; Fan *et al.* 2008). These studies open a new avenue for assessing fetal aneuploidy and provide a foundation for NGS-based analysis of CNA in other states, e.g. cancer and infectious disease. For example, NGS-based sequence analysis for the serum CNA from ductal breast cancer patients can discriminate the patients with tumor stage I from healthy and nonmalignant disease control with a high sensitivity and specificity (Beck *et al.* 2010). Particularly, when Beck *et al.* were studying the sequence profiles of CNA from apparently healthy individuals, a previously unknown hepatitis B virus (HBV) infection was detected, suggesting the capability of this method to uncover occult infections (Beck *et al.* 2009). On the other hand, CNA is an easily accessible material to implement the clinical management of cancer patients on the premise of the development of personalized biomarkers that are being identified and accumulated from the ongoing

cancer genome sequencing projects (Leary *et al.* 2010).

2.5 Aim of the thesis

This thesis aims at characterizing circulating cell-free DNA in plasma and investigating its diagnostic application for fetal chromosomal aneuploidy detection by the use of sequencing technologies.

The first part of the thesis studies the biological characteristics of circulating fetal DNA in maternal plasma by the use of cloning and sequencing strategies. In Chapter 4, the end property and fragment species of fetal specific DNA in maternal plasma are investigated in detail.

One of the NGS platforms, namely, the Illumina platform, is employed to achieve the noninvasive prenatal diagnosis of fetal chromosomal aneuploidies by massively parallel maternal plasma DNA sequencing in the second part of the thesis. Chapter 5 attempts to extend this noninvasive diagnostic approach from the published work for fetal trisomy 21 detection to fetal trisomy 13 and 18 detection. To facilitate the clinical implementation of this promising approach, the effects of the starting volume of maternal plasma and barcoding strategy on the diagnostic performance of fetal trisomy 21 detection are investigated in Chapter 6.

The third part of this thesis focuses on another sequencing mode, i.e., paired-end (PE) sequencing, for the plasma DNA analysis. Chapter 7 demonstrates the feasibility of PE sequencing of maternal plasma DNA and its applications for fragment size analysis and fetal trisomy 21 detection. Chapter 8 dissects the size profiles of plasma DNA provided by PE sequencing to understand the underlying biological mechanisms of circulating cell-free DNA. In Chapter 9, PE sequencing approach is applied to sequence the plasma DNA from hematopoietic stem cell

transplant patients in order to validate the cellular origin and reveal the size characteristics of plasma DNA in these patients.

In Chapter 10, a general conclusion of the studies and the future perspectives of massively parallel maternal plasma DNA sequencing are presented.

SECTION II : MATERIALS AND METHODS

CHAPTER 3:METHODS FOR PREPARING DNA FROM MATERNAL PLASMA FOR SEQUENCING

3.1 Preparation of samples

3.1.1 Patient consent

Unless otherwise specified, all cases involved in this thesis were collected from the Prince of Wales Hospital, Hong Kong. Women with singleton pregnancies who attended the Department of Obstetrics and Gynaecology, Prince of Wales Hospital, Hong Kong were recruited. All study participants gave informed consent, and ethics approval was obtained from the Institutional Review Board.

3.1.2 Preparation of plasma samples

The peripheral blood from pregnant women was collected in ethylenediaminetetraacetic acid (EDTA)-containing tubes before termination of pregnancy, amniocentesis or elective cesarean delivery. The peripheral blood from healthy individuals and patients after bone marrow transplantation was also collected in EDTA-containing tubes. The blood samples were first centrifuged at 1,600 *g* for 10 min at 4°C (Centrifuge 5810R, Eppendorf, Hamburg, Germany) so as to separate the plasma from the peripheral blood cells. The plasma portion was carefully transferred to plain polypropylene tubes and then subjected to centrifugation at 16,000 *g* for 10 min at 4°C (Centrifuge 5415R, Eppendorf) to pellet the remaining cells. The blood cell portion was recentrifuged at 2,500 *g* for 5 min in order to remove any residual plasma (Chiu *et al.* 2001). The harvested cell-free plasma

samples were stored in plain polypropylene tubes at -20°C for future DNA extraction.

3.1.3 Collection of placental tissues

Placental tissues were collected immediately after termination of pregnancy or elective cesarean delivery. They were rinsed briefly with diethylpyrocarbonate (DEPC) (Sigma-Aldrich, St. Louis, MO)-treated water, cut into small pieces and stored in plain polypropylene tubes at -80°C.

3.2 Nucleic acid extraction from plasma and tissues

3.2.1 DNA extraction from plasma samples

Plasma DNA was extracted following the blood and body fluid protocol of the QIAamp DSP DNA blood mini kit (Qiagen, Hilden, Germany) with some modifications. The 800 μL of plasma was extracted per column instead of 200 μL as recommended. For efficient extraction, each 800 μL sample was divided into two aliquots. For each 400 μL of plasma, 40 μL of protease and 400 μL of Buffer AL were added, mixed thoroughly and incubated at 56°C for 10 min. Following the incubation, 400 μL of cold absolute ethanol were added to each sample and mixed thoroughly. The mixture was then transferred to a DSP Spin Column and centrifuged at 6,000 *g* for 1 min. The column was then washed with two buffers (Buffer AW1 and AW2) and spun at 6,000 *g* and 16,000 *g* for 1 min, respectively. To remove any residual wash buffer on the column, another spin at 16,000 *g* for 3 min was required. To elute the extracted DNA, 70 μL of sterile water was added to the column and incubated at room temperature for 1 min, followed by a centrifugation at 16,000 *g* for 1 min. For the starting plasma volume larger than 800 μL, multiple DSP Spin

Columns were used and the eluted DNA samples were combined for subsequent experiments.

3.2.2 DNA extraction from placental tissues

Genomic DNA was extracted from placental tissues using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's tissue protocol. To facilitate enzyme digestion, 25 mg of placental tissues was cut into small pieces using a razor blade and placed in 1.5 mL microcentrifuge tubes. In each tube, 360 μL of Buffer ATL and 40 μL of Proteinase K were added, mixed well and then incubated at 56°C in a shaking incubator for 6 hours until the tissue was observed to be completely dissolved. Following the incubation, 400 μL of Buffer AL was added to the lysate and incubated at 70°C for 10 min. Then, 400 μL of cold absolute ethanol was added and mixed thoroughly by vortexing. The mixture was transferred into a QIAamp Mini Spin Column and centrifuged at 16,000 *g* for 1 min. The filtrate was discarded, and the column was washed with two buffers, Buffer AW1 and AW2, spun at 6,000 *g* for 1 min and 16,000 *g* for 3 min, respectively. To remove any residual wash buffer left on the column, another spin at 16,000 *g* for 1 min was performed. Afterwards, 50 μL of sterile water was added to the column and incubated at room temperature for 1 min, with a subsequent centrifugation at 16,000 *g* for 1 min. The eluted DNA samples were stored at -20°C for subsequent experiments.

3.3 Quantitative measurements of nucleic acids

The DNA extracted from placental tissues was quantified by a Nano-Drop ND-1000 spectrophotometer (NanoDrop Technologies). The DNA extracted from plasma samples was quantified by the beta-globin (*HBB*) and *SRY* QPCR assays on an ABI

7300 Sequence Detector (Applied Biosystems) as previously described (Lo *et al.* 1998b). The QPCR was set up according to the manufacturer's instructions (TaqMan PCR Core Reagent Kit). PCR was set up in a total reaction volume of 50 μL by mixing 5 μL of DNA sample with 1 × Buffer A (Applied Biosystems), 4 mM $MgCl_2$, 200 μM of each of dATP, dCTP and dGTP, 400 μM of dUTP, 300 nM of each of the forward and reverse primers (Integrated DNA Technologies, Coralville, IA), 100 nM of TaqMan probe (Applied Biosystems), 0.5 U of AmpErase UNG (Applied Biosystems) and 1.25 U of AmpliTaq Gold (Applied Biosystems). Each reaction mixture was incubated at 50°C for 2 min to activate uracil N-glycosylase, followed by an initial denaturation at 95°C for 10 min, and 45 cycles of thermal cycling at 95°C for 15 s and 60°C for 1 min. Blank controls were included for contamination detection.

For absolute quantification, a calibration curve made up of serially diluted male blood cell DNA, ranging from 1000 to 1 GE per reaction, was run in parallel and in duplicate with each analysis. The average was reported in the results. A conversion factor of 6.6 pg of DNA per cell was used to calculate the yield of extracted plasma DNA. All amplification data were analyzed by the Sequence Detection Software v1.2.3 (Applied Biosystems). The primer and probe sequences for the *HBB* and *SRY* QPCR assays are listed in Table 3.1.

Targets	Primer/Probe	Sequence (5'-3')
SRY	Forward Primer	TGGCGATTAAGTCAAATTCGC
	Reverse Primer	CCCCCTAGTACCCTGACAATGTATT
	Probe	(FAM)AGCAGTAGAGCAGTCAGGGAGGCAGA(TAMRA)
HBB	Forward Primer	GTGCACCTGACTCCTGAGGAGA
	Reverse Primer	CCTTGATACCAACCTGCCCAG
	Probe	(FAM)AAGGTGAACGTGGATGAAGTTGGTGG(TAMRA)

Table 3.1 Summary of primer and probe sequences for QPCR assays.

3.4 Cloning and sequencing

Cloning and sequencing were performed starting from PCR products in this study. PCR product was purified by the MicroSpin™ S-300 HR column (GE Healthcare, Little Chalfont, U.K.) so as to remove unincorporated primers and primer dimers. The purified PCR product was TA-cloned into the pGEM-T Easy vector for transforming into *Escherichia coli* strain JM109 (Promega, Madison, WI), according to the manufacturer's instructions. The PCR product was ligated to the vector in a 10 μL reaction consisting of 1X Rapid Ligation Buffer, 50 ng of pGEM-T Easy Vector, 3 Weiss units of T4 DNA Ligase and 3.5 μL of purified PCR products. The reaction mixture was incubated at 16°C overnight. Three microliters of ligation product were added to 50 μL of JM109 competent cells and chilled on ice for 20 min. Afterwards, the cells were heat-shocked at 42°C for 45 s and placed on ice for 5 min. The cells were then recovered by adding 950 μL of super optimal broth, catabolite repression (SOC) medium (Invitrogen, Carlsbad, CA) and incubated at 37°C for 2 hr with shaking. Then the cells were pelleted by centrifugation at 1,000 *g* for 10 min, resuspended in 100 μL of SOC medium, plated onto LB/ampicillin/ isopropyl-beta-D-thiogalactopyranoside (IPTG)/X-Gal plates and incubated at 37°C overnight. Clones with successful transformation were selected based on the blue/white screening scheme. Since the presence of DNA insert would disrupt the coding sequence of the *β–galactosidase* gene on the vector, the positive recombinant clones could be identified by their white color on LB plates coated with IPTG and X-Gal. White-colored clones were picked randomly and incubated in 10 μL of distilled water at 95°C for 5 min.

Vector primers SP6 (5'-ATTTAGGTGACACTATAGAA-3') and T7 (5'-TAATACGACTCACTATAGGG-3') (Proligo, Singapore) were used to amplify

the cloned inserts. A 25 μL PCR reaction consisting of 1 × Buffer II, 2 mM $MgCl_2$, 200 μM dNTP mix (Promega, Madison, WI), 100 nM each of the forward and reverse primers (Proligo, Singapore), and 1 U of AmpliTaq Gold polymerase was mixed with 3 μL of the clone solution. The PCR was initiated at 95°C for 10 min, followed by 30 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by a final incubation at 72°C for 5 min. The colony PCR products were checked by agarose gel electrophoresis prior to cycle sequencing in order to confirm the presence of inserts.

Cycle sequencing was performed using a BigDye v1.1 kit (Applied Biosystems). A 20 μL reaction consisting of 4 μL BigDye, 2 μL 5 × Sequencing Buffer, and 320 nM SP6 primer was mixed with 2 μL of the cloned PCR product. The thermal profile was 25 cycles of 96°C for 30 s, 50°C for 15 s, and 60°C for 4 min. After purification by ethanol precipitation, 10 μL of HiDi formamide (Applied Biosystems) were added to each sample and denatured at 95°C for 1 min. Capillary electrophoresis was then performed on an ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems).

3.5 Massively parallel sequencing

Figure 3.1 shows the in-house work flow for plasma DNA sequencing using the Illumina sequencing platform.

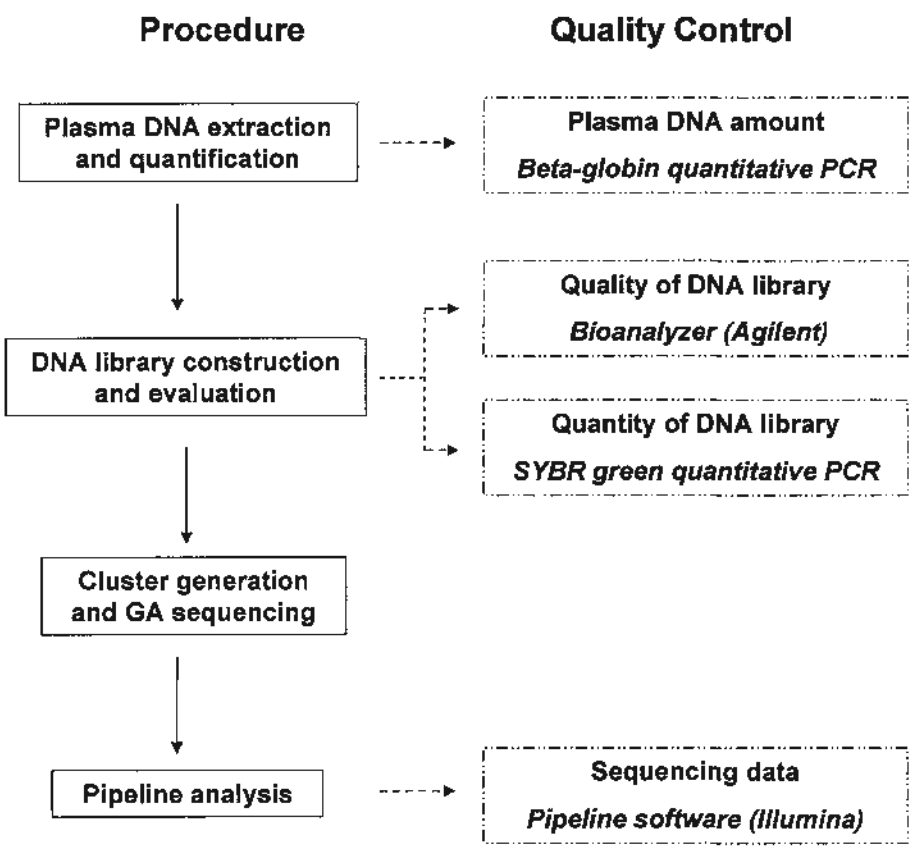


Figure 3.1 In-house work flow for massively parallel plasma DNA sequencing.

The left panel describes the step-by-step work flow of plasma DNA sequencing using the Illumina GA sequencing platform, while the right panel shows the respective quality-checking method for each step.

3.5.1 Library preparation for placental tissue DNA

Figure 3.2 illustrates the recommended work flow of DNA library construction for genomic DNA. Sequencing libraries were constructed from the extracted placental tissue DNA using the standard protocol for the genomic DNA Paired-End Sequencing Sample Preparation Kit (Ilumina). All of the reagent components mentioned below were provided in the kit. Tissue DNA is fragmented by nebulization technique, which breaks up DNA into pieces less than 800 bp in minutes using a disposable device. Firstly, 5 μg of tissue DNA and Tris-EDTA (TE) buffer were mixed in the nebulizer in a total volume of 50 μL. Then 700 μL of nebulization buffer was added to the DNA. After connecting the compressed air source to the inlet port on the top of the nebulizer, nebulization was conducted at 32 pounds per square inch (psi) for 6 min to fragment the tissue DNA. The nebulizer was then centrifuged at 450 *g* for 2 min to collect the droplets from the side of the nebulizer. The fragmented DNA sample was purified by a QIAquick PCR Purification Kit (Qiagen) following the instructions and eluted in 30 μL of elution buffer (EB).

In order to convert the overhangs resulting from fragmentation into blunt ends and also to phosphorylate the 5' ends of the DNA fragments, end repairing was performed in a 100 μL reaction volume containing 30 μL of the nebulized DNA sample, 10 μL of T4 DNA ligase buffer with 10 mM ATP, 4 μL of 10 mM dNTP mix, 5 μL of T4 DNA polymerase, 1 μL of Klenow enzyme, 5 μL of T4 polynucleotide kinase (T4 PNK) and 45 μL of sterile water on the thermal cycler (Eppendorf) for 30 min at 20°C. Following the instructions of the QIAquick PCR Purification Kit (Qiagen), the mixture was purified on a QIAquick column and eluted in 32 μL of EB.

Afterwards, a 50 μL mixture comprising of 32 μL of the DNA from last step, 5 μL of Klenow buffer, 10 μL of 1 mM dATP and 3 μL of Klenow exo$^-$ (5′→ 3′ exonuclease activity lost) was incubated on the thermal cycler for 30 min at 37°C. This step could add a single 'A' nucleotide to the 3' ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction. The reaction mixture was purified by a MinElute PCR Purification Kit (Qiagen) following the manufacturer's instructions and eluted in 10 μL of EB.

The eluted DNA was mixed with 25 μL of 2 × DNA Ligase buffer, and 10 μL of PE Adapter Oligo Mix (Illumina) as well as 5 μL of DNA Ligase, and the mixture was incubated on the thermal cycler for 15 min at 20°C, so as to ligate adapters to the ends of the DNA fragments. The reaction mixture was purified by a QIAquick PCR Purification Kit (Qiagen) following the instructions and eluted in 30 μL of EB.

A 60 mL, 2% agarose gel was prepared and cast with Tris-acetate-EDTA (TAE) buffer. Before the gel electrophoresis, 4.5 μL of SYBR® Green I Nucleic Acid Gel Stain (100 ×, Invitrogen) and 10 μL of Loading Buffer (Qiagen) were added to 30 μL of the DNA from the ligation reaction. At the same time, 1 μL of SYBR® Green I stain and 1.5 μL of Loading Buffer (Qiagen) were added to 5 μL of the Low Molecular Weight DNA Ladder (New England Biolabs). Both mixtures were incubated at room temperature for 10 min. Afterwards, the DNA sample and DNA ladder were loaded onto lanes of the gel, leaving a gap of multiple empty lanes between the sample and the ladder. Gel was run at 150 V for 60 min and then examined on a Dark Reader transilluminator. Using the DNA ladder as a guide, a 2 mm slice of the sample lane at approximately 300 bp was excised with a clean scalpel. The size-selected DNA was purified by a QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's instructions and eluted in 30 μL of EB.

An enrichment PCR is involved to selectively enrich those DNA fragments that have adapter molecules on both ends, and to amplify the amount of DNA in the library. PCR was set up in a total reaction volume of 50 μL by mixing 10 μL of purified DNA, 1 μL of PCR primer PE 1.0 and 1 μL of PCR primer PE 2.0, and 25 μL of Phusion DNA polymerase (Finnzymes Oy) as well as 13 μL of ultra pure water. Each reaction mixture was incubated at 98°C for 30 s, followed by 12 cycles of thermal cycling at 98°C for 40 s, 65°C for 30 s and 72°C for 30s. Blank controls were included for contamination detection. The PCR products were purified by a QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions and eluted in 30 μL of EB.

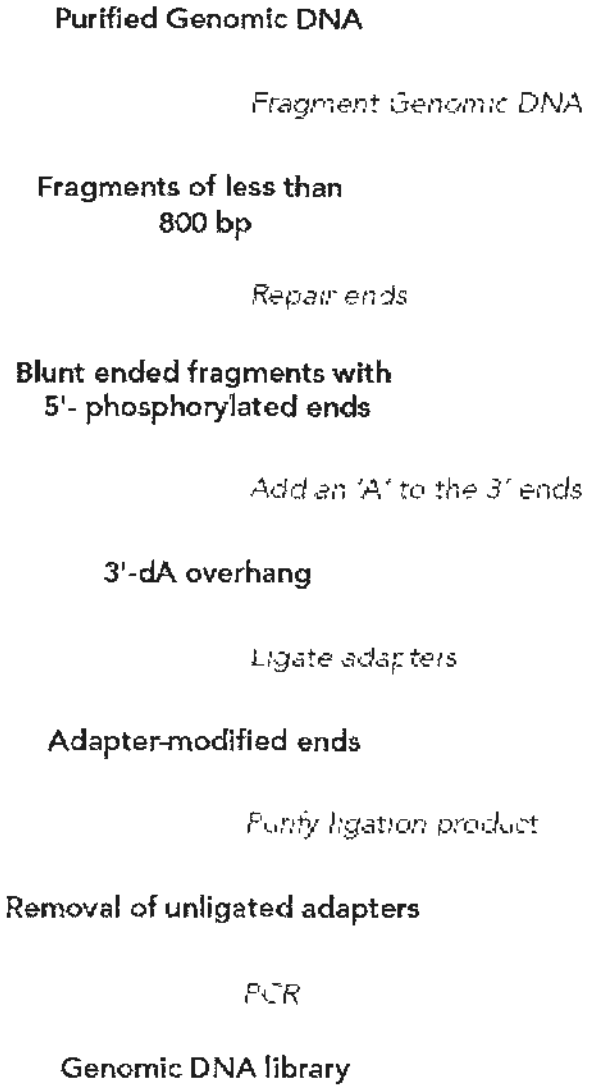


Figure 3.2 Illumina workflow of DNA library construction for genomic DNA

Figure extracted from the Paired-End Sample Preparation Guide, Illumina, 2010

3.5.2 Library preparation for plasma DNA

The DNA library construction for the extracted plasma DNA was mostly according to manufacturer's instructions with several modifications as described below. Firstly, plasma DNA molecules are short fragments in nature (Chan *et al.* 2004); hence, the steps of fragmentation and size selection by gel electrophoresis are omitted. Secondly, as the DNA amount in plasma samples is limited (Lo *et al.* 1998b), the concentrations of reagent components in each reaction are adjusted and the cycle of enrichment PCR is increased from 12 to 15 cycles.

The extracted plasma DNA was concentrated by a SpeedVac Concentrator (Thermo, SAVANT DNA120) into a final volume of 40 μL per case for the subsequent library preparation. End repairing of the plasma DNA fragments was performed in a 50 μL reaction volume containing 40 μL of concentrated plasma DNA, 5 μL of T4 DNA ligase buffer with 10 mM ATP, 2 μL of 10 mM dNTP mix, 1 μL of T4 DNA polymerase, 1 μL of 5-fold diluted Klenow enzyme and 1 μL of T4 PNK on the thermal cycler for 30 min at 20°C. The DNA was purified by the QIAquick PCR Purification Kit (Qiagen) and eluted in 34 μL of EB. Next, a 50 μL reaction volume comprising of 34 μL of the DNA from the last step, 5 μL of Klenow buffer, 10 μL of 1 mM dATP and 1 μL of Klenow exo- was incubated on the thermal cycler for 30 min at 37°C to add an 'A' to the 3' ends of the plasma DNA fragments. The DNA was purified by a MinElute PCR Purification Kit (Qiagen) and eluted in 10 μL of EB, which was then mixed with 15 μL of 2 × DNA Ligase buffer, and 1 μL of 10-fold diluted Illumina adapters as well as 4 μL of DNA Ligase and incubated at 20°C for 15 min to ligate the adapters to the DNA fragments. The adapter-ligated DNA was purified by a QIAquick PCR Purification Kit (Qiagen) and eluted in 23 μL of EB.

For DNA sequencing, Illumina provides multiple kinds of commercially available adapters for various sequencing modes, e.g., single-read (SR) and paired-end (PE) sequencing. Both modes are involved in the studies of this thesis. Therefore, in the experiments, the corresponding Illumina adapters were added to the ligation reaction. The adapter-ligated DNA fragments were then amplified using the enrichment PCR with the Illumina primers that corresponded to the adapters. PCR was set up in a total reaction volume of 50 μL by mixing 23 μL of adapter-ligated DNA and 25 μL of Phusion DNA polymerase along with 1 μL of each primer. Each reaction mixture was incubated at 98°C for 30 s, followed by 15 cycles of thermal cycling at 98°C for 40 s, 65°C for 30 s and 72°C for 30 s. Blank controls were included for contamination detection. The PCR products were purified by a QIAquick MinElute PCR Purification Kit (Qiagen) and eluted in 17 μL of EB.

3.5.3 Validation of DNA library

Bioanalyzer validation

A bioanalyzer (Agilent 2100) was used to check the quality and size of the adapter-ligated DNA libraries. Empirically, for libraries constructed from plasma DNA, there would be a sharp peak appearing at around 260-290 bp, depending on which set of adapter/primers was used (e.g., for the SR library, the sequence length occupied by SR adapter/primer is ~90 bp; for the PE library, it is ~120 bp); while for libraries constructed from tissue DNA, since a size-selection step was involved, the observed size distribution on the bioanalyzer should be equal to the size-selected range plus the sequence length of the adapters/primers. For the plasma samples, the position of the sharp peak on bioanalyzer was regarded as the main size of the DNA molecules in the library; while for the tissue samples, the middle of the size range was regarded as the main size of the DNA molecules in the library.

SYBR Green quantitative PCR

Meyer *et al.* reported a quantitative PCR-based method for the quantification of 454 sequencing library to minimize the initial material for the 454 sequencing platform (Meyer *et al.* 2008). Based on a similar principle, a universal SYBR Green quantitative PCR assay was designed, which could target the common sequences within various Illumina primer sets, to quantify multiple types of constructed DNA libraries.

To establish a standard curve for the SYBR Green quantitative PCR, one of SR libraries was used for a whole DNA library amplification, followed by the cloning and sequencing processing. One microliter of the 1000-fold diluted library was used as a template for a 25 μL PCR reaction consisting of 1 × Buffer II, 2.5 mM $MgCl_2$, 250 μM dNTP mix (Promega, Madison, WI), 1.25 U of AmpliTaq Gold polymerase and 1 μL of each of the Illumina SR primers. The PCR was initiated at 95°C for 10 min, followed by 35 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s, followed by a final incubation at 72°C for 10 min. The resulting whole library amplicon was cloned according to the protocols described above. Several clones were subjected to sequencing on ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems) and their sequences were analyzed to ensure that the clones carried the enrichment PCR priming sites flanking the inserted fragments. One confirmed clone with an insert size of 330 bp was selected, and used as a template for colony PCR using primers of SP6 and T7 as described above. Colony PCR product was purified using the QIAquick PCR purification kit (Qiagen) and quantified on a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). The molecular concentration was calculated from the product size and the mass concentration with a conversion factor of 650 Dalton per basepair of DNA. Then, a series of dilution was conducted

to obtain the quantification standards ranging from 10^9 to 10^2 molecules/uL for the SYBR Green quantitative PCR.

Duplicate measurements were carried out using an ABI 7300 Sequence Detector (Applied Biosystems) in a 50 μL reaction volume containing 2 μL of 10,000-fold diluted DNA library, 1 × Power SYBR Green PCR Master Mix (Applied Biosystems), 200 nM of each primer (Q-PCR-F: 5' - GAT ACG GCG ACC ACC GAG AT - 3'; Q-PCR-R: 5' - CAA GCA GAA GAC GGC ATA CGA G - 3') and 0.5 U AmpEras UNG. Cycling conditions included an incubation step at 50°C for 2 min to activate the AmpEras UNG and initial denaturation at 95°C for 10 min, followed by 40 cycles of denaturation at 95°C for 15 s and annealing and extension at 60°C for 1 min. Finally, dissociation curve analysis was performed at 95°C for 15 s, then 60°C for 30 s, followed by a slow ramp to 95°C. The library DNA concentrations were calculated according to the following formula: library concentration [molecules/μL] = readout of SYBR Green quantitative assay [molecules/μL] × dilution factor (10,000) × (size of the standard [bp]/main size of the DNA molecules in the library [bp]).

3.5.4 Cluster generation and SBS

Based on the measurement from SYBR Green quantitative PCR, each of the DNA libraries was diluted to 10 nM. The diluted DNA library was denatured in 0.1N NaOH at room temperature for 5 minutes, and then further diluted to a final 36 pM concentration in 1 mL of pre-chilled hybridization buffer. The single-stranded DNA molecules were then introduced onto the Illumina flow cell using the Cluster Station (Illumina).

For the SR sequencing, 36 pM of DNA was subjected to hybridization onto the SR sequencing flow cell (Illumina) with a SR Cluster Generation Kit v1 (Illumina), followed by 36 cycles of sequencing on a GA II (Illumina) with a 36-Cycle SBS Sequencing Kit v1 (Illumina).

Illumina sequencing platform allows one to perform PE sequencing to sequence both ends of DNA fragments. The procedures of PE sequencing include the cluster generation and SBS of Read 1 and cluster regeneration and SBS of Read 2. First, 36 pM of DNA library was subjected to hybridization onto the PE sequencing flow cell (Illumina) using a Paired-End Cluster Generation Kit v1 (boxes 1 and 3 for Read 1), followed by the first round of 36 cycles of sequencing on a GA II (Illumina) with a 36-Cycle SBS Sequencing Kit v1 (Illumina). Paired-End Module (Illumina) is a requisite part for PE sequencing, which regulates reagents to pump into the flow cell for the regeneration of the second read. The extended sequencing primer for Read 1 was stripped off and the complementary strands were bridge-amplified to form clusters *in situ* using Paired-End Module with a Paired-End Cluster Generation Kit v1 (boxes 2 and 4 for Read 2). Then, the second round of 36 cycles of sequencing was performed on a GA II (Illumina) with another 36-Cycle SBS Sequencing Kit v1 (Illumina).

3.5.5 Image analysis and base calling

The images taken per sequencing cycle were automatically transferred to a dedicated server, where the image processing and base calling were done using the programs in the GAPipeline software package provided by Illumina.

3.6 Statistical analysis

Statistical analyses were performed with SigmaStat 3.0 software (SPSS) and MedCalc software (version 9.6.4.0).

SECTION III : CHARACTERIZATION OF CIRCULATING FETAL DNA BY CLONING AND SEQUENCING

CHAPTER 4: CHARACTERIZATION OF CIRCULATING FETAL DNA BY CLONING AND SEQUENCING

4.1 Introduction

The discovery of fetal-derived DNA in maternal plasma has opened up new possibilities of noninvasive prenatal diagnosis and monitoring (Lo *et al.* 1997). During the past decade, plenty of sensitive and reliable noninvasive prenatal diagnosis assays have been developed to detect genetic characteristics of fetuses in early gestation (Chiu *et al.* 2008; Chiu *et al.* 2002b; Costa *et al.* 2002; Lo *et al.* 1998a; Lo *et al.* 2007a; Lo *et al.* 2007b; Tong *et al.* 2006).

Apart from its clinical applications for prenatal diagnosis, investigators also have made efforts to unravel the biological features of circulating fetal DNA in maternal plasma. It has been reported that fetal DNA accounts for 3.4-6.2% of total DNA circulating in the plasma of pregnant women and the concentrations of fetal DNA increase as gestation progresses (Lo *et al.* 1998b). The postpartum clearance of fetal DNA is rapid with a mean half-life of about 16 minutes (Lo *et al.* 1999c). With respect to the cellular origin, the accumulation of evidence favors the hypothesis that the circulating fetal DNA is derived from fetal and/or placental cells undergoing apoptosis and necrosis (Chim *et al.* 2005; Masuzaki *et al.* 2004; Tjoa *et al.* 2006). Chan *et al.* investigated the size distributions of cell-free DNA in maternal plasma and found that the circulating fetal DNA consisted predominantly of short DNA fragments and was generally shorter than maternally-contributed DNA (Chan *et al.* 2004).

From sequence analysis, the total circulating DNA in healthy individuals and malignant patients have been characterized to some extent (Beck *et al.* 2009; Li *et al.*

1989; Suzuki *et al.* 2008; van der Vaart *et al.* 2008). For maternal plasma DNA, it would be well worth performing the sequence analysis on the fetal specific DNA fragments so as to further elucidate the biological features of circulating fetal DNA molecules in maternal plasma. Such information, in turn, will be beneficial to the assay development for noninvasive prenatal diagnosis by maternal plasma DNA analysis.

The aim of this study is to investigate the fetal specific DNA fragments in maternal plasma by the use of touch down ligation-mediated PCR (LM-PCR) coupled with cloning and sequencing strategies. The *SRY* and Testis Specific Protein, Y-linked (*TSPY*) genes located on the Y chromosome were chosen as targets to characterize the fetal specific DNA fragments in maternal plasma in terms of the nature of the DNA ends and the fragment species with cleavage sites.

4.2 Methods

4.2.1 Subjects

Blood samples from pregnant women in the third trimester of pregnancy were recruited according to the procedure described in Chapter 3.1.1.

4.2.2 Sample preparation

Plasma was harvested from blood samples as described in Chapter 3.1.2. DNA was extracted from the maternal plasma according to procedures described in Chapter 3.2.1.

4.2.3 QPCR assays

To determine the input amount of maternal plasma DNA for the downstream procedures, all extracted DNA was subjected to *HBB* QPCR as described in Chapter 3.3. Instead of the *SRY* QPCR assay described in Chapter 3.3, another *SRY* QPCR assay was designed, which targeted the region nearby the gene-specific primers for the following *SRY* touch down PCR assay. The QPCR was set up according to the manufacturer's instructions (TaqMan PCR Core Reagent Kit) with a reaction volume of 50 μL containing 5 μL 10 × buffer A, 300 nM of each amplification primer (*SRY*-F, 5'- CCG TTT CAC ACT GAT ACT TAG AGT TAC A -3'; *SRY*-R, 5'-GCG TAT TCA ACA GCG ATG ATT ACA-3), 150 nM of TaqMan minor-groove-binding (MGB) probe (Applied Biosystems) (5'-(FAM) GAG AGC GGG AAT ATT (MGBNFQ)-3'), 4 mM $MgCl_2$, 0.2 mM each dATP, dCTP, and dGTP; 0.4 mM dUTP; 1.25 U AmpliTaq Gold and 0.5 U AmpErase UNG as well as 5 μL of plasma DNA. Each reaction mixture was incubated at 50°C for 2 min to activate uracil N-glycosylase, followed by an initial denaturation at 95°C for 10 min, and 45 cycles of thermal cycling at 95°C for 15 s and 60°C for 1 min on an Applied Biosystems 7300 Sequence Detector (Applied Biosystems).

4.2.4 Experimental design

SRY and *TSPY* genes located on the Y chromosome were chosen as targets for the touch down LM-PCR. Two sets of primers were designed for the *SRY* assay, which amplified *SRY* sequences in two opposite directions. For *TSPY* assay, two regions were chosen and four sets of directional primers were designed for each region (Figure 4.1 A). In order to capture the low abundance of fetal specific DNA in the overwhelming maternal background, the touch down PCR method was adopted. As shown in Figure 4.1 B, two round nested PCR were performed with pairs of adapter primers (AP) and gene-specific primers (GSP).

The experimental procedures are shown in Figure 4.2. Firstly, three maternal plasma samples involving female fetuses were used to assess the chromosome Y specificity of the touch down PCR assay. Next, three maternal plasma samples from male pregnancies were subjected to the polished and unpolished comparison. Using the *SRY* assays, the cloned sequences obtained from both procedures were compared to infer the nature of the DNA ends of the fetal DNA fragments. Lastly, four maternal plasma samples involving male fetuses were analyzed by both *SRY* and *TSPY* assays. By studying the fragment species and the cleavage sites of the sequenced fetal DNA fragments, we were able to reveal the characteristics of circulating fetal DNA in maternal plasma.

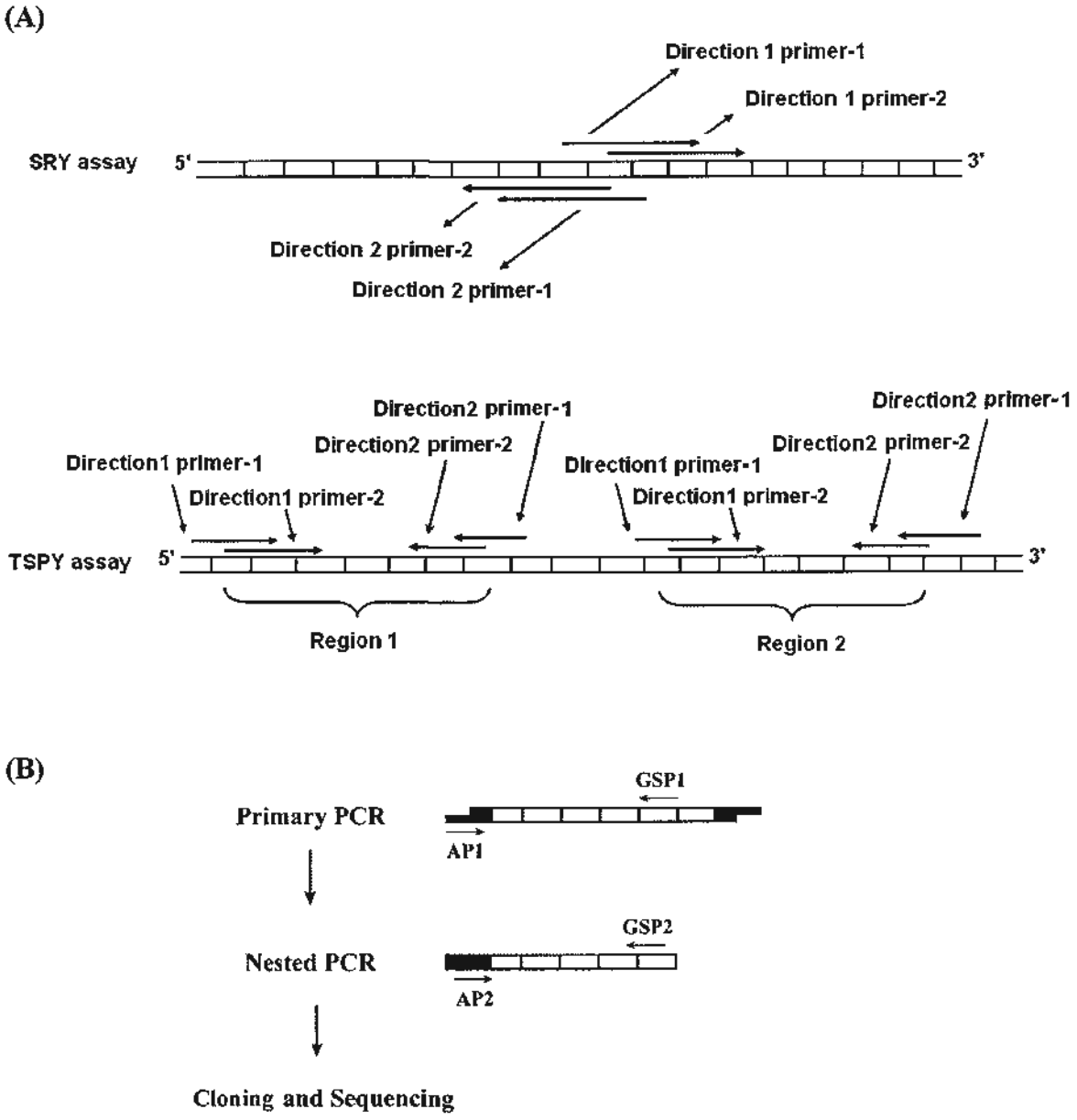
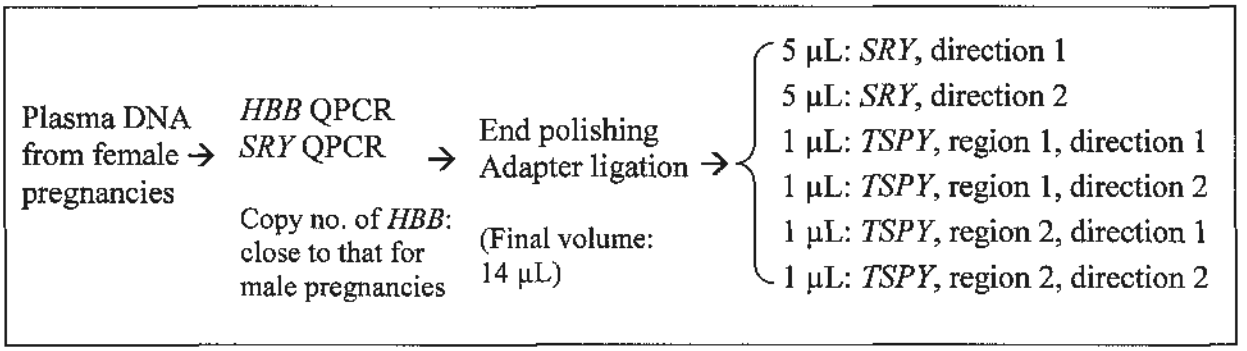
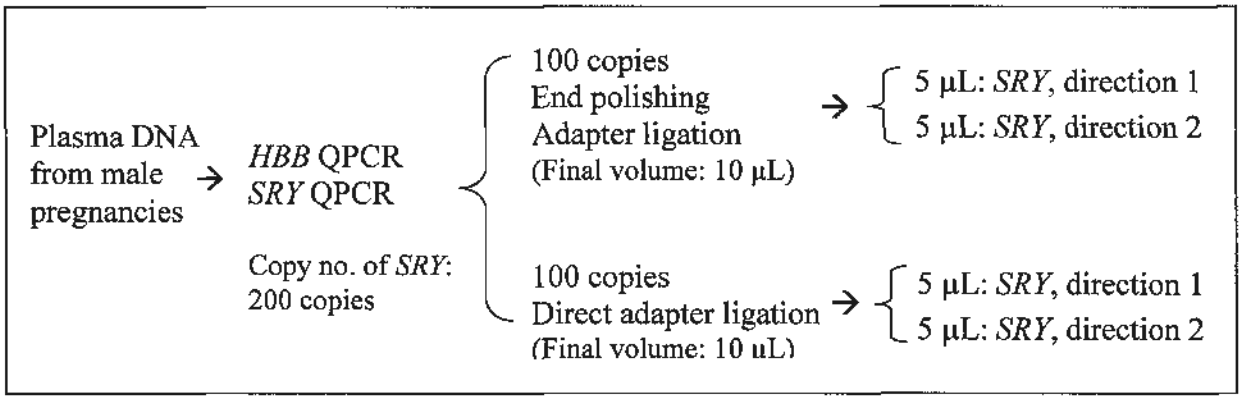


Figure 4.1 Schematic diagram of primer design for the *SRY* and *TSPY* assays.

A



B



C

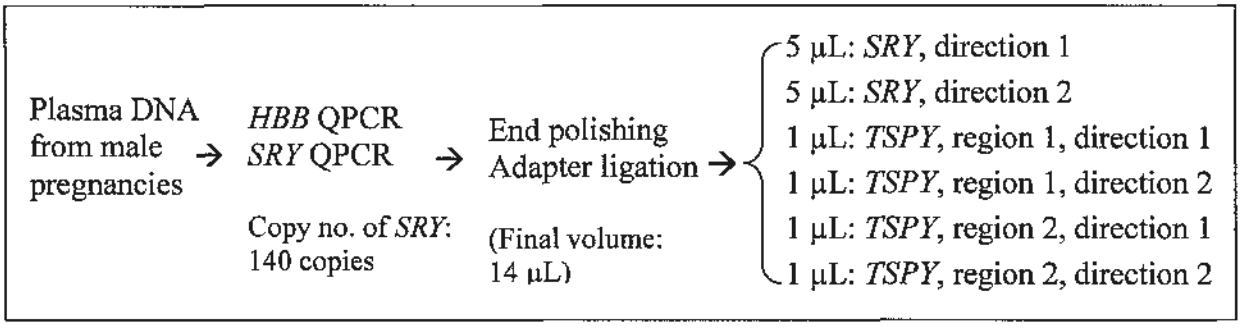


Figure 4.2 Experimental design.

(A) Specificity of the designed assays. The same amount of plasma DNA from female pregnancies as those from male pregnancies determined by the *HBB* QPCR assay was used. ***(B)*** Polished and unpolished assay comparison. Maternal plasma samples from male pregnancies were quantified by the *SRY* QPCR assay and divided into two aliquots for comparative analysis. ***(C)*** Characterization of fetal DNA fragments. The copy number of maternal plasma samples from male pregnancies was determined by the *SRY* QPCR assay. Then, 50 copies and 10 copies of plasma DNA were input as starting material for the *SRY* and *TSPY* touch down LM-PCR assays, respectively.

4.2.5 DNA polishing and ligation

DNA polishing

Plasma DNA was concentrated by a SpeedVac Concentrator (Thermo, SAVANT DNA120) and then the DNA ends were polished using a Quick Blunting™ Kit (New England Biolabs) according to the manufacturer's instruction. The plasma DNA was mixed with 1.25 μL of 10 × Blunting Buffer, 1.25 μL of 1 mM dNTP Mix, and 0.5 μL of Blunting Enzyme Mix in a final volume of 12.5 μL. The reaction was incubated at room temperature for 15 min. The enzyme in the blunting reaction was immediately inactivated by heating at 70°C for 10 min.

Oligo annealing and DNA ligation

The adapters were provided in a GenomeWalker™ Universal Kit (Clontech), or prepared by annealing two partially complementary oligonucleotides (5'-GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGG T-3'; 5'-PO_4-ACCAGCCC-N_2H-3') in a reaction volume of 10 μL consisting of 1 μL of each oligonucleotide (100 μM, Tech Dragon Limited) and 8 μL annealing buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl). The annealing reaction was incubated at 95°C for 3 min, followed by reducing the annealing temperature to 25°C with 1°C reduction per cycle.

The end-polished plasma DNA was ligated with adapters using the Quick Ligation™ Kit (New England Biolabs). The ligation reaction containing 10 μL of adapters (10 μM), 25 μL 2 × Ligation Buffer, 2.5 μL Quick T4 DNA Ligase and 12.5 μL end-polished plasma DNA was incubated at room temperature for 10 min, and then

at 16°C overnight. The reaction mixture was purified by a MinElute PCR Purification Kit (Qiagen) following the manufacturer's instructions and eluted in EB.

Polished and unpolished assays

To investigate the end property of circulating fetal DNA in maternal plasma, the DNA fragments with and without end polishing were analyzed. In the unpolished assay, the DNA fragments were directly subjected to the ligation with adapter as described above, bypassing the end-repairing step.

4.2.6 Touch down LM-PCR

Touch down PCRs were performed using an Advantage® 2 Polymerase Mix (Clontech) according to the instructions. The primer sequences of GSP and AP are listed in Table 4.1. Primary PCRs were conducted in a 25 μL reaction consisting of 1 × Advantage 2 PCR buffer, 0.2 mM dNTP, 200 nM adapter primer AP1, 200 nM gene-specific primers 1, and 1 U Advantage 2 Polymerase Mix as well as the adapter-ligated DNA templates. After an initial denaturation at 95°C for 3 min, 10 cycles of denature at 95°C for 45 s, annealing at 69°C (for *SRY* assays) or 72°C (for *TSPY* assays) for 45 s and extension at 68°C for 1 min were carried out, followed by 30 cycles of thermal cycling at 95°C for 45 s, 64°C (for *SRY* assays) or 67°C (for *TSPY* assays) for 45 s and 68°C for 1 min and an extension phase at 68°C for 10 min. Touch down was conducted during the first 10 cycles, where the annealing temperature was decreased by 0.5°C/cycle. The primary PCR products were 100-fold diluted, and used as template for the secondary PCR. In the secondary PCR, the same reagent components as the primary PCR were used except that the AP1 and GSP1 were replaced by the AP2 and GSP2. The reaction was incubated at 95°C for 3 min, followed by 10 cycles of denature at 95°C for 45 s, annealing at 73°C (for *SRY*

assays) or 75°C (for *TSPY* assays) for 45 s and extension at 68°C for 1 min. Touch down was also conducted with the annealing temperature decreased by 0.5°C/cycle. This was then followed by 26 cycles of thermal cycling at 95°C for 45 s, 67°C (for *SRY* assays) or 69°C (for *TSPY* assays) for 45 s and 68°C for 1 min and an extension phase at 68°C for 10 min. PCR products were examined on 1.5% agarose gels stained with GelRed™ (Biotium).

(A)

Gene	Gene-specific primer	Sequences
SRY	*SRY*-direction1 primer-1	5'-GGTACTCTGCAGCGAAGTGCAACTGGACAAC-3'
	SRY-direction1 primer-2	5'-GACAACAGGTTGTACAGGGATGACTGTACG-3'
	SRY-direction2 primer-1	5'-CGTACAGTCATCCCTGTACAACCTGTTGTC-3'
	SRY-direction2 primer-2	5'-GTTGTCCAGTTGCACTTCGCTGCAGAGTACC-3'
TSPY	*TSPY*-region1-direction1 primer-1	5'-AGCAGGCTGTGCCTGGCCCTGGGCCCATGA-3'
	TSPY-region1-direction1 primer-2	5'-CCATGACCCCAGAGTCTGCACTGGAGGAGC-3'
	TSPY-region1-direction2 primer-1	5'-CATTGGCCCAGAAGCCAGGGACGCTCT-3'
	TSPY-region1-direction2 primer-2	5'-ATGACGGCGCCTCTGCGGTCTAGGTGGGG-3'
	TSPY-region2-direction1 primer-1	5'-GATCATACATGGAAGCAGATCTGAG-3'
	TSPY-region2-direction1 primer-2	5'-AGATCTGAGAAATCCCCTACCCCAGCCTCT-3'
	TSPY-region2-direction2 primer-1	5'-AGTCTTCCTGGCCTCACCTCCAGGC-3'
	TSPY-region2-direction2 primer-2	5'-CAGGCTGACCATGTAGCTCAGCATGTCTT-3'

(B)

Adapter Primer	**Sequences**
Adapter Primer 1 (AP1)	5'-GTAATACGACTCACTATAGGGC-3'
Nested Adapter Primer 2 (AP2)	5'-ACTATAGGGCACGCGTGGT-3'

Table 4.1 Primer sequences for touch down LM-PCR.

(A) Sequences of gene-specific primers (GSP). ***(B)*** Sequences of adapter primers (AP)

4.2.7 Cloning and sequencing

For each assay, 3.5 μL of PCR products were subjected to cloning and sequencing processing according to the procotols described in Chapter 3.4. At least 16 individual white recombinant colonies were randomly picked for each assay and sequenced on an ABI 3100 Genetic Analyzer (Applied Biosystems).

4.2.8 Sequence analysis

The DNA sequences were extracted, and subsequently aligned to reference sequences using the Seqscape V2.0 software (Applied Biosystems). The reference sequences referred to the nucleotide contents by extending the gene specific primers to 1000 bp upstream or 1000 bp downstream. Only the sequence reads containing the sequences of GSP2 and AP2 were retrieved. A fragment species was defined as a uniquely identified sequence read starting from GSP2 and ending before AP2. The cleavage sites in fetal DNA fragments, defined as the nucleotides at the junction between the adapter sequence and the target DNA fragments, were analyzed from all fragment species. The fragment length and the end nucleotides were documented. For the non-specific sequences that could not be assembled to the reference sequences, the identities and chromosomal locations were determined using the Basic Local Alignment Search Tool (BLAST) of the National Center for Biotechnology Information (NCBI).

4.3 Results

4.3.1 Validation of touch down LM-PCR

To determine whether the touch down LM-PCR assays were able to amplify chromosome Y specific DNA, three maternal plasma samples from female pregnancies and four maternal plasma samples from male pregnancies were examined using both *SRY* and *TSPY* GSPs. Total plasma DNA concentrations were determined by *HBB* QPCR assay to ensure that all plasma samples were properly prepared. The *HBB* measurements showed that the concentrations of all extracted plasma DNA ranged from 1219 to 1842 copies/mL, being comparable to the previous data (Lo *et al.* 1998b). Positive signal could only be detected in the plasma samples from women bearing male fetuses using the *SRY* QPCR assay, and the fetal DNA concentrations were from 99 to 452 copies/mL, also in line with the previous data (Lo *et al.* 1998b).

The specificity of the designed assays was checked according to the procedures described in Figure 4.2A using three maternal plasma samples from female pregnancies. Although PCR bands were observed on agarose gel after electrophoresis of the LM-PCR products, all clones obtained from female cases could not be aligned to the target regions. BLASTing against the human reference genome indicated that all of them were derived from chromosomes other than Y chromosome. Target specific clones were only obtained in the plasma samples from male pregnancies, demonstrating that the touch down LM-PCR assays were able to capture and amplify the fetal specific DNA from maternal plasma.

4.3.2 Characterization of DNA ends of fetal DNA fragments

To investigate the end property of fetal DNA fragments in maternal plasma, a comparative analysis based on the experimental procedures including or excluding the end-polishing step was performed using *SRY* assays. Three plasma samples from male pregnancies were involved and the plasma DNA from each case was divided into two aliquots, each processed according to the two procedures (Figure 4.2B) independently. In the polished assay, T4 DNA polymerase and T4 Polynucleotide Kinase would convert plasma DNA with incompatible 5' or 3' overhangs to 5' phosphorylated, blunt-ended DNA so as to facilitate the downstream blunt-end ligation with adapters (Figure 4.3). In the parallel experiment, the other aliquot of plasma DNA was directly ligated with adapters, followed by touch down LM-PCR (Figure 4.3).

More than 500 clones were picked and analyzed, of which 79 fragment species were identified. Table 4.2 summarizes the numbers of DNA fragments species identified through the procedures with and without the end-polishing step, respectively. From the *SRY* direction 1 assay, 25 and 19 fragment species were found through two procedures, respectively; while from the *SRY* direction 2 assay, the numbers were 25 and 10, respectively. A definite number of fetal DNA fragment species arising from the unpolished assay indicated the existence of 5' phosphorylated, blunt-ended fetal DNA fragments in maternal plasma. Thus, there would be two forms of DNA ends for the fetal DNA molecules in maternal plasma, namely, the 5' phosphorylated, blunt-end and the staggered end. Notably, inclusion of the end-polishing step allowed the detection of more fetal DNA fragment species by sequencing in general.

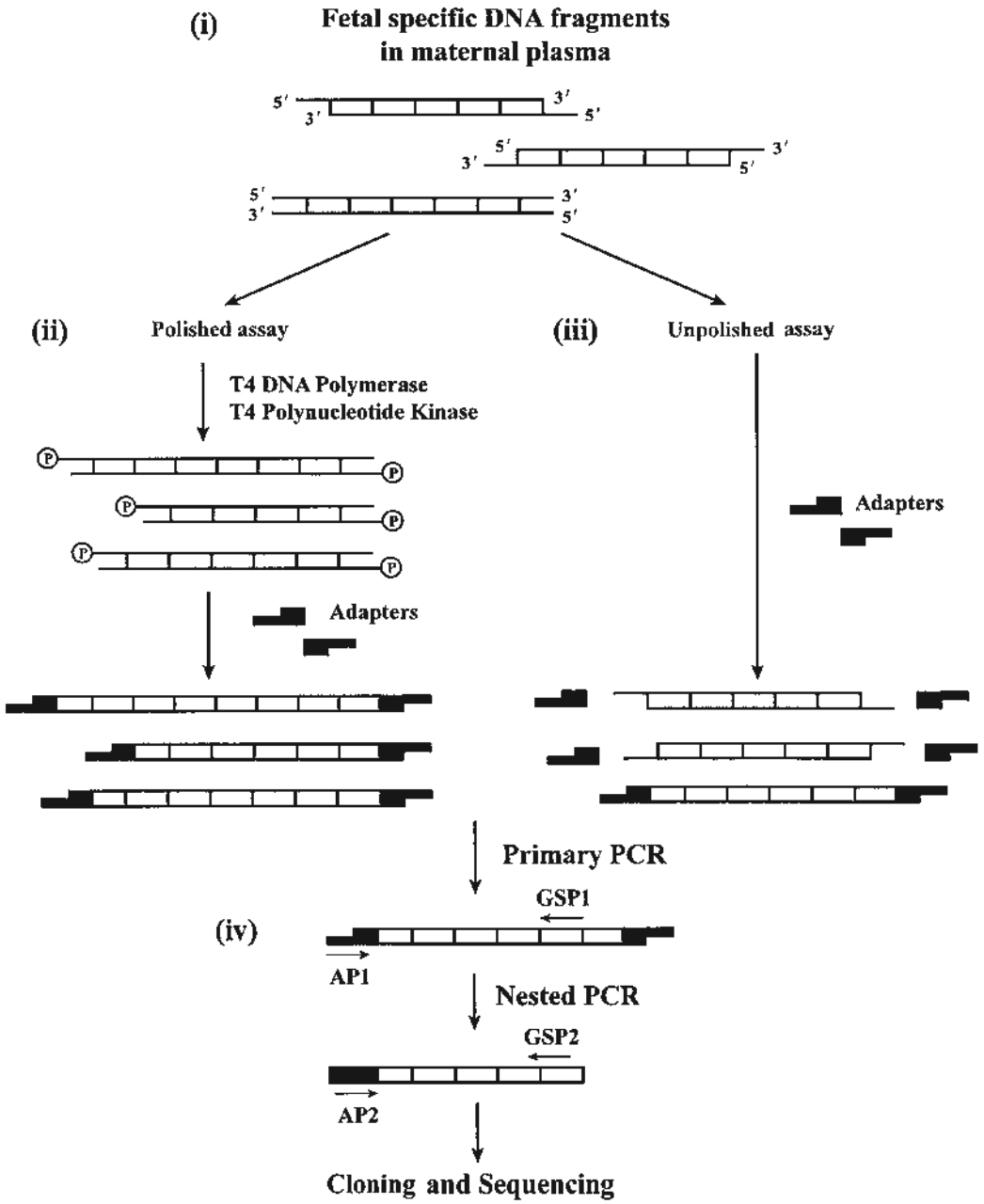


Figure 4.3 Schematic diagram of the experimental procedures for the polished and unpolished assays.

(i) Fetal DNA fragments in maternal plasma may bear diverse end patterns, e.g., 5' overhang, 3'overhang and blunt ends. (ii) In the end-polishing step, T4 DNA polymerase and T4 polynucleotide kinase were added in the reaction mixture. The 3' to 5' exonuclease activity of the T4 DNA polymerase removes 3' overhangs and the 5' to 3' polymerase activity of the enzyme fills in the 5' overhangs. Meanwhile, the T4 polynucleotide kinase catalyzes the transfer of the g-phosphate from ATP to the 5'-OH group of double-stranded fragments. With end polished, all original fragments with diverse end patterns would be unified to the 5' phosphorylated, blunt-ended

fragments. (iii) On the other hand, without such a step, only the 5' phosphorylated, blunt-ended fragments by nature could be successfully ligated with adapters. (iv) Both aliquots would be subjected to the touch down LM-PCR and subsequent cloning and sequencing.

Case no.	*SRY*-Direction 1		*SRY*-Direction 2	
	Polished	Unpolished	Polished	Unpolished
M3835	8	4	8	3
M3998	6	8	7	4
M4010	11	7	10	3

Table 4.2 Number of DNA fragment species identified from the polished and unpolished assays.

4.3.3 Fragment species and cleavage sites of fetal DNA fragments

According to the procedures described in Figure 4.2C, touch down LM-PCR products from four pregnant women carrying male fetuses were cloned and sequenced to characterize the fetal DNA fragments. More than 600 clones were screened, among which 359 clones were retrieved. All assembled sequences were subjected to BLAST against the NCBI human genome database to affirm the DNA fragments were located within the target specific regions. Thirty-nine fragment species were obtained from the *SRY* assay. One hundred and forty one fragment species were obtained from the *TSPY* assays, with 63 and 78 from region 1 and region 2, respectively. As *TSPY* is a multi-copy gene whereas *SRY* is a single-copy gene on the Y chromosome (Ali *et al.* 2003), it is not surprising that more *TSPY* fragment species were cloned.

The cleavage sites of fetal DNA fragments from the four maternal plasma samples are shown respectively in Figure 4.4. The cleavage sites distributed dispersedly alongside the reference sequences with few cleavage sites shared by different maternal plasma samples. Among all fragment species that we obtained from these four samples, there were totally 55 and 68 unique fragment species from the direction 1 and 2 assays, respectively (the end nucleotide that was shared by multiple samples were only counted as once). Assuming that there was no specific cleavage pattern, the contents of A/T and G/C for the end nucleotides would be equal to 50%. To determine whether there would be any base that is preferentially cleaved, I looked into the base composition around the cleavage sites of these unique fragment species. Figure 4.5 shows the base composition for the adjacent 5 bases at the 5' and 3' termini. For the 5' end, the contents of G/C for the first five bases were 54.41%, 73.53%, 47.06%, 58.82% and 61.76%, respectively, while for the 3' end, they were

41.82%, 56.36%, 47.27%, 74.55% and 67.27%, respectively. With the use of a chi-square test, I found that the second base at the 5' end and the fourth base at the 3' end had a higher G/C content (chi-square test, $P < 0.05$). For the other bases, no significant difference was observed.

(A)

TTACAGTCCAGCTGTGCAAGAGAATATTCCCGCTCTCCGGAGAAGCTCTTCCTTCCTTT
GCACTGAAAGCTGTAACTCTAAGTATCAGTGTGAAACGGGAGAAAACAGTAAAGGCAAC
GT↓CCAGGATAGAGTGAAGCGACCCATGAACGCATTCATCGTGTGGTCTCGCGATCAGA
GGCGCAAGATGGCTC↓TAGAGAATCCCAGAATGCGAAACTCAGAGATCAGCAAGCAGC
TGGGATACCAGTGGAAAATGCTTACTGAAGCCGAAAAATGGCCATTCTTCCAGGAGGCA
CAGAAATTACAGGCCA TGCACAGAGAGAAA↓TACCCGAATTATAAG TAT↓CGACC↓T↓
↓CGT CGGAAGGCGA↓AGATGCT↓G↓↓CCGAA↓GAATTGC↓AGTTTGCTTCCCGCAGATC
CCGCTTCGGTACTCTGCAGCGAAGTGCAACTGGACAACAGGTTGTACAGGGATGACTG
TACGAAAGCC↓A↓CA CACTCA↓↓AGAATGG A↓GCACC↓AGC↓TAGGC CACTTACCGC
CCATC↓A↓ACG↓CAGCCAGCTCACCGCAGCAACGGG↓ACCGC↓TACAGCCACTGGA↓CA
AAGCTGTAGG↓ACAATCGGG T AACA↓TTGGC TACAAAGACCTACCTAGATGCTCCTT
TTTACGATAACTTACAGC↓CCTCACTTTCTTATGTTTAGTTTCAATATTGTTTTCTTTTCT
CTGGCTAATAAAGGC CTTATTCATTTCA

→ *SRY* Direction 1 Primer
← *SRY* Direction 2 Primer
↓= M3529 ↓=M3530 ↓= M3533 ↓= M3552

(B)

GGAGGAGGCGGTGCTGCTGTTGGATGACATAATGGCGGAGGTGGAGGTGGTGGCGGA
GGTGGAGGTGGTGGCGGAGGAGGAGGGCC ↑ TCGTGGAGCGGCGGGAGGAGGCCCAG
CGGGCACAGCAGGCTGTGCCTG ↑ GCCCTGGGCCCATGACCCCAGAGTCT ↑ GCACTGG
A ↑ GGA GCTGC↓↓ ↑ T ↑ G↓ G CCGT TCA ↑ ↑ GG T ↑ GG A ↑ ↑ ↑ GC T ↑ GG
A ↑ GCC ↑ GG↓TTAA ↑ TG C↓CC↓AAGCC ↑ AGGA AG↓GCCTT ↑ TT ↑ C ↑ TCGGCA
↑ GCGGGAAAAGATGGAGCGGAGGCGCAAGCC↓CCACC↓↓TAGACCGCAGAGGCGCCGT
↓CATCCAGAGCGTCCCTGGCTTCTGG GCCAATGTTGTATCCTTCTCAGTGTTTCTT CG
GCCTTTCTAGTGGAGAGGTGCTCTCG↓GGGAAGTGTAAGTGACCGATGGGCAGCTCGG
CGTCGATGTGACTCTTTGGGGAACAAAGGGGAGTTGCCACGGACCAGTGTG

→ *TSPY* Region 1 - Direction 1 Primer
← *TSPY* Region 1 - Direction 2 Primer

↓= M3529 Direction 1 ↓=M3530 Direction 1 ↓= M3533 Direction 1 ↓= M3552 Direction 1
↑ = M3529 Direction 2 ↑ =M3530 Direction 2 ↑ = M3533 Direction 2 ↑ = M3552 Direction 2

(C)

Figure 4.4 Fetal DNA fragment species with cleavage sites from 4 maternal plasma samples.

(A) *SRY* target region (NT_011896.9). ***(B)*** *TSPY* target region 1 (NT_086998.1). ***(C)*** *TSPY* target region 2 (NT_086998.1). Four maternal plasma samples are represented in different colors. Each arrow represents a unique cleavage site for a particular sample in relation to the strand direction. Since there are multiple copies of *TSPY*, one of the representative genomic regions is presented here.

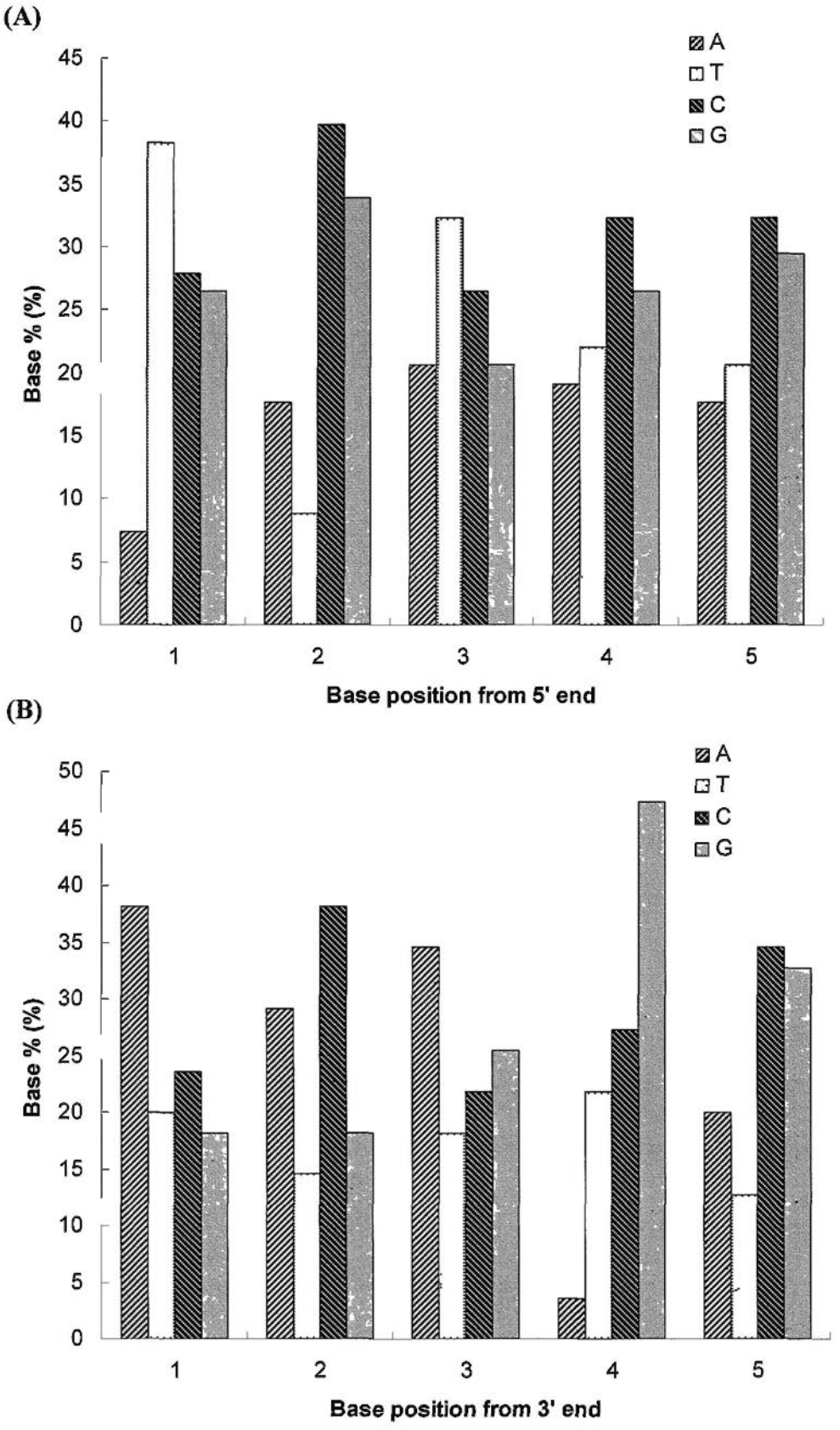


Figure 4.5 Cleavage site of fetal specific DNA fragments in maternal plasma.

Base composition for the adjacent 5 bases at the ***(A)*** 5′ and ***(B)*** 3' termini are shown.

4.3.4 Size distribution of fetal DNA fragments

Unlike the standard QPCR that flanked a target region with a fixed amplicon, the current strategy immobilized one end of fetal DNA fragments but released the other to attain the sequence information. Therefore, although one end of the fetal DNA was fixed by the gene specific primers, the length of fetal DNA fragment species would be informative of the whole size distribution, without the restriction of the target size. The length distribution of all obtained fetal DNA fragment species is shown in Table 4.3. There was no significant difference between the length distribution of fetal DNA fragment species derived from the analyzed *SRY* and *TSPY* regions (Mann-Whitney Rank Sum Test, $P = 0.102$). The relatively short length of fetal DNA fragment species is accordant with the previous observation (Chan *et al.* 2004). No fragment species longer than 400 bp was observed.

Gene specific primer	Number of fragment species	Size distribution (bp)	
		Median	Range
SRY-direction 1	23	82	36-259
SRY-direction 2	16	90	57-316
TSPY-region 1-direction 1	33	76	33-231
TSPY-region 1-direction 2	30	103.5	57-267
TSPY-region 2-direction 1	14	102	31-201
TSPY-region 2-direction 2	64	123	38-327

Table 4.3 Length distribution of fetal DNA fragment species from the *SRY* and *TSPY* assays.

4.4 Discussion

In this study, a touch down LM-PCR assay is designed to clone and sequence the fetal specific DNA in maternal plasma. One can anchor a specific target by a gene specific primer but with freedom to determine the sequence information at the other end. Three aspects of fetal specific DNA are studied in this chapter, i.e., the nature of the DNA end, the end sequences and the relative size distribution.

It is generally believed that circulating fetal DNA is derived from placental trophoblasts (Guibert *et al.* 2003; Ohashi *et al.* 2002; Tjoa *et al.* 2006), a layer of cells that helps the embryo attach to the uterine wall and forms part of the placenta. Necrotic or apoptotic cells from the placenta are probable sources of fetal DNA in maternal plasma (Jones *et al.* 1980; Smith *et al.* 1997). In apoptotic cells, chromosomal DNA is initially cleaved into large fragments of 50–300 kb with intact nucleosomes, followed by degradation into nucleosomal units (180-200 bp) resulting from the activation of endonucleases. Necrosis, on the other hand, produces much larger DNA fragments (Jahr *et al.* 2001; Nagata *et al.* 2003). From our result, there was rarely any fetal DNA fragment species longer than 330 bp, indicating that fetal DNA is unlikely to be derived from necrotic cells, which should generate a much longer fragment size.

During apoptosis, DNA is cleaved into blunt-end, double-stranded fragments carrying a 5' phosphate and a 3' hydroxyl group (Gavrieli *et al.* 1992; Rosl 1992; Staley *et al.* 1997). The unpolished assay could identify the fetal DNA fragments carrying 5' phosphorylated, blunt DNA ends, and the resulting fragment species demonstrated that a part of the fetal DNA fragments in maternal plasma existed with this form of fragment ends. However, pre-treatment of plasma DNA with the T4

DNA polymerase and T4 polynucleotide kinase that altered the state of DNA ends could increase the number of fragment species obtained, implying that a proportion of fetal DNA fragments appeared to have non-blunt ends. Therefore, we speculate that the fetal DNA is present with both blunt and staggered ends in maternal plasma. It is possible that after release from the apoptotic placental cells, the fetal DNA may be subjected to the additional degradation by multiple nucleases in plasma (Tamkovich *et al.* 2006), resulting in various end forms.

The results of cleavage sites, i.e., the nucleotide information of the ends of fetal DNA fragments in maternal plasma, are more complex and seem not to be consistent with the reports from other groups. It has been known that the primary nuclease responsible for oligonucleosomal DNA fragmentation during apoptosis is specific for the cleavage of double-stranded DNA with a preference for A/T-rich region (Khodarev *et al.* 2000; Widlak *et al.* 2000). In a similar work performed by Suzuki *et al.*, the authors showed a particular pattern of end nucleotides of the circulating DNA in healthy individuals, i.e., their 5' and 3' ends were rich in C and G, respectively, along with a gradual increase of A/T content in the adjacent bases (Suzuki *et al.* 2008). However, neither of these observations was obtained in our sequence data. Possible explanations include a fundamental difference in cleavage sites between the circulating DNA in adults and fetal specific DNA in maternal plasma, and the limited assay numbers (only two specific targets on the Y chromosome). It is also possible that the multiple nucleases in plasma degrade the circulating fetal DNA into diverse cleavage sites with no specific pattern observed. Further studies using NGS technologies would hopefully resolve these issues.

Knowing the characteristics of fetal DNA in maternal plasma will benefit the diagnostic application of maternal plasma DNA for noninvasive prenatal diagnosis.

Since the fragment species of fetal DNA molecules vary in cleavage sites and fragment size, it would be challenging for a locus-specific method to precisely quantify the fetal DNA molecules, as such a method would only capture fixed amplicon with a certain fragment size. Investigators therefore seek for a locus-independent platform/technology to achieve a high-precision assessment for fetal DNA analysis. The recently developed NGS platform acts as one of such powerful tools. Chiu *et al.* (Chiu *et al.* 2008) and Fan *et al.* (Fan *et al.* 2008) have demonstrated the use of NGS platforms to massively sequence maternal plasma DNA for fetal chromosomal aneuploidy detection. This promising technology is adopted as an analytical platform for the studies in the following chapters.

SECTION IV : MASSIVELY PARALLEL MATERNAL PLASMA DNA SEQUENCING FOR NONINVASIVE PRENATAL DIAGNOSIS

CHAPTER 5: MATERNAL PLASMA DNA SEQUENCING FOR FETAL TRISOMY 13 AND 18 DETECTION

5.1 Introduction

Since the discovery of fetal cell-free DNA molecules in maternal plasma (Lo *et al.* 1997), maternal plasma DNA analysis has become a valuable molecular diagnostic tool for noninvasive prenatal diagnosis. However, the overwhelming maternal DNA background and low fetal DNA concentration in maternal plasma (Lo *et al.* 1998b) make it challenging for locus-specific DNA assays to detect the small increase in sequences derived from a trisomic chromosome (Chiu *et al.* 2009a). Recently, NGS technologies have been adopted to analyze maternal plasma DNA for fetal trisomy 21 (T21) detection (Chiu *et al.* 2008; Fan *et al.* 2008). By randomly counting DNA fragments in a locus-independent fashion, one could precisely quantify and therefore detect the small proportional increments, i.e. overrepresentation, in maternal plasma chromosome 21 (chr21) molecules contributed by the extra chr21 of the T21 fetus. These studies open a new avenue for assessing fetal aneuploidy precisely and provide a foundation for NGS-based analysis of cell-free DNA in other clinical settings.

Following T21, Edwards syndrome (trisomy 18, T18) and Patau syndrome (trisomy 13, T13) are the second and third most common autosomal trisomies, affecting 1 in 6,000 births and 1 in 10,000 births, respectively (Driscoll *et al.* 2009). Hence, it would be of diagnostic interest to extend the evaluation to the noninvasive prenatal detection of fetal trisomy 18 and 13 using the same approach. However, in our previous study, as reflected by the coefficient of variant (CV), we noticed that the precision for measuring the genomic representation (GR) varied among human

chromosomes and tended to be worse for chromosomes with GC contents at either end of the spectrum (Chiu *et al.* 2008). It was observed that the CVs for chromosome 13 (chr13) and chromosomes 18 (chr18) were larger than that for chr21 (Chiu *et al.* 2008), suggesting that the precision of detecting fetal T13 and T18 by maternal plasma DNA sequencing analysis might be not as good as for T21.

In this study, the maternal plasma samples from three T18 pregnancies and one T13 pregnancy as well as four euploid male pregnancies are recruited and subjected to plasma DNA sequencing. The objective of this chapter is to analyze the sequencing data from these samples to investigate whether the same approach could be applied to diagnose fetal T18 and T13 accurately. The effect of the GC content on the detection sensitivity is studied and a region-selection method is proposed to minimize the quantification bias caused by GC bias.

5.2 Methods

5.2.1 Subjects

Peripheral blood was collected according to the description in Chapter 3.1.1. Blood samples were collected from 8 pregnant women in the first or second trimester, among which there were three women with T18 male pregnancies, one woman with T13 male pregnancy and four women with euploid male pregnancies.

5.2.2 Sample preparation

Plasma was harvested from ~10 mL of blood samples as described in Chapter 3.1.2. DNA was extracted from the plasma samples according to procedures described in

Chapter 3.2.1. The extracted plasma DNA was subjected to QPCR as described in Chapter 3.3.

5.2.3 Massively parallel sequencing of maternal plasma DNA

The massively parallel SR sequencing of plasma DNA was performed on the Illumina GA I system, following the procedures described in Chapter 3.5. The library preparation and sequencing process were performed by Yuan Gao at the Center for High Performance Computing, Virginia Commonwealth University.

5.2.4 Sequence alignment

All 36 bp sequence reads were aligned to the repeat-masked human genomic reference sequences (NCBI Build 36, version 48) downloaded from the Ensembl Genome Browser (http://www.ensembl.org) using the ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) program in the GAPipeline-0.2.2.5 software suite (Illumina), as described in our previous paper (Chiu *et al.* 2008). Basically, a result output file (*.eland_result.txt) was generated after running ELAND, in which code U0 on the third field of the output file indicated that the best match found was a unique match in the repeat-masked human reference genome. A sequence with codes 1 in the fourth, 0 in the fifth and 0 in the sixth fields (hence U0-1-0-0) indicated that it had just a single exact match in the repeat-masked human reference genome without any nucleotide mismatch. U0-1-0-0 sequence reads are thought to be the most exact unique matches and hence selected for the subsequent quantitative analysis. U0-1-0-0 sequences in each chromosome were sorted and counted using the awk utility in Linux for further analysis.

5.2.5 *Z*-score calculation

To quantify the GR of each chromosome in plasma DNA, the number of U0-1-0-0 sequences mapped to each chromosome was counted and then expressed as a percentage of all U0-1-0-0 sequences generated for the sample (expressed as %U0-1-0-0), which is the obtained %GR of each chromosome in maternal plasma. *Z*-scores were calculated as previously described (Chiu *et al.* 2008). The mean and standard deviation (SD) of the %GR of target chromosomes (e.g., chr13, chr18, chr21 and chrX) were calculated using the data from male euploid pregnancies and were denoted as $mean_{chr}$ and SD_{chr}, respectively. A chrN (standing for the target chromosome) *z*-score for a test sample would be generated by subtracting the mean %chrN of a reference set of euploid pregnancies from the %GR of chrN (%chrN) of the test case and divided by the SD of the %chrN values among the reference set according to the equation:

chrN *z*-score$_{\text{test case}}$= (%chrN$_{\text{test case}}$– mean %chrN$_{\text{reference controls}}$) / SD %chrN$_{\text{reference controls}}$.

A *z*-score value greater than 3, representing a %chrN value greater than that of the 99.9th percentile of the reference set for a one-tailed distribution, was used as the cut-off to determine if overrepresentation of chr13/18/21 plasma DNA molecules and hence fetal T13/T18/T21 was present.

5.2.6 Calculation of the genomic representation of each chromosome in the reference human genome

The expected %GR in the reference human genome is calculated according to the previously established method (Chiu *et al.* 2008). Since in maternal plasma, the majority of DNA fragments are of maternal origin, a haploid female genome was used to calculate the expected chromosome size. The reference sequences (NCBI Build 36, version 48, repeat-masked) for each human chromosome were downloaded

from the Ensembl Genome Browser (http://www.ensembl.org). The expected %GR of each chromosome was obtained by dividing the repeat-masked nucleotide counts per chromosome by the total repeat-masked nucleotide counts of all chromosomes except chrY. In some scenarios, a male genome is required for reference. The expected chromosome size for a diploid male genome was calculated similarly, except that one copy of chrX and one copy of chrY were included together with two copies of autosomes.

5.2.7 Calculation of fetal DNA fraction from %chrY

The %GR of chrY (%chrY) was calculated and was used to determine the fetal DNA concentration in maternal plasma samples collected from male pregnancies. Our group has previously reported that a small fraction of reads would be misaligned to chrY in pregnancies with female fetuses (Chiu *et al.* 2008). Hence, the maternal plasma %chrY value in a pregnancy with a male fetus is a composite of the amount of chrY sequences contributed by the male fetus and those sequences from the maternal background DNA that were misaligned to chrY (Chiu *et al.*, submitted). In our cumulative data, the mean %chrY value of plasma samples obtained from four adult male individuals (containing 100% male DNA) was 0.157%. The mean %chrY value in the plasma of all the women carrying euploid female fetuses (containing 100% female DNA) was 0.007%. Hence, for a male pregnancy with a certain fetal DNA fractional concentration in the maternal plasma (indicated as F), the contributions of the male fetus and the maternal background DNA to the maternal plasma %chrY value could be represented as 0.157F and 0.007(1-F), respectively. Thus, the fetal DNA fractional concentration (F) can be derived from the equation: %chrY = 0.157F + 0.007(1-F).

5.2.8 GC content counting and region-selection method

The G+C content was calculated by an in-house perl program, in which each nucleotide (i.e., A, T, G and C) within the target chromosome (or region) in the reference genome was counted, respectively, and then the GC content was obtained by dividing the G+C counts by the total A+T+G+C counts in the particular chromosome (or region). If a region was required to be deleted from the whole dataset, the sequence reads aligned to that region would be discarded according to the chromosomal coordinates shown in the *.eland_result.txt files. Afterwards, the %GRs were recalculated based on the remaining sequence reads. The mean and SD of %chrN were recalculated using the same euploid samples, and the *z*-scores for the test samples were regenerated as described above.

5.3 Results

5.3.1 Massively parallel sequencing of maternal plasma DNA

A median of 10.3×10^6 raw reads (SD, 1.1×10^6) from each sample were obtained. Since the majority of maternal plasma DNA is of maternal origin, a repeat-masked haploid reference genome of a female was used as reference genome. A median of 2.3×10^6 U0-1-0-0 reads per sample (SD, 2.1×10^5), representing ~23% of the raw reads, were obtained. With a read length of 36 bp, 2.3×10^6 unique reads would be equivalent to 2.8% coverage of the haploid human genome. The clinical information and sequence reads of eight cases are shown in Table 5.1.

Case no.	Fetal sex	GA (weeks + days)	Input DNA (ng)	Karyotype	Total sequenced counts	Total U0-1-0-0 counts	Chr13 U0-1-0-0 counts	Chr18 U0-1-0-0 counts	ChrY U0-1-0-0 counts
M1819	M	17+3	11	47XY+13	10,729,645	2,163,727	78,352	62,602	822
M3294	M	13+3	17	47XY+18	10,952,000	2,402,010	77,970	72,471	561
M3455	M	13+2	11	47XY+18	14,320,253	2,543,034	92,383	80,807	805
M3845	M	12+6	20	47XY+18	11,529,426	2,418,882	82,528	72,830	501
M4198	M	13+6	20	46XY	9,891,031	2,347,481	79,267	68,479	647
M4206	M	18+2	20	46XY	9,096,160	2,028,401	69,120	59,237	488
M4480	M	20+3	20	46XY	9,157,938	1,877,598	63,121	54,844	886
M4523	M	13	20	46XY	8,493,969	1,913,611	62,043	54,859	624

Table 5.1 Clinical details and sequencing counts of eight maternal plasma samples in the current sequencing run.

5.3.2 Distribution of maternal plasma DNA sequences among the human chromosomes

The expected %GR of each chromosome is plotted alongside the %U0-1-0-0 per chromosome for the four sequenced maternal plasma DNA samples obtained from the euploid male pregnancies (Figure 5.1A). The obtained %chrX was lower than the expected value owing to only one dose of chrX contributed by the male fetus (Figure 5.1A). On the other hand, the signals from chromosome Y were detected and expressed as %chrY, ranging from 0.024% to 0.047% in those euploid male pregnancies (not visible in Figure 5.1A, due to the too small values). Linear regression analysis was performed to compare the expected %GRs and the mean of the obtained %GRs of the autosomes from the four euploid male pregnancies. As shown in Figure 5.1B, the slope was 0.9765 and R^2 was 0.9921. These data suggested that the overall DNA molecules in the maternal plasma (inclusive of both maternal and fetal DNA) distributed quite evenly across the human genome.

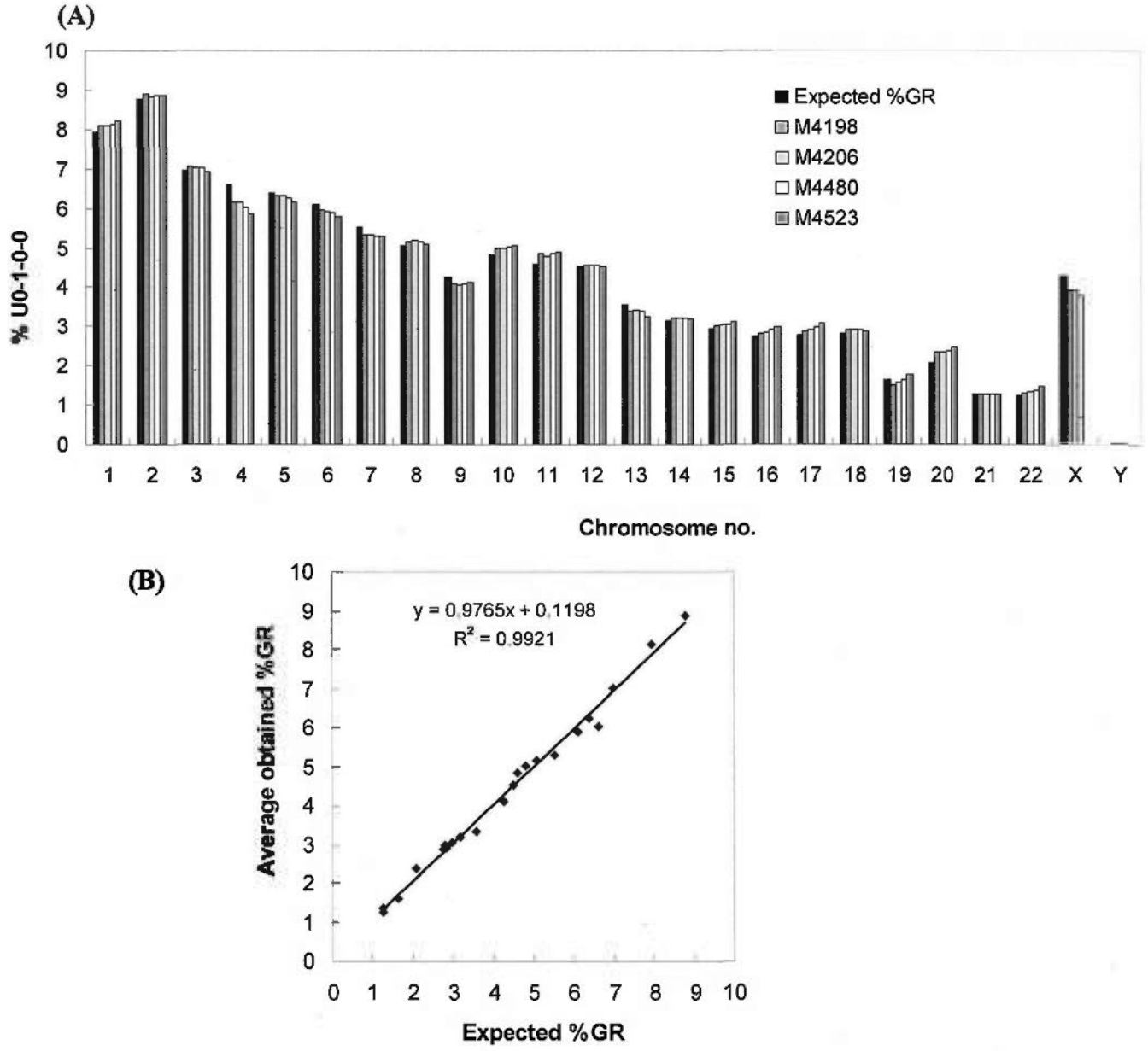


Figure 5.1 Distribution of maternal plasma DNA sequences among the human chromosomes.

(A) Bar chart of %U0–1–0–0 sequences per chromosome for the 4 maternal plasma samples from the euploid male pregnancies. The percentage of genomic representation of each chromosome as expected for a repeat-masked reference haploid female genome is plotted for comparison (black bars). *(B)* Linear regression plots of the average obtained %GRs (%U0–1–0–0) against the expected %GRs of the autosomes. The values on y axis were calculated from the four euploid male cases in the current dataset. The chrX is excluded, as the obtained %GR of chrX is affected by the fetal DNA proportion in maternal plasma from the male pregnancies.

5.3.3 Measurement precision

Although quite close to the expected %GRs, the obtained %GRs of all autosomes deviated from the expected values in varying degrees (Figure 5.1A). Here the ratio of the obtained %GR to the expected %GR for each chromosome, termed the "measured ratio", was introduced to assess whether the sequencing data represented, without bias, the expected chromosomal contribution. Assuming that the plasma DNA derived from both the maternal and fetal genomes distributed evenly across the entire genome, the measured ratios for all autosomes in the plasma from a euploid male pregnancy would be around 1. The human genome has a GC content that is below 50%, but different between the individual chromosomes (Kel-Margoulis *et al.* 2003). The GC content of each chromosome in the reference genome (NCBI 36.48, repeat-masked version) is listed in Table 5.2. As shown in Figure 5.2A, the average measured ratios were variable among the autosomes but correlated with the chromosomal GC content (Pearson correlation analysis, $r = 0.621$, $P = 0.00203$).

In terms of the chromosomal GC content, the autosomes plus chrX can be distributed into five groups (Kel-Margoulis *et al.* 2003) (Table 5.2). As previously reported (Chiu *et al.* 2008), there was a quantification bias in measuring the %GRs of difference chromosomes and a tendency of being worse for the chromosomes with GC contents at either end of the GC-abundance spectrum, namely group I and group V, when employing the Illumina sequencing platform for plasma DNA sequencing. In this sequencing run, there were four euploid male cases, which could be used to constitute an intra-flow cell reference group. As the SD of a data set was in fact reflecting the precision of its measurement, the coefficient of variation (CV = SD/mean × 100%) of measuring the %GR of each chromosome was calculated from the four euploid male fetuses to evaluate the measurement precision. The CV plot

showed the same distribution pattern as previously reported (Chiu *et al.* 2008), but were generally lower than previous data (Figure 5.2B). It was possible that the relatively smaller size of reference cases in the current dataset did not adequately reflect the analytical precision of the reference population. Since both the previous and current datasets were generated by the same core lab (Center for High Performance Computing at the Virginia Commonwealth University) according to the same library construction and sequencing protocols, all euploid male cases together (10 from the previous dataset and 4 from the current dataset) were combined as an inter-flow cell reference group for downstream analyses. As a result, the CVs calculated from this new reference group were similar to the previous data (Figure 5.2B). The sample information and sequence reads of previously sequenced 10 maternal plasma samples are given in Appendix I.

Chromosome no.	Chromosomal GC content (%)	Group
4	36.53	I
13	37.10	I
6	38.60	II
5	38.61	II
3	38.79	II
X	38.96	II
18	39.07	II
8	39.45	II
2	39.67	II
12	39.80	II
7	39.94	II
21	40.20	II
14	40.33	II
9	41.03	III
10	41.41	III
1	41.52	III
11	41.89	III
15	42.35	III
20	44.65	IV
16	45.22	IV
17	46.05	IV
22	49.89	V
19	50.65	V

Table 5.2 GC content of each chromosome.

Chromosomes are grouped based on the similarity in the GC content (Kel-Margoulis *et al.* 2003) and listed in an ascending order of chromosomal GC content.

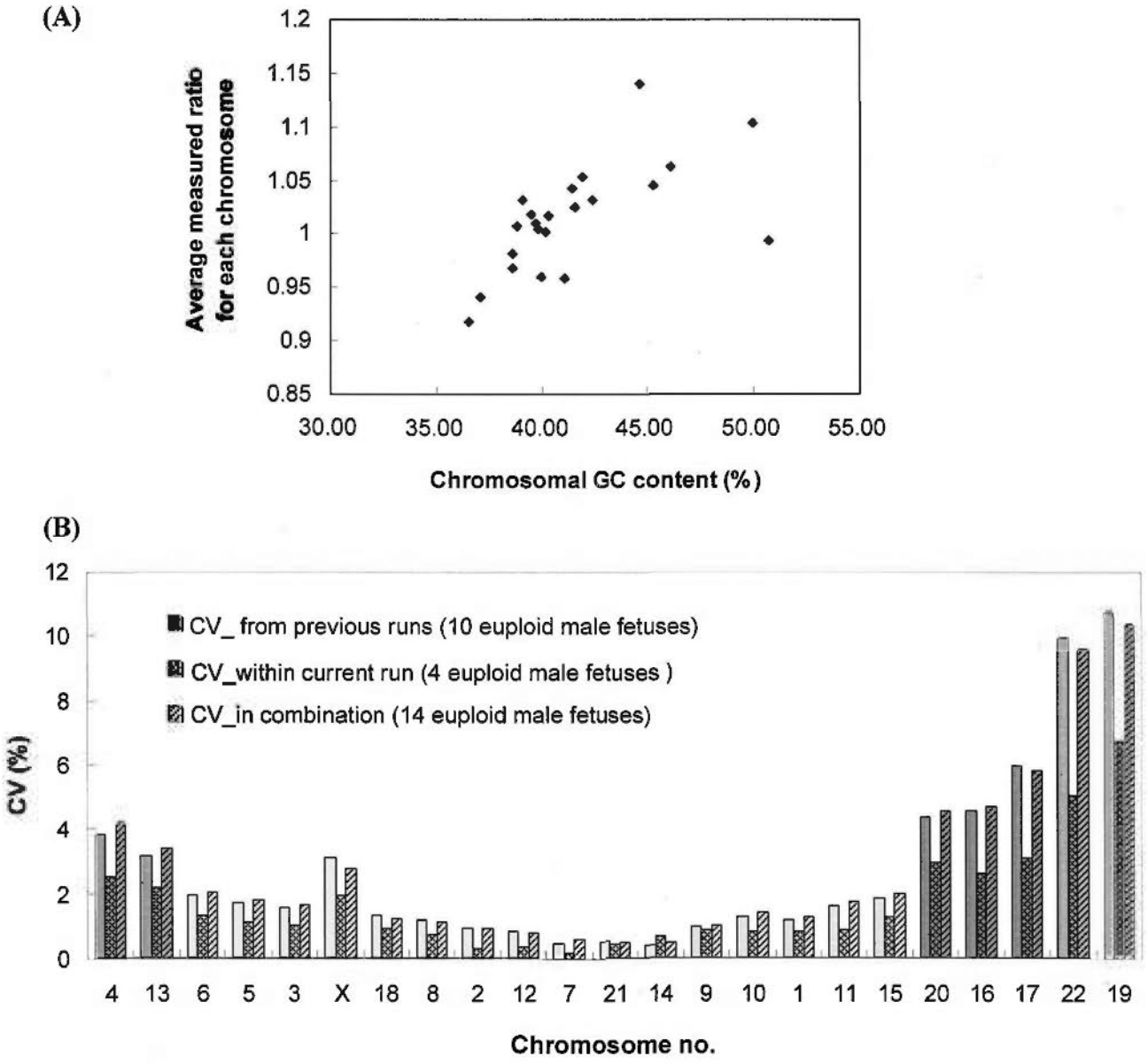


Figure 5.2 Sequencing bias and variation among chromosomes.

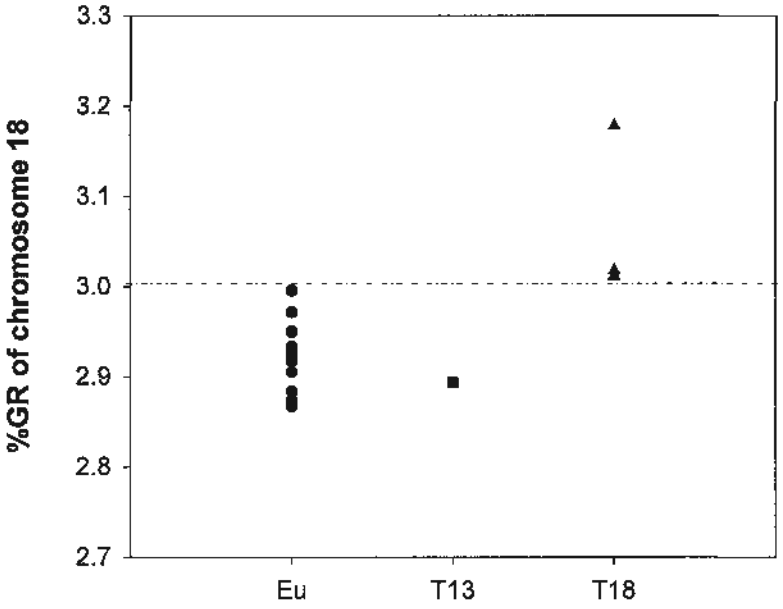
(A) Correlation between the measured ratios and chromosomal GC content for autosomes. The values on the y axis were calculated from the four euploid male cases in the current dataset. The chrX is excluded, as the %chrX is affected by the fetal DNA proportion in maternal plasma from male pregnancies, apart from its GC content. *(B)* CV per chromosome for different reference groups. Chromosomes are grouped according to their GC contents (Table 5.2) and each group is represented by one color. Group I chromosomes have the lowest GC contents while group V chromosomes have the highest GC contents.

5.3.4 Fetal trisomy 18 detection

I proceeded to test if fetal T18 would lead to quantitative perturbation in the %GR of chr18 (%chr18). As shown in Figure 5.3A, %chr18 for the three T18 fetuses were observed to be larger than those for the non-T18 fetuses. To objectively quantify the degree of such overrepresentation, chr18 *z*-scores were calculated using the 14 euploid male fetuses as a reference group. The chr18 *z*-scores for the three T18 cases were 2.59, 7.06 and 2.41, respectively, and all non-T18 cases had a chr18 *z*-core less than 3 (Figure 5.3B). With a *z*-score of 3 being a diagnostic cutoff, one of three T18 cases could be correctly identified with a specificity of 100%.

The extent of the overrepresentation of an at-risk chromosome in maternal plasma for trisomy cases is governed by the fractional fetal DNA concentration. As calculated from %chrY, the fetal DNA proportions for the 3 T18 cases were 10.90%, 16.44% and 9.14%, respectively. The two T18 male fetuses that failed to be detected by chr18 *z*-scores had relatively lower fetal DNA concentrations (around or less than 10%). Although these two T18 cases had higher %chr18 values (Figure 5.4A), they could not be distinguished from the euploid group with 99.9% confidence, thus resulting in *z*-scores of less than 3. The slightly larger CV of chr18 (1.23, Figure 5.3) would be another reason for the failure for trisomy 18 detection.

(A)



(B)

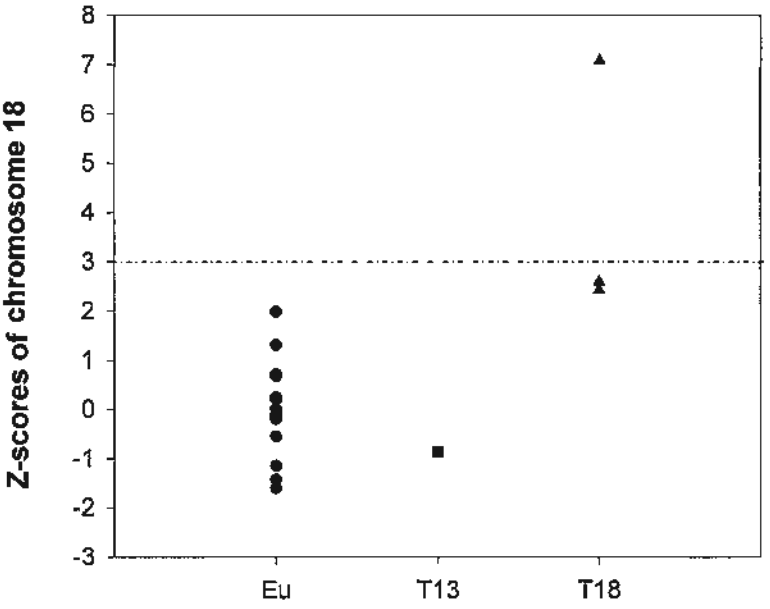


Figure 5.3 Fetal trisomy 18 detection by maternal plasma DNA sequence analysis.

(A) %GR of chr18 for the test samples (1 T13 male fetus and 3 T18 male fetuses) and reference samples (14 euploid male fetuses, indicated as "Eu"). The dashed line represents the boundary of %chr18 between T18 and non-T18 cases. ***(B)*** Chr18 *z*-scores for the test samples and reference samples. The dashed line represents the diagnostic cutoff with a *z*-score of 3.

5.3.5 Fetal trisomy 13 detection

Next, I tested if fetal T13 would lead to quantitative perturbation in the %GR of chr13 (%chr13). As shown in Figure 5.5A, the %chr13 for T13 and non-T13 fetuses were blurred. After translating the %chr13 into the chr13 *z*-score using the same reference groups as above, the chr13 *z*-score for the T13 fetus was only 1.48 (Figure 5.4B). Using a *z*-score of 3 as a diagnostic cutoff, this T13 fetus was not correctly classified.

One theoretical reason for a non-diagnostic %chr13 value for a T13 fetus is the low fetal DNA concentration of such a case in maternal plasma. However, the fetal DNA proportion calculated from %chrY for this case was 20.66%, which was not particularly low, suggesting that there could be other factors affecting T13 detection. On the other hand, the broad distribution of %chr13 values among the controls reflected the large analytical variation in %chr13 measurement (Figure 5.4A). The CV for chr13 was 3.41% (Figure 5.2B), ranking 17th among autosomes plus chrX. The imprecision of %chr13 measurement implied that the detection of fetal T13 by maternal plasma DNA sequencing analysis would be challenging. Further optimization and other strategies are needed to be developed.

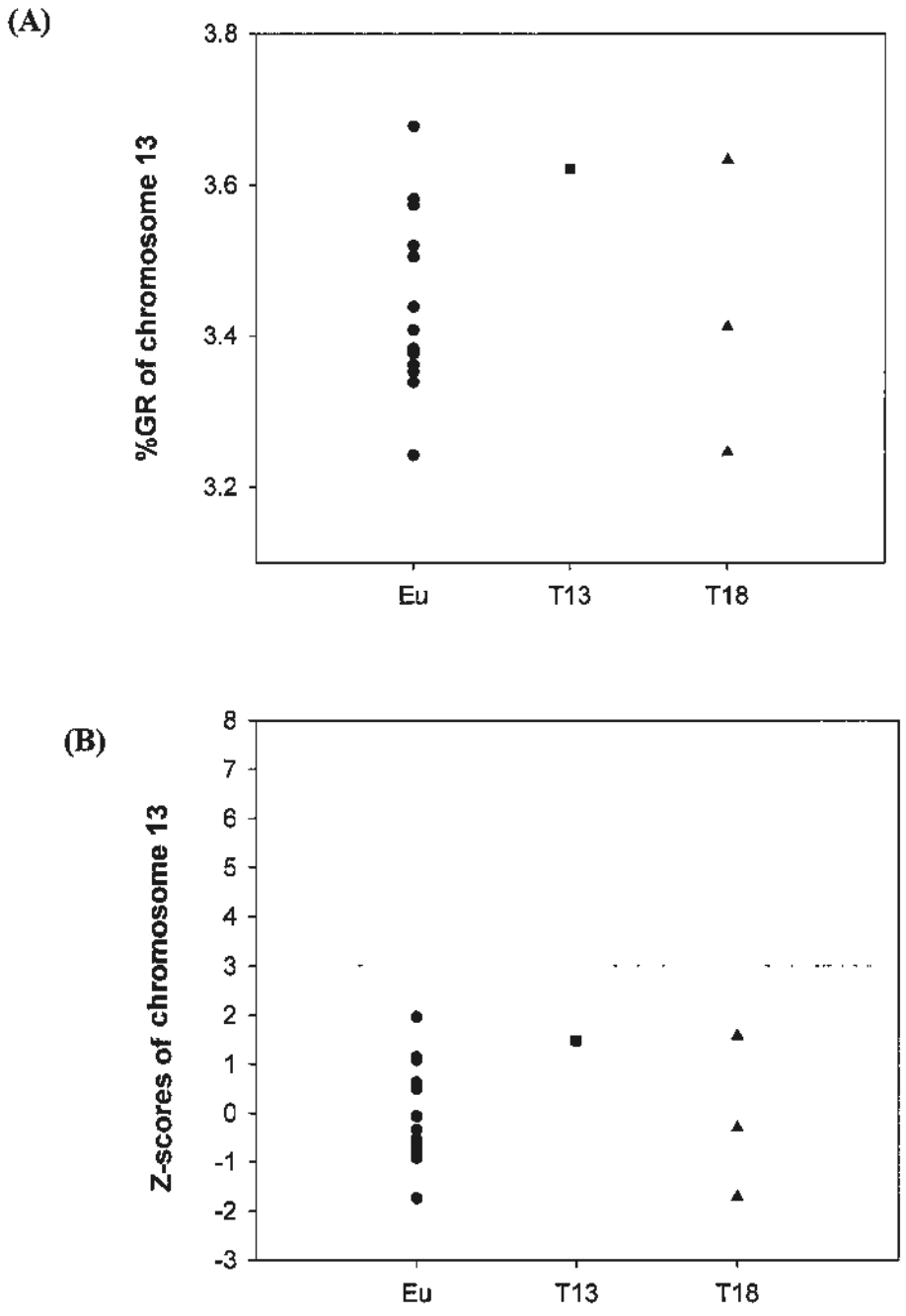


Figure 5.4 Fetal trisomy 13 detection by maternal plasma DNA sequence analysis.

(A) %GR of chr13 for the test samples (1 T13 male fetus and 3 T18 male fetuses) and reference samples (14 euploid male fetuses, indicated as "Eu"). ***(B)*** Chr13 *z*-scores for the test samples and reference samples. The dashed line represents the diagnostic cutoff with a *z*-score of 3.

5.3.6 Region-selection method to minimize GC bias

DNA sequence analysis demonstrates that chr13 contains a central 'gene desert' region of 37.8 Mb, where the gene density drops to only 3.1 genes per Mb (Dunham *et al.* 2004). The GC content of this region in the reference chr13 (NCBI Build 36.48, repeat-masked) drops to 33.22%. In contrast, the most gene-rich regions are at either end of the long arm of this chromosome, which possess the regional GC contents of 39.48% and 39.66%, respectively. Accordingly, the long arm of chr13 could be divided into three regions: region 1 (16.9~52.9 Mb), region 2 (52.9~90.7 Mb) and region 3 (90.7~114.1 Mb). The sequence read density from a maternal plasma sample has obvious discrepancies across the three regions but correlates with the GC content (Figure 5.5). The mean, SD, and CVs for the three regions were calculated respectively from the sequence data of 14 reference samples. It was found that the mean of the obtained %GR for region 2 substantially deviated from the expected value and the CV for region 2 was largest among three regions (Table 5.3), suggesting this region contributed mostly to the large variation in %chr13 quantification.

Therefore, a region-selection method was proposed to minimize the effects caused by the variations in GC content on the measurement of %chr13 from plasma DNA sequencing. The U0-1-0-0 sequence reads mapped to region 2 were discarded, whereas those mapped to region 1 and region 3 were retained to represent the %GR of the so-called "new" chr13. Encouragingly, the CV for chr13 decreased from 3.41% to 1.03% after removing region 2 (Table 5.3), demonstrating the effectiveness of this region-selection method in lowering the measurement imprecision of %chr13. The %chr13 for that T13 case became higher than for the non-T13 cases (Figure 5.6A). As a result, the chr13 *z*-score for the T13 case increased from 1.48 to 6.43

(Figure 5.4B and 5.6B), correctly identifying the T13 fetus with a *z*-score of 3 as the diagnostic cutoff.

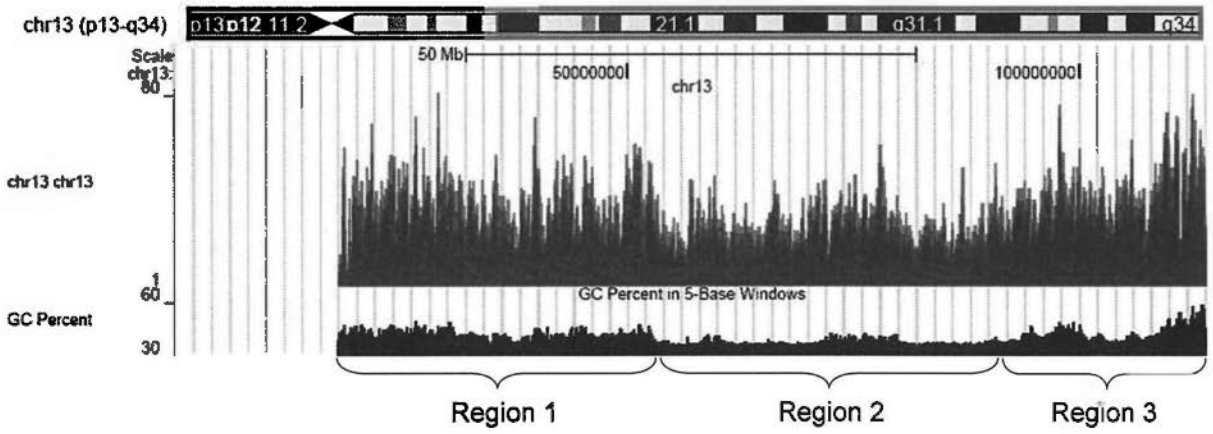


Figure 5.5 Sequence read distribution and regional GC content.

The short arm of chromosome 13 is heterochromatic, while the long arm is euchromatic and contains most or all of the protein-coding genes of the chromosome. The long arm of chromosome 13 measures 97.2 Mb (chr13:16,900,000-114,142,980, 13q11-13q34) and can be divided into three regions in terms of gene density (Dunham *et al.* 2004). The sequence read density (blue bars) and GC content (black bar) of each bin (10,000 bp) are shown for a plasma DNA sample from a male euploid pregnancy.

Region	Location (Mb)	Nucleotide content (bp)	GC content (%)	Expected %GR	Obtained %GR [a]		
					Mean	SD	CV (%)
Region 1	16.9~52.9	18,097,024	39.48	1.26	1.27	0.017	1.34
Region 2	52.9~90.7	19,702,811	33.22	1.38	1.22	0.100	8.16
Region 3	90.7~114.1	13,135,476	39.66	0.92	0.95	0.007	0.73
All	16.9~114.1	50,935,311	37.10	3.56	3.45	0.12	3.41
Region 1+3	16.9~52.9, 90.7-114.1	31,232,500	39.55	2.21 [b]	2.25	0.023	1.03

a: the obtained values were calculated from the reference group consisting of 14 euploid male fetuses.
b: the expected %GR of chr13 (region 1 and 3) was updated by deleting region 2 from the nucleotide content of the haploid reference genome.

Table 5.3 Summary statistics of the three regions within the long arm of chromosome 13.

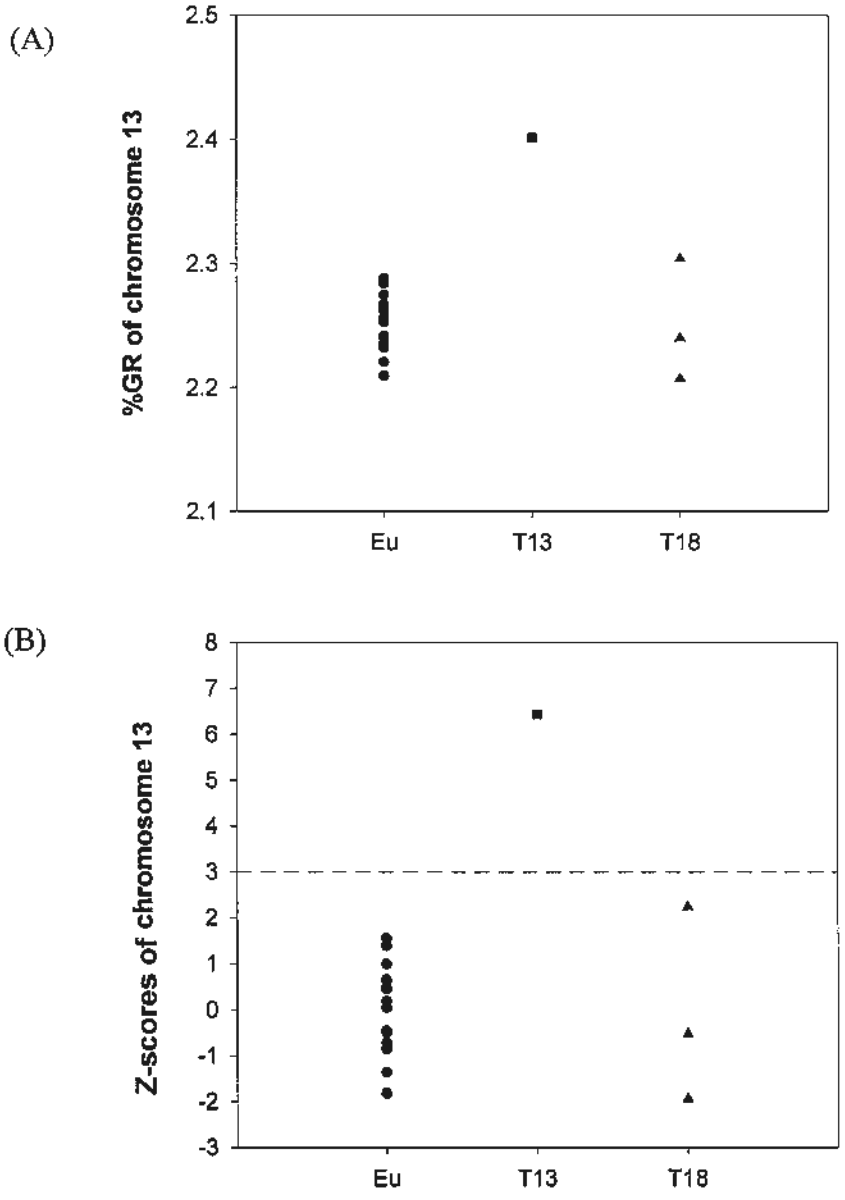


Figure 5.6 Fetal trisomy 13 detection by region-selection method.

(A) %GR of the new chr13 for the test samples (1 T13 male fetus and 3 T18 male fetuses) and reference samples (14 euploid male fetuses, indicated as "Eu"). ***(B)*** New chr13 *z*-scores for the test samples and reference samples. The dashed line represents the diagnostic cutoff with a *z*-score of 3.

5.4 Discussion

Massively parallel sequencing of maternal plasma DNA has been demonstrated to be a promising approach for fetal trisomy 21 detection. In this chapter, I extend the evaluation to noninvasively detect fetal triosmy 18 and 13 by applying the same approach. As predicted from the *z*-score equation, the detection of a trisomic fetus by *z*-score calculation is dependent on the %GR of the at-risk chromosome for the test case and the spread (i.e. SD) of the distribution of %GR of that chromosome among controls (Chiu *et al.*, submitted). The first variable is governed by the fractional fetal DNA concentration, and the latter one depends on the analytical and biological variations for measuring the %GR among euploid pregnancies and whether the size of the control group adequately reflects these variations (Chiu *et al.*, submitted). From the results, both variables impacted on fetal trisomy 18 detection. However, the T13 case, even with a high fetal DNA proportion, still escaped detection, probably as a consequence of the lack of the precision for measuring the %GR of chr13 when using the Illumina sequencing platform. In spite of the limited case number, the results presented here suggested that it would be more challenging for the noninvasive prenatal diagnosis of fetal trisomy 13 by maternal plasma DNA sequencing analysis, when compared with fetal trisomy 21.

A correlation between the measured ratio and chromosomal GC content was observed from these results (Figure 5.2A). Also, the precision for measuring %GR of chromosomes was variable and associated with the chromosomal GC content (Figure 5.2B). These observations regarding GC bias are consistent with previous reports (Chiu *et al.* 2008; Fan *et al.* 2008). The GC bias of the Illumina sequencing platform for genomic sequencing has been reported by several groups (Dohm *et al.* 2008; Hillier *et al.* 2008). For maternal plasma DNA sequencing, it is unclear at this point

where the GC bias exactly stems from. It is possible that such GC bias arises from the PCR artifacts during DNA library preparation (with a 15-cycle enrichment PCR included) or cluster generation (solid-phase bridge amplification for the formation of clonal clusters) or the sequencing process (the incorporation bias in A/T/G/C nucleotides). Another possibility is that there may be biological reasons, i.e., certain regions within genome are prone to be more fragmented. However, the recent data from other sequencing platforms (e.g. the SOLiD sequencing platform), where the different pattern of GC bias was observed (Chiu *et al.* 2009b), suggest that the non-uniform representation is more likely explained by analytical bias than biological factors.

A distinctive feature of chromosome 13 is that the GC content within the whole chromosome varies largely (Dunham *et al.* 2004). This observation may partially explain why chromosome 13 is more susceptible to the GC bias of the Illumina sequencing platform and consequently has a large CV. Based on the rationale of standardizing the GC content within chromosome 13, I demonstrated that the region-selection method could effectively increase the sensitivity of aneuploidy detection and enhance the precision for the noninvasive prenatal diagnosis of trisomy 13. On the other hand, chromosome 18 is one of the low-GC chromosomes and contains 24 gene deserts with relatively low GC content (Nusbaum *et al.* 2005). Theoretically, such a region-selection method could also be applied to the measurement of chromosome 18-derived DNA sequences by removing these GC-low regions, so as to improve the diagnostic performance of fetal trisomy 18 detection.

Built on the same conception, other comprehensive and in-depth analyses could be carried out with the help from bioinformaticians. For example, one could divide the whole genome into multiple bins and consider the GC content of each bin as a

variable. Afterwards, one could recalculate the %GR and *z*-scores either by retrieving the regions with a defined range of GC content or normalizing the sequence read count of each bin in terms of its GC content. In addition, the correlation between the measured ratio and the chromosome GC content (Figure 5.2A) suggests another way to eliminate the GC bias by taking this relationship into account. Both approaches have recently been achieved (Chu *et al.* 2009; Fan *et al.* 2010) and will facilitate the diagnostic application of the NGS-based methods for fetal trisomy 13 and 18 detection.

The GC content of template DNA is believed to play an important role in PCR (Benita *et al.* 2003). The variation in sequence composition of the human chromosomes would inevitably cause the quantification bias for different chromosomes when using the NGS platforms that require an amplification for library preparation. With the emergency of the third-generation sequencing platforms, such as the Helicos platform (Harris *et al.* 2008), nanopore sequencing (Lund *et al.* 2009) and so on, which are free of PCR amplification, maternal plasma DNA sequencing analysis would hopefully overcome such quantification bias and clinical application would become more robust.

CHAPTER 6: FURTHER INVESTIGATION INTO MASSIVELY PARALLEL SEQUENCING OF MATERNAL PLASMA DNA FOR CLINICAL IMPLEMENTATAION

6.1 Introduction

In the pioneering work of applying NGS platform to noninvasive prenatal diagnosis, our group have demonstrated the use of the Illumina GA platform for the prenatal diagnosis of trisomy 21 by maternal plasma DNA sequencing analysis from 28 first and second trimester pregnancies. All 14 trisomy 21 fetuses and 14 euploid fetuses were correctly identified (Chiu *et al.* 2008).

Although promising, further investigations are needed to prepare such technology for future clinical implementation. In the work presented in previous study (Chiu *et al.* 2008) and the last chapter, DNA from 4~5 mL of maternal plasma were collected to obtain a sample containing a representative profile of DNA molecules in maternal plasma. As plasma occupies 50%~55% of blood, 4~5mL of maternal plasma necessitates collecting 8~10 mL maternal blood, thereby making it time-consuming and labor-intensive for blood processing and plasma DNA extraction. Nevertheless, starting with a limited volume of plasma would increase the risk of failure in library preparation in view of the multiple column-purification steps involved. More importantly, on account of the low abundance of fetal DNA in maternal plasma (Lo *et al.* 1998b), less starting molecules in the maternal plasma samples would be less representative of the overall plasma DNA profiles for detecting the quantitative perturbation in the %GR of the target chromosome, thus giving rise to false diagnoses. Hence, there is a need to investigate the effect of the input plasma volume

on the genomic representation of maternal plasma DNA and the resultant diagnostic performance when using massively parallel maternal plasma DNA sequencing.

On the other hand, the NGS platforms impose inherent limitations on the high cost per sample and low sample throughput. Taking the Illumina sequencing system as an example, the sequencing is performed on a flow cell physically containing eight chambers (lanes), which means that at most eight samples could be processed in parallel. Apart from sample preparation, a typical sequencing run on an Illumina GA II currently costs about US$500 in reagents per lane and requires 3~4 days to complete both the sequencing and the Illumina pipeline analysis phases. These limitations present obvious obstacles in the clinical implementation of maternal plasma DNA sequencing. One approach to overcome these limitations is to barcode samples with sample-specific sequence tags (i.e., indices) prior to sequencing and to identify each sample after sequencing by their unique sequence tags. Multiplexed DNA sequencing has been pursued since the beginning of Sanger sequencing (Church *et al.* 1988) and has been applied to Roche's 454 platform (Meyer *et al.* 2007) and the Illumina GA platform (Cronn *et al.* 2008). For maternal plasma DNA sequencing, one could potentially reduce the cost and enlarge the sample throughput by barcoding individual patients' samples such that one sequencing reaction could simultaneously generate diagnostic information for multiple cases (Chiu *et al.* 2008).

The objective of this study is to address each of the issues mentioned above. By gradually reducing the input volume of maternal plasma, the effect of plasma volume on the chromosomal representation of maternal plasma DNA is studied. Barcoding strategy is employed to process multiple maternal plasma samples in parallel. Two levels of multiplexing, 4-plex and 8-plex sequencing, are evaluated.

6.2 Methods

6.2.1 Subjects

The maternal plasma samples prefixed with "U" were recruited from UK with the kind help from Prof Kypros Nicolaides and Dr Ranjit Akolekar at the Harris Birthright Research Centre for Fetal Medicine, King's College Hospital, London, UK. The local ethical approval was obtained. Specimens were transported frozen to Hong Kong on dry ice. Other maternal plasma samples were collected from the Department of Obstetrics and Gynaecology at the Prince of Wales Hospital, Shatin, Hong Kong, according to the description in Chapter 3.1.1.

6.2.2 Sample preparation

Plasma was harvested from blood samples as described in Chapter 3.1.2. DNA was extracted from plasma samples according to procedures described in Chapter 3.2.1. For the evaluation of starting plasma volume, a total of ~10 mL of maternal plasma per case was collected and extracted. For the multiplexed sequencing, 3.2 mL of maternal plasma of each case was collected and extracted. The extracted plasma DNA was subjected to QPCR as described in Chapter 3.3.

6.2.3 Massively parallel sequencing of maternal plasma DNA

The massively parallel SR sequencing of plasma DNA was performed on the Illumina GA II system, following the procedures described in Chapter 3.5.

6.2.4 Massively parallel multiplexed sequencing of maternal plasma DNA

With the introduction of barcodes, one can sequence multiple samples simultaneously in a sequencing chamber (lane), whereby the analytic throughput could be increased. Two levels of multiplexing, 4-plex and 8-plex, were studied whereby DNA from four or eight maternal plasma samples, respectively, were co-sequenced in each specimen lane. As each sequencing flow cell has eight specimen lanes, 32 and 64 specimens can be sequenced in each run of the 4-plex and 8-plex protocols, respectively.

Multiplexing was achieved by introducing a characteristic 6-bp index barcode to the DNA molecules of each plasma sample through a triple-primer PCR amplification (http://www.illumina.com/). Sequenced reads from the respective plasma samples could be identified by additionally sequencing the index sequences. Indexed DNA libraries were prepared according to manufacturer's instructions with the use of the Multiplexing Sample Preparation Oligonucleotide kit (Illumina). The protocol for library preparation is similar to the description in Chapter 3.5.2, except that in the enrichment PCR step the triple primers (Illumina) were input and 18 cycles of PCR amplification were performed. After library preparation, the amplified products from multiple subjects were mixed at equal amount and then 36 pM of the mixed DNA was introduced to one lane of a flow cell. After cluster generation, 36 cycles of sequencing were performed on the GA II (Illumina), followed by an additional 7 cycles of sequencing to read out the 6-bp index sequences.

6.2.5 Sequence analysis for massively parallel sequencing of maternal plasma DNA

For the SR sequencing data, the sequence reads were aligned to the repeat-masked reference genome as described in Chapter 5.2.4, except that the Pipeline suite was

updated to GAPipeline-1.0. *Z*-score calculation was performed according to the description in Chapter 5.2.5.

6.2.6 Sequence analysis for massively parallel multiplexed sequencing of maternal plasma DNA

For the multiplexed sequencing data, the sequence reads were aligned to the repeat-masked reference genome using the ELAND program in the GAPipeline-1.0 software suite. In the output sequence files, there would be a column indicative of the index sequences to facilitate the identification of each barcoded sample. An in-house algorithm was compiled by Prof Hao Sun and Mr Zhang Chen based on the previously described strategy (Chiu *et al.* 2008). Briefly, the algorithm would decode each sample according to the characteristic 6-bp index sequences and identify sequenced reads that were aligned perfectly with no nucleotide mismatch to a single location in the human genome. These were termed U0-100 reads for each sample. The %chr21, %chrX and %chrY were then calculated. The subsequent *z*-score calculation was similar to that in Chapter 5.2.5.

6.2.7 Calculation of the genomic representation of each chromosome in the reference human genome

The expected %GR in the reference human genome was calculated according to the description in Chapter 5.2.6.

6.2.8 Calculation of fetal DNA fraction from %chrY

The fetal DNA fraction was calculated from %chrY according to the description in Chapter 5.2.7.

6.2.9 Computer simulation

Computer simulation is carried out to evaluate the effect of the multiplexing level of maternal plasma DNA sequencing on the genomic representation of the target chromosome, i.e., chromosome 21. An in-house algorithm was kindly compiled by Prof Hao Sun and Mr Peiyong Jiang. In a sequence dataset generated by the monoplex protocol, usually there would be 7~10 million raw sequence reads (indicating how many DNA molecules had been sequenced). The algorithm would first count the exact number of raw sequence reads for each case, and would then calculate its 1/2, 1/4, 1/ 6, 1/8, 1/10 and 1/12 to represent the expected number of sequence reads generated by the 2-plex, 4-plex, 6-plex, 8-plex, 10-plex and 12-plex protocols, respectively. A subset containing the resultant number for each multiplexing level was randomly sampled from the monoplex sequence data, and such random sampling process was repeated for 1000 times. In each subset, the U0-1-0-0 sequence reads were identified and selected, and the genomic representation of chr21 U0-1-0-0 reads (%chr21) was then calculated as described above.

6.3 Results

6.3.1 Effect of plasma volume on massively parallel maternal DNA sequencing

6.3.1.1 Effect of reduction in plasma volume on plasma DNA sequence profile

To determine whether the reduction in plasma volume would distort the overall genomic representation of maternal plasma DNA, ~20 mL blood was collected from three pregnant women carrying euploid male fetuses (1 in the 1st trimester and 2 in the 2nd trimester). About 10 mL of maternal plasma was harvested from each case and subsequently divided into three portions (1.2 mL, 2.4 mL and 4.8 mL, respectively). The plasma DNA from each portion was extracted separately. The amount of plasma DNA quantified by the *HBB* QPCR showed less than 1 ng for the 1st trimester case (P003) with 1.2 mL of plasma extracted. If starting with such low DNA amount, one might worry about the potential failure of the library preparation, taking the multiple steps of column purification into account. Hence, I compared the quality and quantity of DNA libraries constructed from the three portions of the same case before the sequencing steps. The bioanalyzer results (Figure 6.1) and the concentrations measured by the SYBR Green quantitative PCR assay (Table 6.1) demonstrated that there was no deficiency for the cases with low plasma volume or small plasma DNA amount (Wilcoxon Signed Rank Test, $P = 0.750$ between 4.8 mL and 1.2 mL plasma as starting material), indicating that DNA library could be successfully constructed with the limited plasma input.

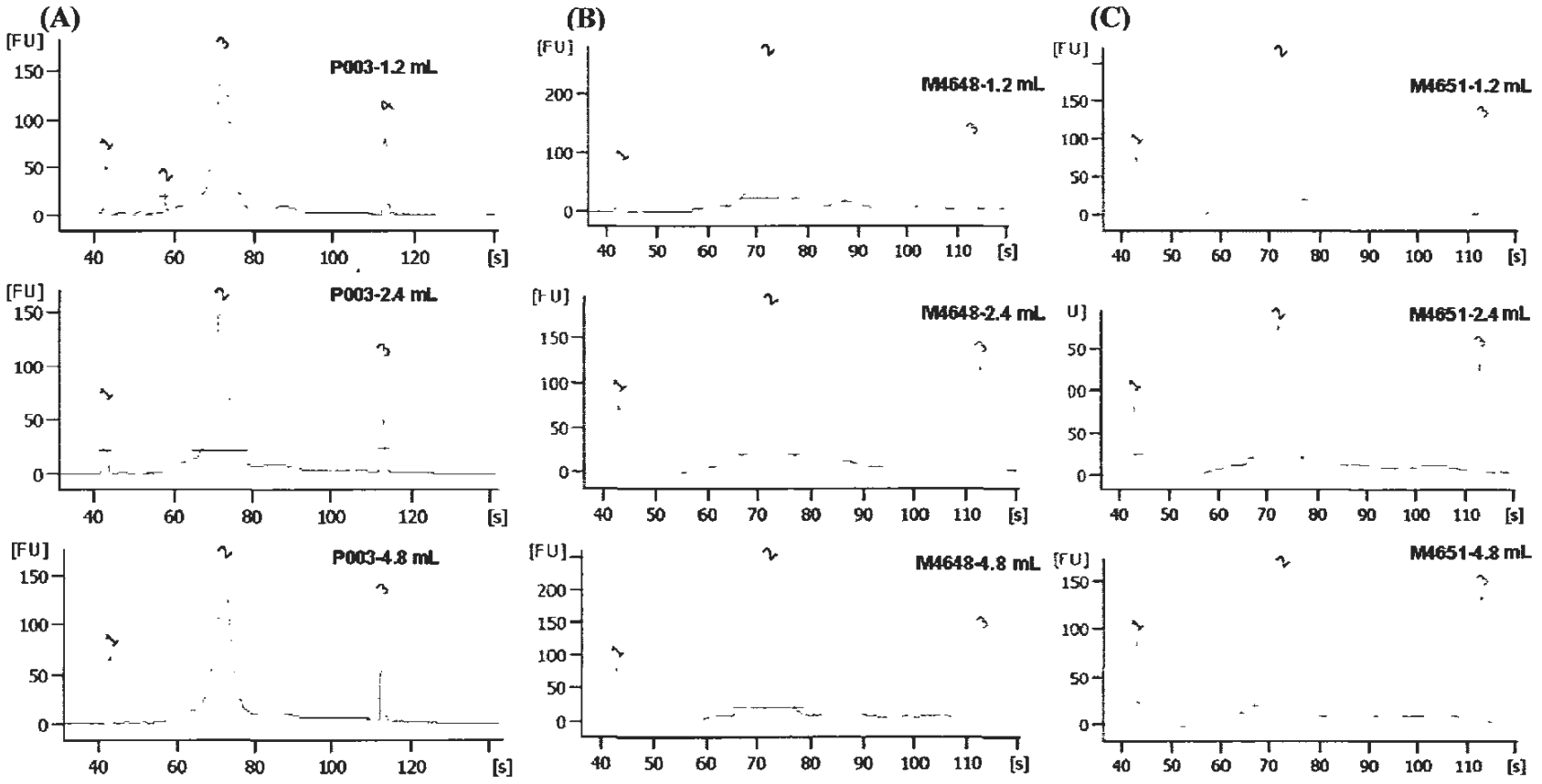


Figure 6.1 Bioanalyzer results of DNA libraries constructed from plasma DNA extracted from various plasma volumes.

(A) Case P003 collected from a male pregnancy in the 1st trimerster; *(B)* Case M4948 collected from a male pregnancy in the 2nd trimerster; *(C)* Case M4951 collected from a male pregnancy in the 2nd trimerster. In the electropherogram from the Agilent Bioanalyzer, the x axis represents the runtime (s), and the y axis represents the fluorescence units (FU).

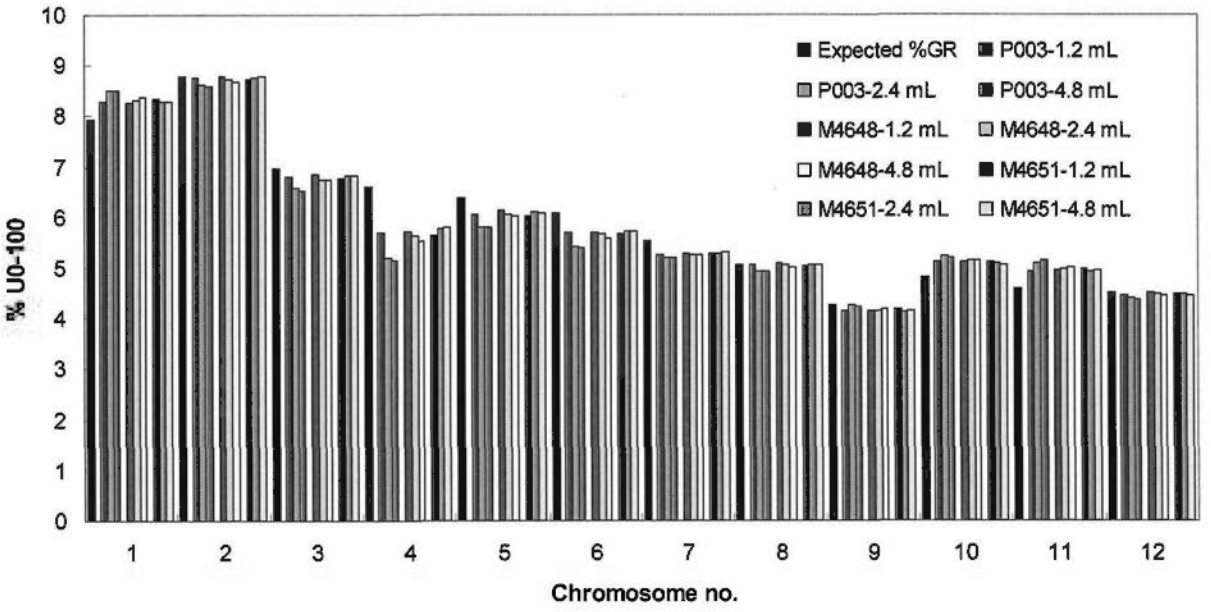
Case no.	Fetal sex	GA (weeks + days)	Extract volume (mL)	Input DNA (ng)	Karyotype	Concentration of library (nM)[a]	Total sequenced counts	Total U0-1-0-0 counts
P003	M	12+2	1.2	0.92	46XY	4,900	6,758,778	1,852,171
P003	M	12+2	2.4	5.82	46XY	5,419	6,776,667	2,071,178
P003	M	12+2	4.8	18.41	46XY	5,523	7,224,780	2,261,332
M4648	M	21+3	1.2	6.4	46XY	8,617	8,694,325	2,362,194
M4648	M	21+3	2.4	11.8	46XY	7,134	8,712,930	2,485,334
M4648	M	21+3	4.8	20.0	46XY	7,350	8,157,785	2,561,712
M4651	M	17+5	1.2	3.5	46XY	8,077	8,589,307	2,621,028
M4651	M	17+5	2.4	8.1	46XY	8,160	8,605,502	2,624,705
M4651	M	17+5	4.8	20.0	46XY	7,542	8,879,256	2,618,743
M4649	M	18+5	1.2	4.7	46XY	7,508	8,066,793	2,497,132
K28	M	13	1.1	2.61	46XY	4,843	7,744,243	2,266,894
K48	M	13	1	3.95	46XY	5,025	7,055,908	2,169,397
K31	M	13	1	2.86	47XY+21	5,087	7,068,780	2,124,639
K334	M	13	1	2.79	47XY+21	5,356	5,782,333	1,715,115

a: the mole numbers are estimated by SYBR Green quantitative PCR.

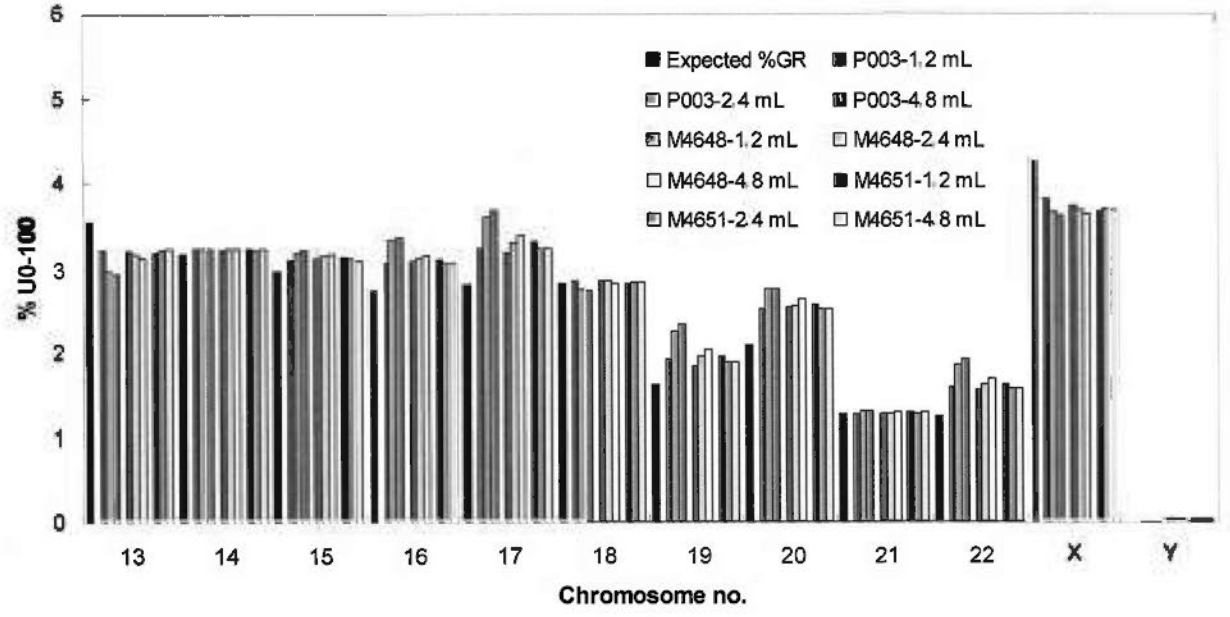
Table 6.1 Clinical details and sequencing counts of maternal plasma samples for plasma volume evaluation.

I next assessed whether the DNA extracted from various plasma volumes could closely resemble the expected %GR of each chromosome. A median of 8.6×10^6 raw reads were obtained from each sample, ~30% of which were U0-1-0-0 reads (the upper panel of Table 6.1). As shown in Figure 6.2A and 6.2B, the %U0-1-0-0 of each chromosome for each portion was generally close to the expected value. By plotting the obtained %GR against the expected %GR for all autosomes, the high concordance between the two variables was found regardless of the variations in starting volumes (Figure 6.2C, D and E). Since the majority of DNA molecules in maternal plasma are of maternal origin (Lo *et al.* 1998a), the robust detection and the consistent quantitative measurement of the fetal-specific DNA molecules in maternal plasma would be determinative factors in the evaluation of maternal plasma DNA sequencing with reduced plasma input. As all three cases were collected from pregnant women carrying male fetuses, I used the sequence reads from chrY as a fetal-specific marker to assess the variability in fetal DNA detection as plasma input decreased. The statistical comparison revealed that there was no significant difference in %chrY as the plasma volume decreased (Wilcoxon Signed Rank Test, $P = 0.750$ between 4.8 mL and 1.2 mL plasma as starting material). Taken together, it would be feasible to start the massively parallel sequencing of maternal plasma DNA from a relatively low plasma volume, i.e., as low as 1.2 mL, as evidenced by the undistorted genomic representation and the effective detection of the fetal-specific signals.

(A)


Expected %GR
P003-1.2 mL
P003-2.4 mL
P003-4.8 mL
M4648-1.2 mL
M4648-2.4 mL
M4648-4.8 mL
M4651-1.2 mL
M4651-2.4 mL
M4651-4.8 mL
% U0-100
Chromosome no.

(B)


Expected %GR
P003-1.2 mL
P003-2.4 mL
P003-4.8 mL
M4648-1.2 mL
M4648-2.4 mL
M4648-4.8 mL
M4651-1.2 mL
M4651-2.4 mL
M4651-4.8 mL
% U0-100
Chromosome no.

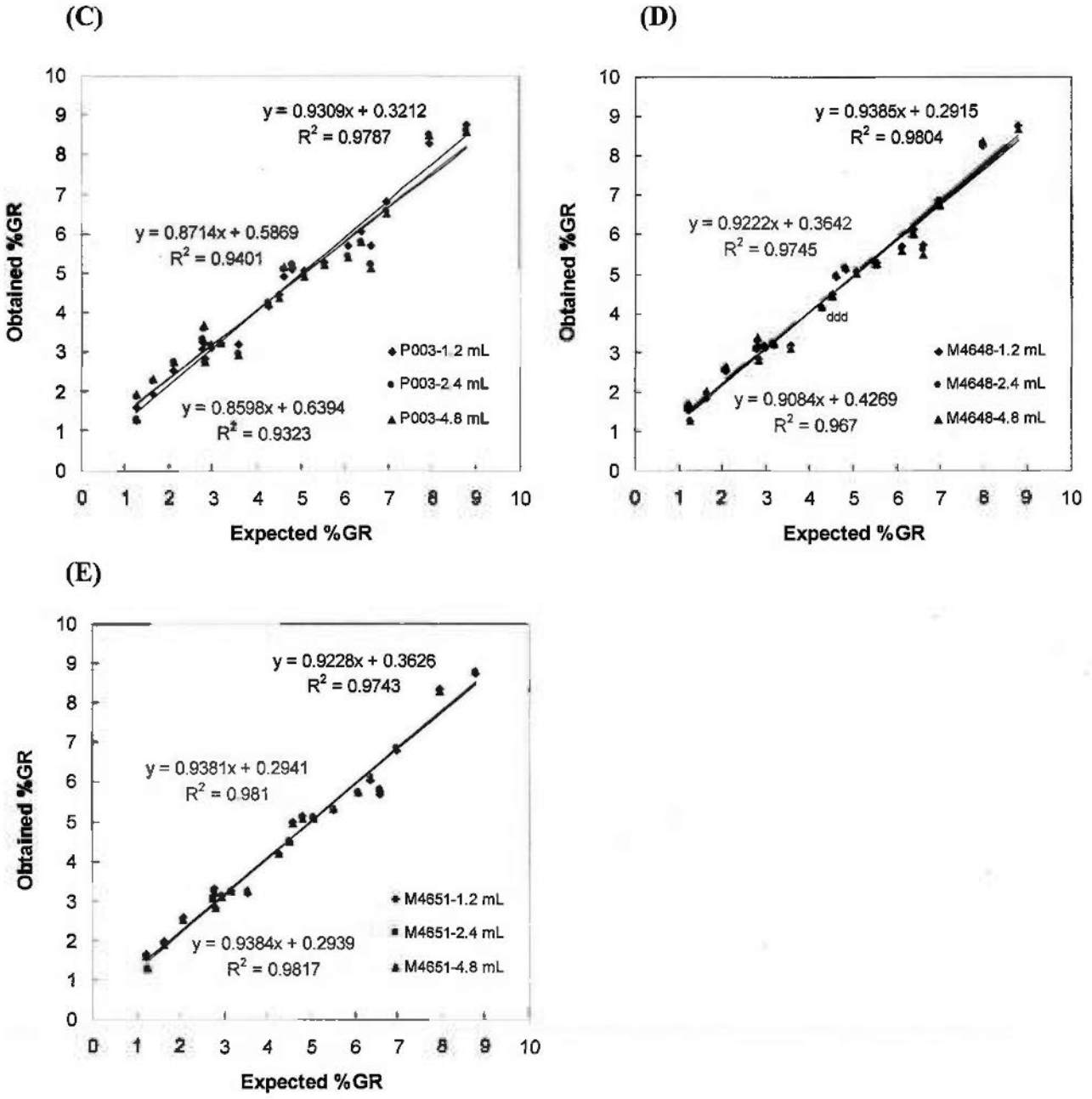


Figure 6.2 Genomic representation of sequenced plasma DNA from various volumes.

(A) and ***(B)*** Bar chart of %U0–1–0–0 sequences for chromosome 1-12 and for chromosome 13-22 along with chrX and chrY, respectively, for each of the three portions per case. ***(C)***, ***(D)*** and ***(E)*** Linear regression plots of the obtained %GRs (%U0–1–0–0) against the expected %GRs for all autosomes in the case P003, M4648 and M4651, respectively. The blank, pink and blue dots represent the data from the plasma DNA samples extracted from 1.2 mL, 2.4 mL and 4.8 mL of plasma, respectively.

6.3.1.2 Fetal trisomy 21 detection with ≤ 1.2 mL maternal plasma

To further evaluate the accuracy of fetal trisomy 21 diagnosis by massively parallel sequencing of maternal plasma DNA extracted from a relatively low plasma volume, 1~1.2 mL of plasma was additionally collected from 3 women each carrying a euploid fetus and 2 women each carrying a T21 fetus. The clinical details and sequencing count of these maternal plasma samples are shown in the lower panel of Table 6.1. The median number of sequence reads generated per sample was 7.1×10^6 and the median U0-1-0-0 count was 2.2×10^6. The percentage of U0-1-0-0 sequences to each chromosome was plotted against the expected %GR for all euploid cases starting from ≤ 1.2 mL of maternal plasma (Figure 6.3A), showing the similarity to the expected value for each chromosome. The %chr21 values in the two T21 pregnancies were higher than those in the euploid pregnancies (Figure 6.3B). Using all euploid cases with ≤ 1.2 mL of plasma to constitute a reference group, the %chr21 values were transformed into chr21 *z*-scores. Both T21 fetuses had a chr21 *z*-score of > 3 (5.95 and 8.16, respectively) (Figure 6.3C), achieving the successful detection of fetal trisomy 21.

Similar to the observations depicted in Chapter 5, a nonuniform distribution of sequence reads among chromosomes was observed (Figure 6.2A and 6.2B, and Figure 6.3A). The obtained %GRs deviated from the expected values to different extents. By plotting the measured ratios of the observed %GR to the expected %GR against the chromosomal GC content, a positive correlation between the two variables was once again found (Pearson Product Moment Correlation, $r = 0.943$, $P = 5.166 \times 10^{-11}$, Figure 6.4A). However, the correlation coefficients for two datasets (in Chapter 5 and this chapter) were different. Since these two datasets were generated separately in two labs using different versions of the Illumina sequencing platforms

(i.e., GA I and GA II, respectively), there were probably inter-batch and/or inter-equipment variations in measuring %GR of each chromosome. In addition, the CVs were plotted to show the measurement precision for the autosomes and chrX. The CV plot showed the same pattern as that in Chapter 5, tending to be larger at the either end of the GC-abundance spectrum (Figure 6.4B). The internal GC bias of the Illumina sequencing platform and its influence on the imprecise measurement of certain chromosomes seemed to persist regardless of batches or equipment versions.

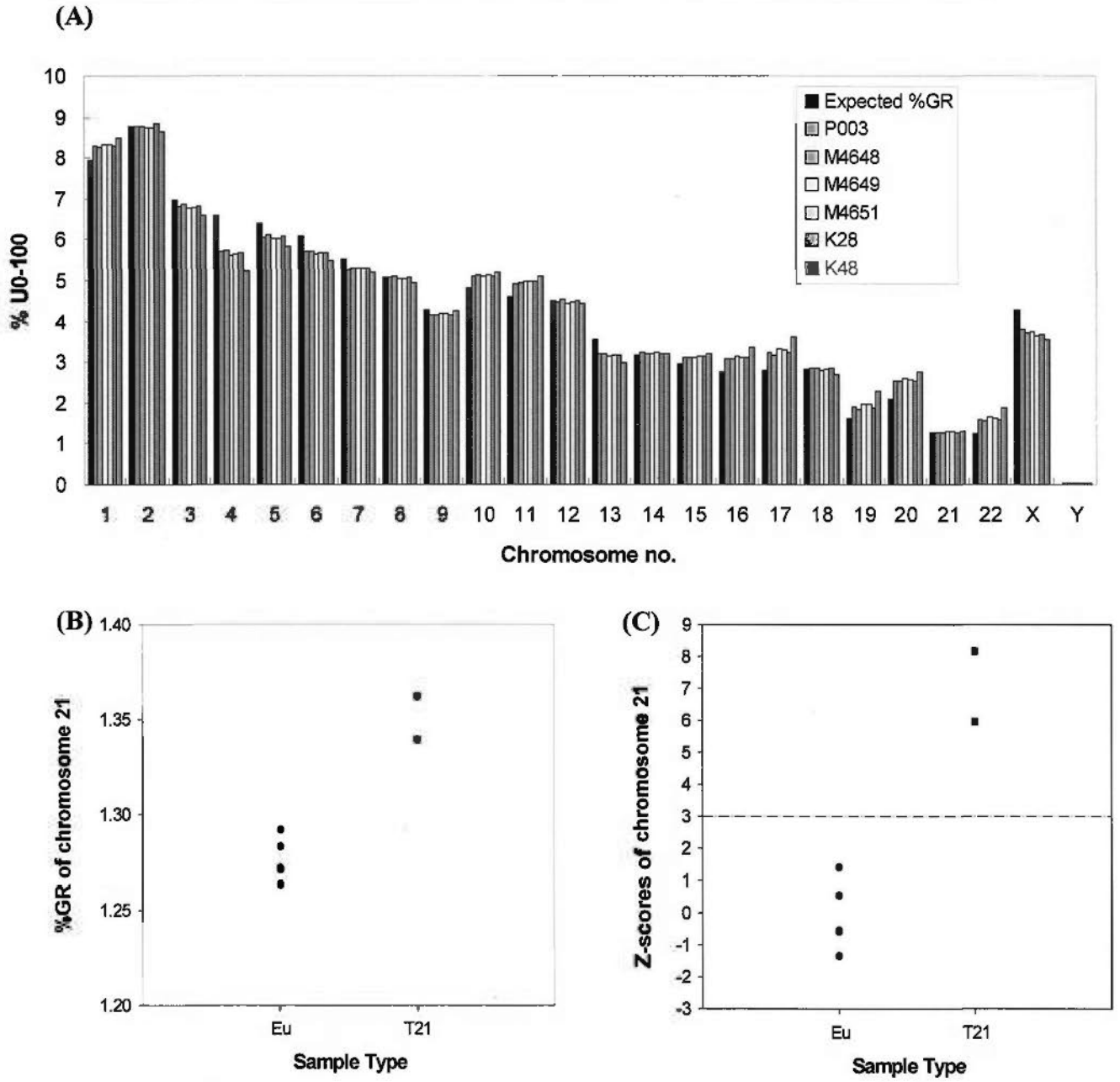


Figure 6.3 Genomic representation and the prenatal detection of fetal trisomy 21 with ≤ 1.2 mL maternal plasma.

(A) Bar chart of %U0-1-0-0 sequences per chromosome for the 6 euploid male cases starting from ≤ 1.2 mL of maternal plasma. The percentage of genomic representation of each chromosome as expected for a repeat-masked reference haploid female genome is plotted for comparison (black bars). *(B)* %GR of chr21 for the euploid and T21 pregnancies; *(C)* *Z*-scores of chr21 for the euploid and T21 cases. The dashed line represents a *z*-score of 3 as a diagnostic cutoff.

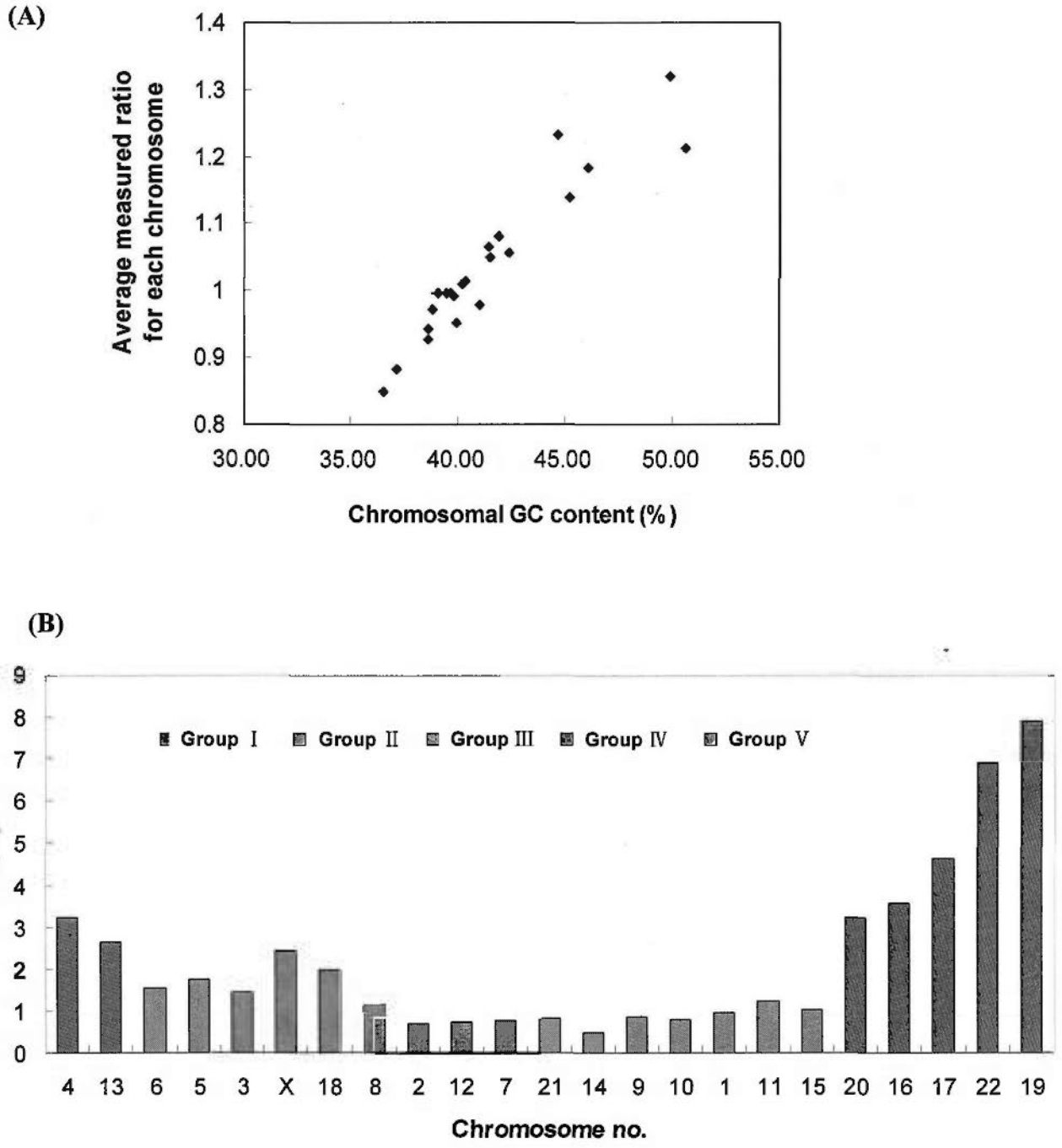


Figure 6.4 Sequencing bias and variation among chromosomes with ≤ 1.2 mL maternal plasma

(A) Correlation between the measured ratios (of the obtained %GR to the expected %GR) and chromosomal GC content for autosomes. The values on the y axis were calculated from 6 euploid male cases with ≤1.2 mL of maternal plasma extracted. ***(B)*** CVs for autosmes plus chrX. Chromosomes are grouped according to their GC contents. Each group is represented by one color. The x axis is ordered with increasing GC contents.

6.3.2 Multiplexed sequencing of maternal plasma DNA for fetal trisomy 21 detection

6.3.2.1 Multiplexed massively parallel sequencing of maternal plasma DNA

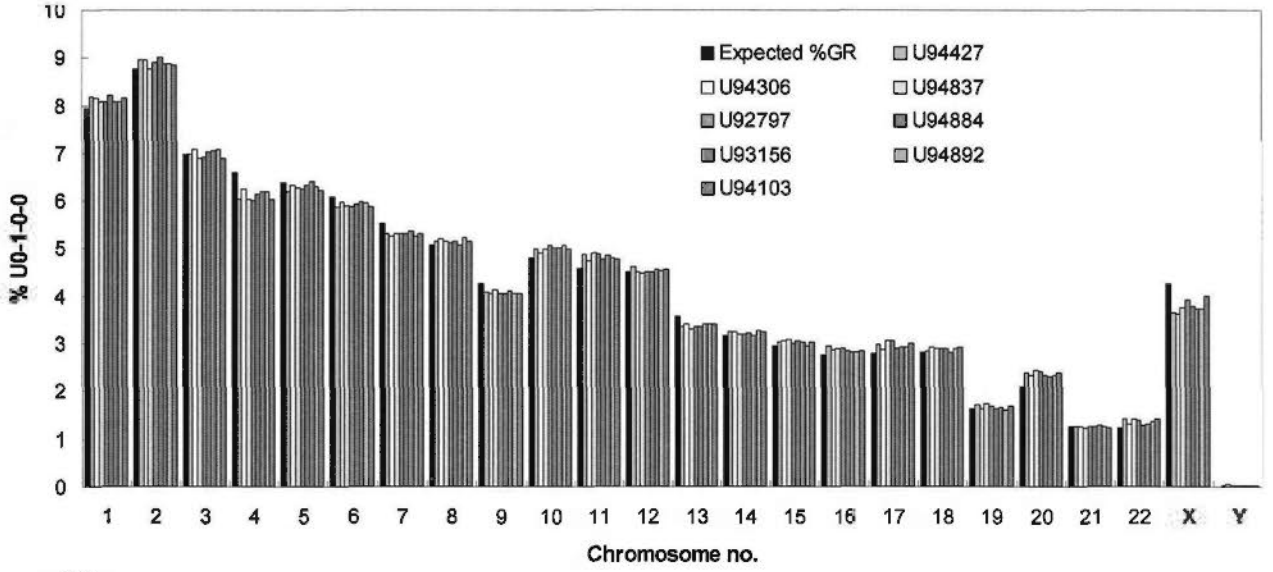
With the introduction of multiplexing, more than one sample could be simultaneously sequenced in a specimen lane of an Illumina flow cell. To evaluate the diagnostic performance of multiplexed massively parallel sequencing of maternal plasma DNA for fetal trisomy 21 detection, 52 maternal plasma samples from 45 women carrying karyotypically normal fetuses and 7 women carrying T21 fetuses in first and second trimesters were collected and then subjected to sequencing by the 8-plex and 4-plex protocols, respectively. The sample information and sequence reads for 52 maternal plasma samples are summarized in Appendix II. A mean of 0.85 million (SD, 136,791) and 1.72 million (SD, 403,607) raw sequence reads were obtained per maternal plasma samples sequenced by the 8-plex and 4-plex protocols, respectively. After sequence alignment, a mean of 0.26 million (SD, 40,961) and 0.56 million (SD, 131,946) unique perfectly matched reads were retrieved, respectively. In these scenarios, the obtained 36-bp U0-1-0-0 reads would be equivalent to 0.3% and 0.6% of the human genome by the 8-plex and 4-plex protocols, respectively.

6.3.2.2 Measurement precision

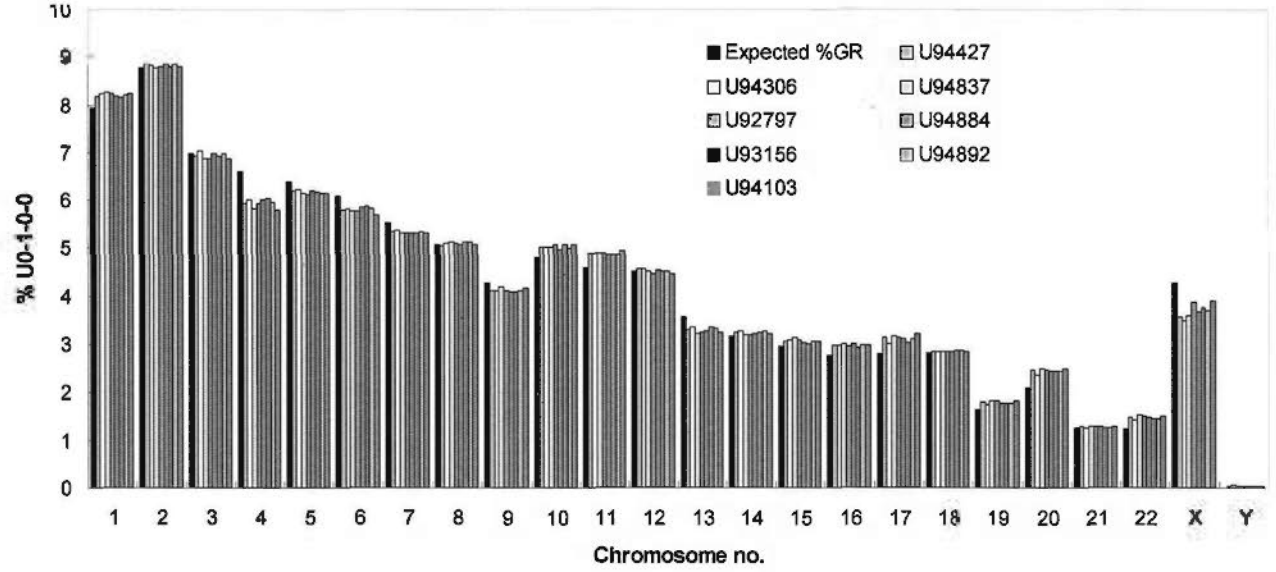
Because multiplexed sequencing is another sequencing mode, I first checked the sequence distribution profile of plasma DNA for each chromosome. Similarly, the quantitative representation of each chromosome was not uniform for both protocols (Figure 6.5A and 6.5B). For T21 diagnosis, 8 euploid pregnancies of male fetuses were selected at random to constitute the reference group for both 8-plex and 4-plex

analysis. On the whole, irrespective of the multiplexing levels, the measurements of %GR for chromosomes with a low or high GC content were less precise than those with an intermediate GC content by maternal plasma DNA sequencing analysis. Notably, it was observed that the CVs among chromosomes from the 8-plex data were generally larger than the 4-plex data, suggesting a worse precision profile for the 8-plex when compared with the 4-plex protocol (Figure 6.5C). In particular, the CVs for measuring the %GR of chromosome 21 from the 8-plex and 4-plex data were 1.53 % and 1.12%, respectively.

(A)


Expected %GR
U94427
U94306
U94837
U92797
U94884
U93156
U94892
U94103
% U0-1-0-0
Chromosome no.

(B)


Expected %GR
U94427
U94306
U94837
U92797
U94884
U93156
U94892
U94103
% U0-1-0-0
Chromosome no.

(C)

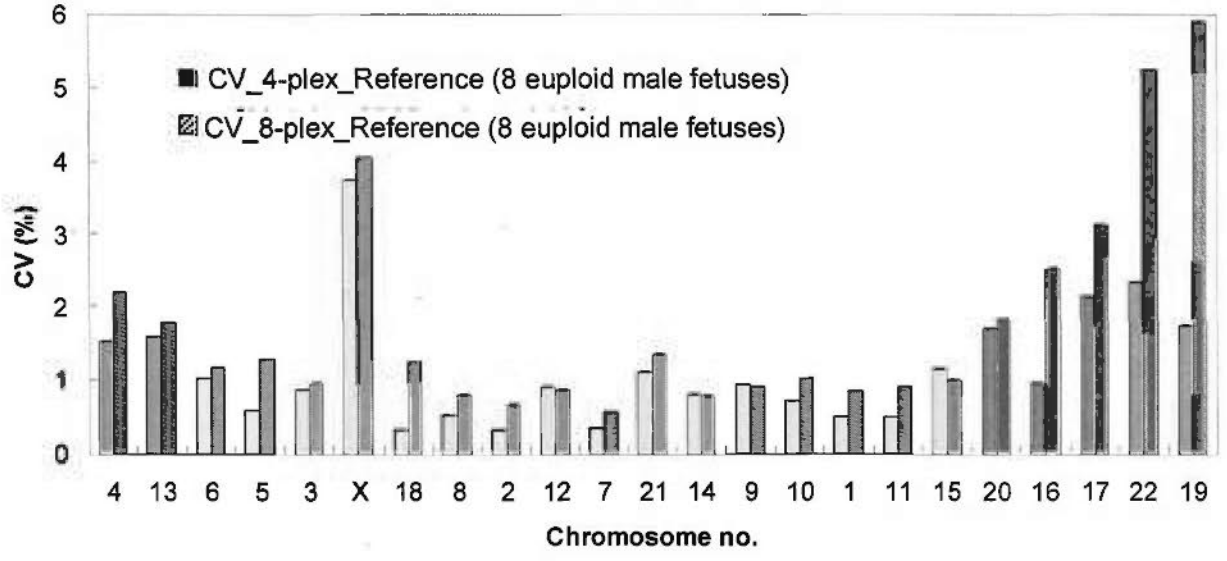

CV_4-plex_Reference (8 euploid male fetuses)
CV_8-plex_Reference (8 euploid male fetuses)
CV (%)
4 13 6 5 3 X 18 8 2 12 7 21 14 9 10 1 11 15 20 16 17 22 19
Chromosome no.

Figure 6.5 Sequencing bias and variation among chromosomes by the 8-plex and 4-plex protocols.

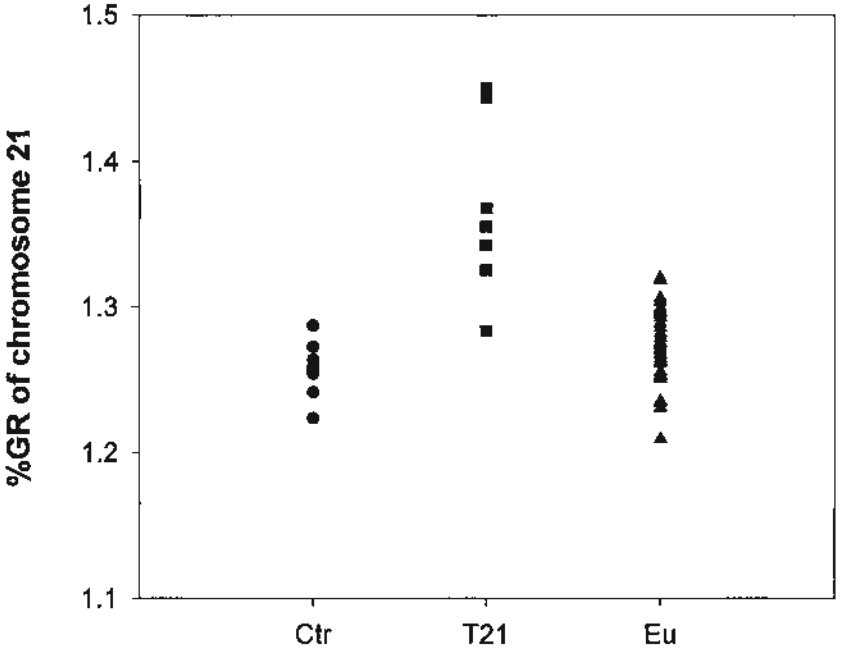
(A) Bar chart of %U0-1-0-0 sequences per chromosome for 8 reference samples (maternal plasma samples from euploid male pregnancies) by the 8-plex protocol. ***(B)*** Bar chart of %U0-1-0-0 sequences per chromosome for the same 8 reference samples (maternal plasma samples from euploid male pregnancies) by the 4-plex protocol. The percentage of genomic representation of each chromosome as expected for a repeat-masked reference haploid female genome is plotted for comparison (black bars). ***(C)*** CVs for the autosmes plus chrX for the 8-plex and 4-plex protocols. Chromosomes are grouped according to their GC contents. Each group is represented by one color. The x axis is ordered with increasing GC contents.

6.3.2.3 Trisomy 21 detection

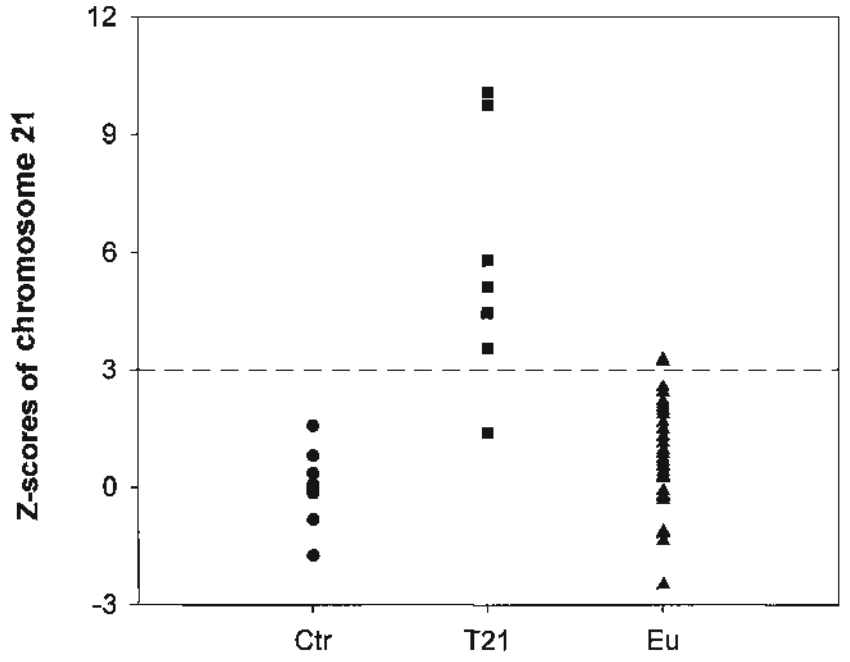
The %chr21 values and chr21 *z*-scores of the sequenced samples are shown in Figure 6.6. Here, I use our predefined diagnostic cutoff, namely, a *z*-score value of 3, for fetal trisomy 21 detection. For the 8-plex protocol, all but one T21 case showed chr21 *z*-scores larger than the diagnostic cutoff of 3, whereas all but one euploid case showed chr21 *z*-scores less than 3 (Figure 6.6A and 6.6B). These results translate to a sensitivity of 85.7% and a specificity of 97.8%, respectively. For the 4-plex protocol, all euploid cases had *z*-scores < 3, whereas all T21 cases had *z*-scores > 3, showing a 100% accuracy for the prenatal diagnosis of fetal trisomy 21 (Figure 6.7C and 6.7D).

As mentioned above, the degree in the overrepresentation of chr21 for a T21 pregnancy is dependent on the fetal DNA concentration. The fetal DNA concentrations of male pregnancies could be estimated by %chrY values. Although the case number involved in the current cohort was limited (three of seven T21 pregnancies were carrying male fetuses), there was a tendency towards larger chr21 *z*-scores if there were higher fetal DNA concentrations in the plasma samples from T21 pregnancies (Figure 6.7E).

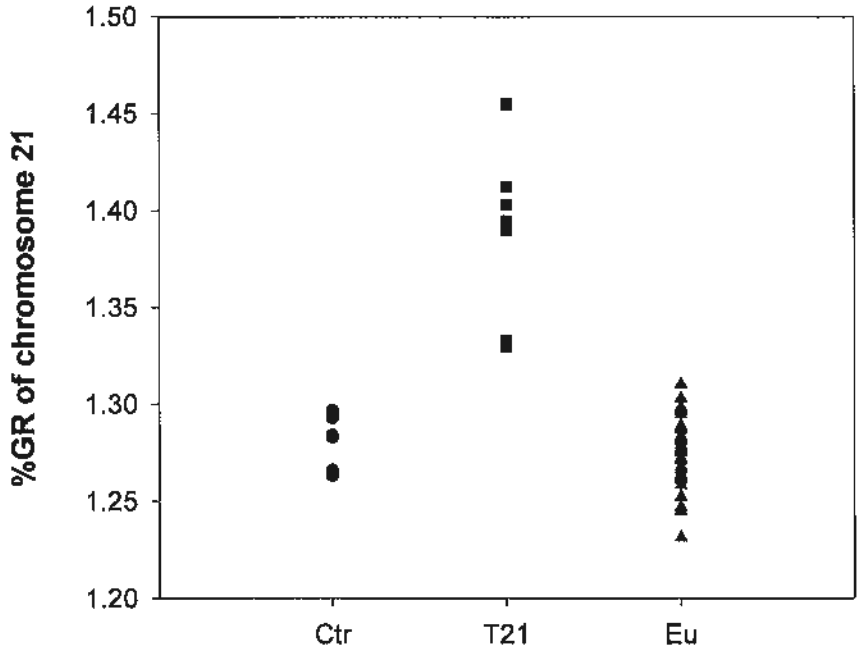
(A)


%GR of chromosome 21
1.5
1.4
1.3
1.2
1.1
Ctr
T21
Eu

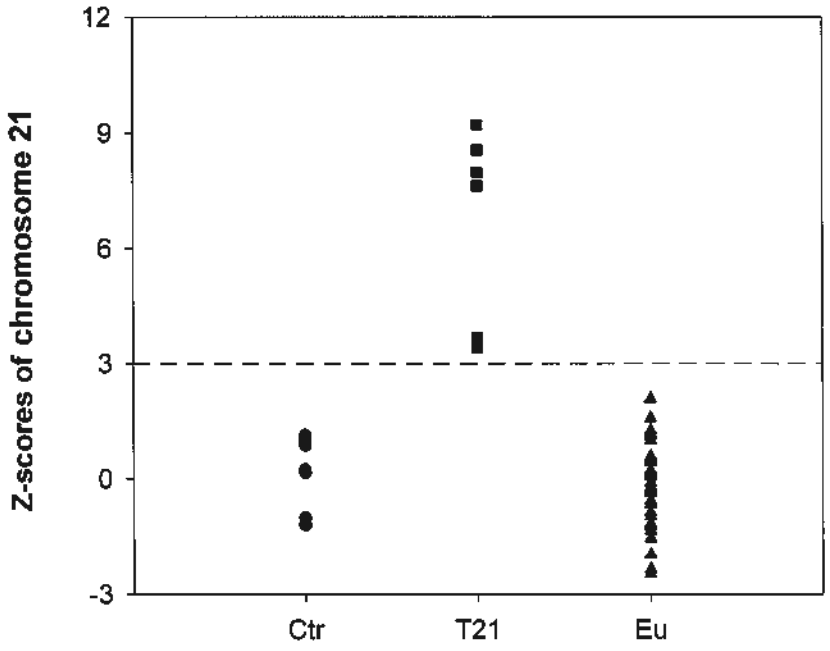
(B)


Z-scores of chromosome 21
12
9
6
3
0
-3
Ctr
T21
Eu

(C)


1.50
1.45
1.40
1.35
1.30
1.25
1.20
%GR of chromosome 21
Ctr
T21
Eu

(D)


12
9
6
3
0
-3
Z-scores of chromosome 21
Ctr
T21
Eu

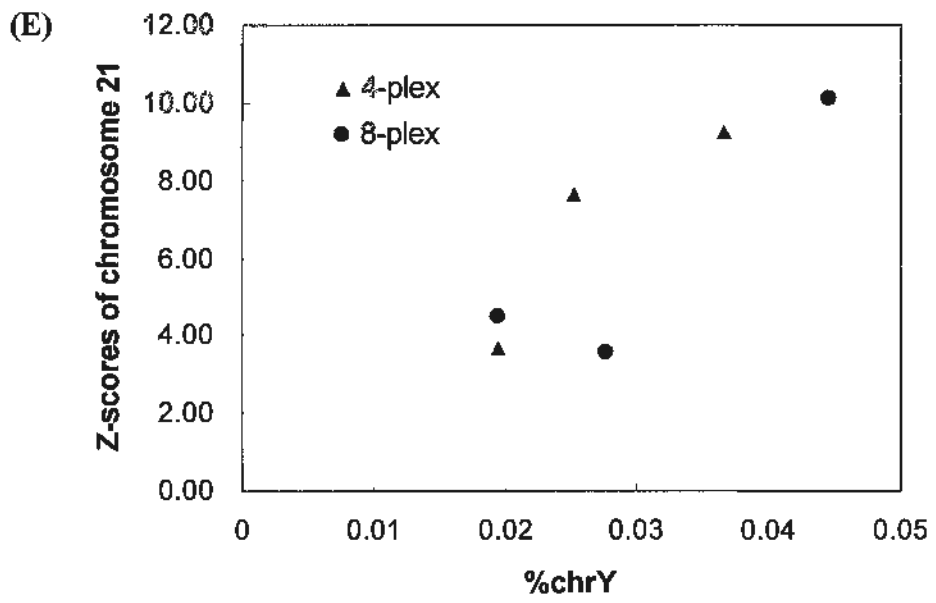


Figure 6.6 Fetal trisomy 21 detection by multiplexed sequencing of maternal plasma DNA.

(A) %GR of chr21 by the 8-plex protocol. ***(B)*** Chr21 *z*-scores by the 8-plex protocol. ***(C)*** %GR of chr21 by the 4-plex protocol. ***(D)*** Chr21 *z*-scores by the 4-plex protocol. The x axis indicates the three sample types, i.e., euploid controls (Ctr), T21 cases and euploid cases (Eu). ***(E)*** Correlation between the fetal DNA concentration (represented by %chrY) and the chr21 *z*-score for T21 male pregnancies. The dashed line represents the *z*-score of 3 for T21 diagnosis.

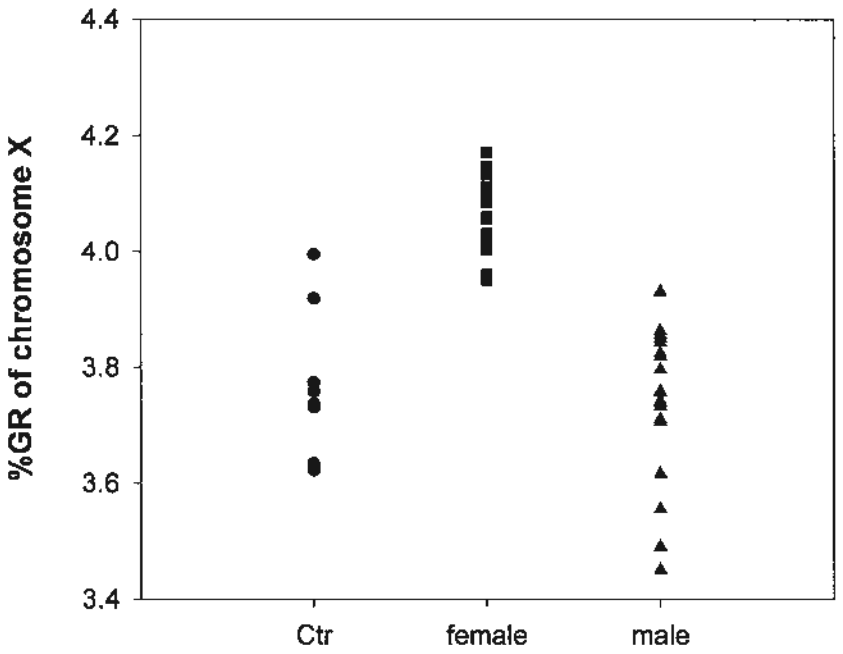
6.3.2.4 Fetal sex determination

Fetal sex could be determined by %chrX and %chrY values from maternal plasma DNA sequencing. In theory, the %chrX values of female pregnancies would be larger than those of male pregnancies owing to two doses of chrX contributed by the female fetus while the %chrY values of female pregnancies should be 0% due to the absence of chrY in the female fetus. The former one could be transformed into chrX *z*-scores by using the euploid male fetuses as controls. For the latter one, our group previously reported that a small fraction of reads would be misaligned to chrY in pregnancies with female fetuses (Chiu *et al.*, 2008a), therefore, an optimal cutoff of %chrY should be defined to determine the fetal sex.

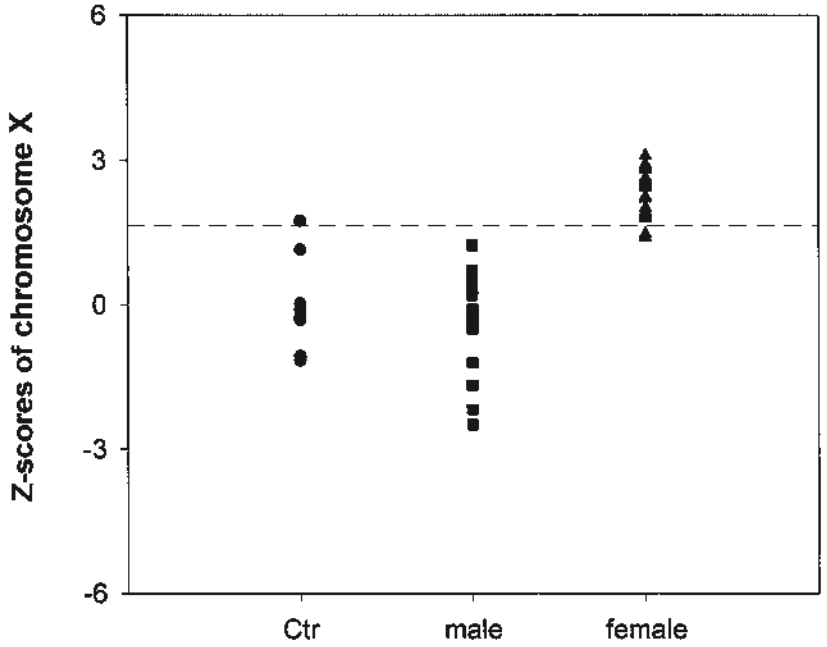
Figure 6.7 shows the diagnostic results of fetal sex determination with chrX *z*-scores and %chrY values. The %chrX value of male pregnancies is dependent on the fetal DNA concentration. With higher fetal DNA concentrations in the maternal plasma from male pregnancies, the %chrX values become smaller. As the eight reference samples varied in the fetal DNA concentration (%chrY range for the 8-plex protocol, 0.018%~0.043%, %chrY for the 4-plex protocol, 0.017%~0.045%), their %chrX values fluctuated accordingly, probably increasing the SD of %chrX in the reference group and resulting in a larger CV for the %chrX measurement in Figure 6.6. Hence, to sensitively detect the chrX overrepresentation in female pregnancies, a loose diagnostic cutoff, namely, a *z*-score of 1.65 (denoting a %chrX value greater than that of the 95th percentile of the reference set for a one-tailed distribution), was used. For the 8-plex and 4-plex protocols, 22 and 20 of 24 female fetuses could be successfully identified, thus showing detection rates of 91.7% and 83.3%, respectively (Figure 6.7C and 6.7D). On the other hand, there was a clear separation of %chrY values between female and male pregnancies, thus showing a great

discrimination power between them. Using 0.010% and 0.012% (the optimal cutoff values determined with the use of receiver operating characteristic (ROC) curve analysis) as %chrY cutoffs for the 8-plex and 4-plex protocols, respectively, all female and male fetuses could be correctly identified (Figure 6.7E and F). In addition, the case U93243 showed false positive signals on the *SRY/HBB* assay (Appendix II), but was proven to be a female fetus in terms of %chrY.

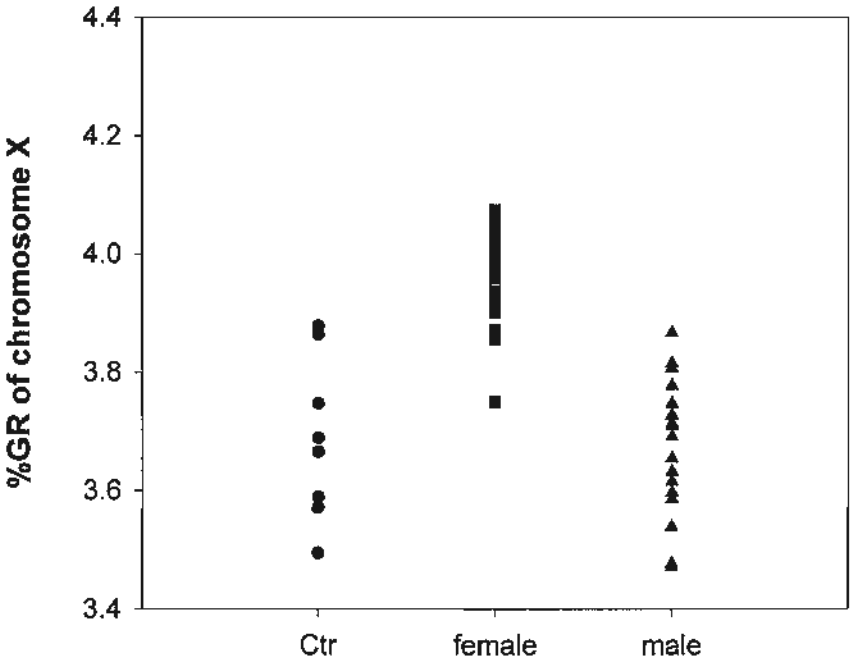
(A)


%GR of chromosome X
4.4
4.2
4.0
3.8
3.6
3.4
Ctr
female
male

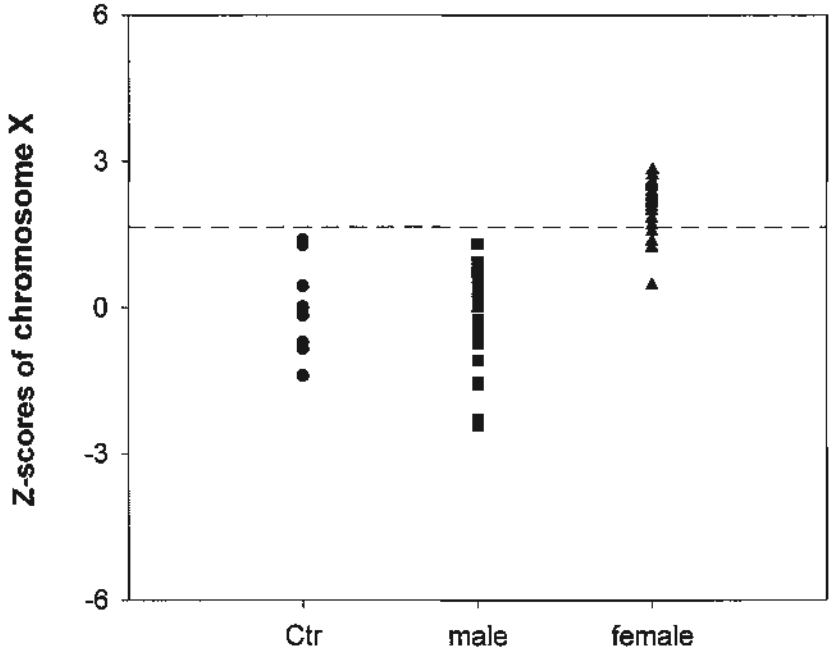
(B)


Z-scores of chromosome X
6
3
0
-3
-6
Ctr
male
female

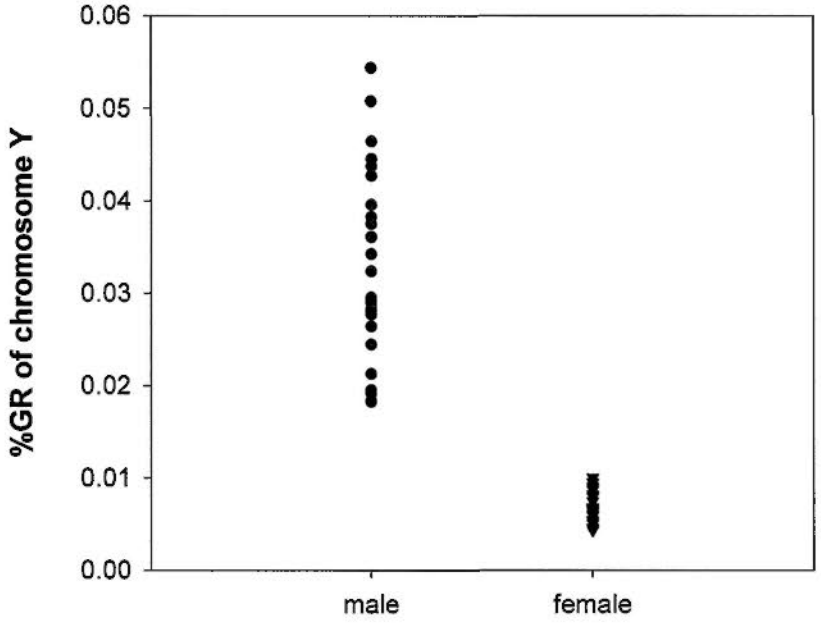
(C)


4.4
4.2
4.0
3.8
3.6
3.4
%GR of chromosome X
Ctr
female
male

(D)


6
3
0
-3
-6
Z-scores of chromosome X
Ctr
male
female

(E)



(F)

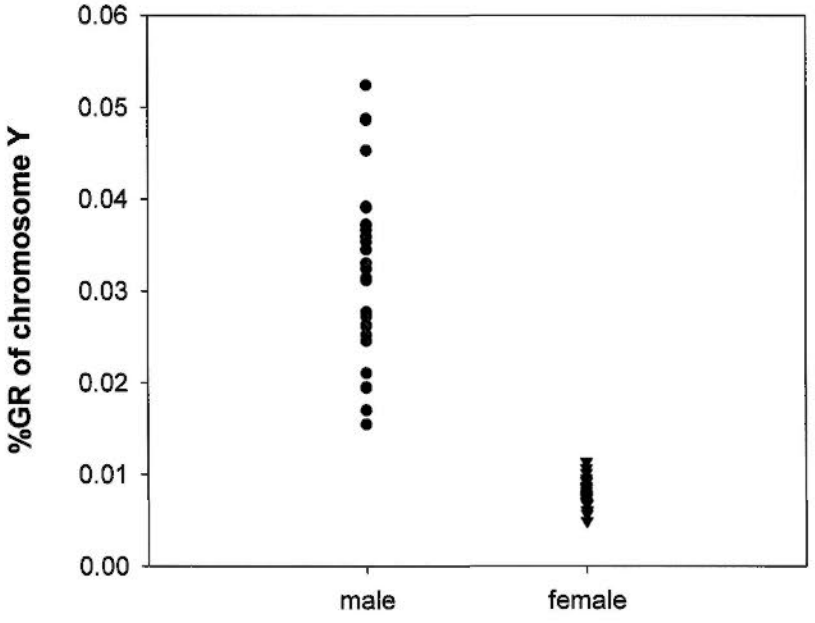


Figure 6.7 Fetal sex determination by multiplexed sequencing of maternal plasma DNA.

(A) %GR of chrX by the 8-plex protocol. ***(B)*** ChrX *z*-scores by the 8-plex protocol. ***(C)*** %GR of chrX by the 4-plex protocol. ***(D)*** ChrX *z*-scores by the 4-plex protocol. ***(E)*** %GR of chrY (%chrY) by the 8-plex protocol. ***(F)*** %GR of chrY by the 4-plex protocol. The dashed line represents the *z*-score of 1.65 as a diagnostic cutoff.

6.3.2.5 Analysis of fetal DNA%

From the sequencing data, the %chrY value could be used to estimate the fetal DNA fraction in maternal plasma. Before library construction, all maternal samples were subjected to the *SRY/HBB* assay, which has been one of the standard assays used in the field (Lo *et al.* 1998b). After sequencing, the fetal DNA fractions in the plasma from the 28 male pregnancies were determined based on the %chrY values obtained from the 4-plex protocol. I compared the fetal DNA fractions calculated by %chrY values with those measured by the *SRY/HBB* assay for all 28 male pregnancies (inclusive of euploid and T21 pregnancies). The median fractional concentration of fetal DNA estimated by %chrY values was 17.14%, ranging from 5.62% to 30.27%, while the median fractional concentration measured by the *SRY/HBB* assay was 5.8%, ranging from 0.3% to 15.0%. The fetal DNA fractions estimated in both ways were correlated (Pearson Product Moment Correlation, $r = 0.395$, $P = 0.0376$) but the sequencing approach resulted in significantly higher fetal DNA% than the *SRY/HBB* assay (Wilcoxon Signed Rank Test, $P < 0.001$). The median fractional concentration of fetal DNA in maternal plasma estimated by %chrY was found ~2.5 times higher than those estimated by the *SRY/HBB* assay for the male pregnancies, in agreement with the data reported by Lun *et al.,* who used a microfluidics digital PCR platform to reveal that the fractional concentration of fetal DNA in maternal plasma was $\geqq 2$ times higher than previously reported using QPCR (Lun *et al.* 2008a). Furthermore, the fetal DNA concentrations in the three male T21 pregnancies were 8.33%, 12.15% and 19.77%, respectively. There was no significant difference in fetal DNA concentration between the euploid and T21 pregnancies in the current sample set (Mann-Whitney Rank Sum Test, $P = 0.373$), but no conclusion could be drawn at this point due to the small sample size in T21 group.

6.3.2.6 Effect of multiplexing level on quantitative representation

The %chr21 values for the false positive case (U93619) and the false negative case (U94376) by the 8-plex protocol were 1.319% and 1.284%, respectively, which later turned to be 1.261 and 1.329, respectively, by the 4-plex protocol. It was unlikely that these two samples were swapped up during the 8-plex sequencing in view of their distinct sequence tags (index). Besides, both raw read number and U0-1-0-0 read number for the two cases by the 8-plex protocol were shown within the normal range (Appendix II). For further validation, their DNA libraries were additionally sequenced by the monoplex protocol (i.e., one sample per lane). In total, 6.81 million and 8.54 million raw sequence reads were obtained for U93619 and U94376, respectively. Of these, 2.09 million and 2.72 million were retrieved as U0-1-0-0 reads, respectively. The %chr21 values for U93619 and U94376 by the monoplex protocol were 1.264 and 1.322, respectively, which were close to those from the 4-plex data but different from those from the 8-plex data. These results suggested that when the number of sequenced molecules decreased, the quantitative representation of chr21 would increasingly stray from the actual value, increasing the chance of false diagnoses.

To explore the effect of the multiplexing level on the measurement of %chr21, a computer simulation was conducted. The simulation was based on the sequence data from U93619 by the monoplex protocol, which showed a %chr21 value of 1.264%. As predicted, the results from the simulated multiplexed sequencing demonstrated that the variability in %chr21 values expanded gradually as the multiplexing level increased (Figure 6.8). For the results from the 8-plex simulation, there was 8.6% chance for the %chr21 value being larger than 1.296% (the expected %chr21 value for a maternal plasma sample obtained from a trisomy 21 pregnancy containing 5% fetal DNA was defined as an arbitrary criterion); whereas such probabilities were

2.2% and 0.2% for the results from the 4-plex and 2-plex simulation, respectively (Figure 6.8).

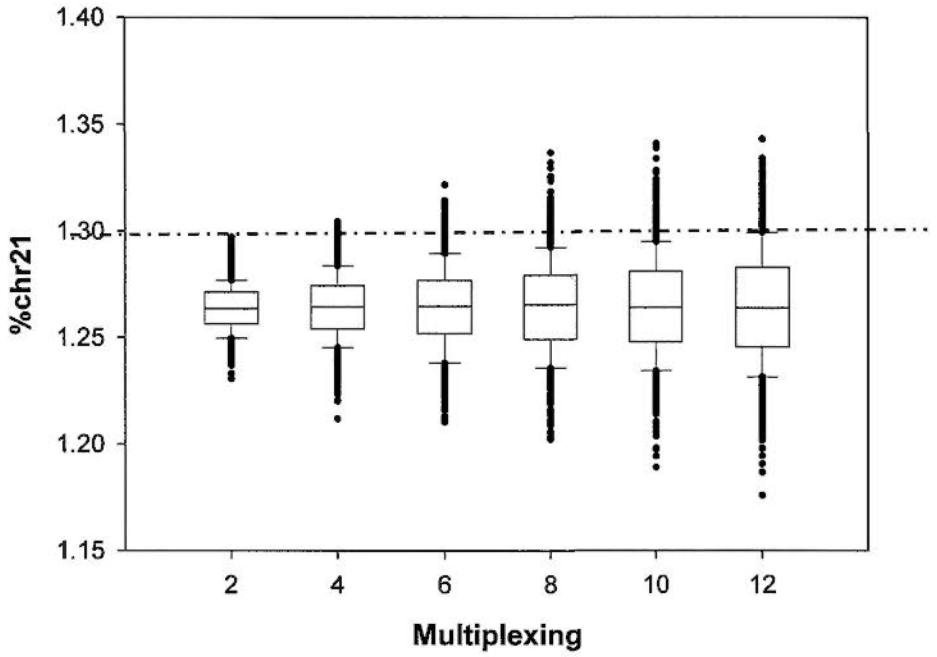


Figure 6.8 Effect of multiplexing level on the quantitative representation of chr21 sequences in maternal plasma.

The result is based on computer simulation with 1000 times' random sampling from the monoplex sequencing data of the euploid case U93619. The *lines inside the boxes* denote medians. The *boxes* mark the interval between the 25th and 75th percentiles. The *whiskers* denote the interval between the 10th and 90th percentile. • indicate data points outside the 10th and 90th percentiles. The dash line indicates the %chr21 value of 1.296%.

The extent of chr21 overrepresentation in maternal plasma for the T21 cases is governed by the fractional fetal DNA concentration. Our group previously sequenced maternal plasma DNA from 10 T21 pregnancies with fetal DNA concentrations ranging from 5.9% to 27.2% by the use of monoplex protocol (Chiu *et al.* 2008). The sequence data from those cases with varying fetal DNA fractions could be used to evaluate the effect of the multiplexing level on the quantitative representation of %chr21 in T21 cases by conducting the same computer simulation as above. The sample information of the previous maternal plasma samples involving T21 fetuses is listed in Appendix III. The results of computer simulation are shown in Figure 6.9. The %chr21 values for the T21 cases (M4386, M1519, M3228 and M3438, from the previous sample set) were generally larger than that for the euploid case (U93619 from the current sample set), but varied among samples because of the different fetal DNA concentrations. Similar to the previous observation, the distribution of %chr21 values broadens as the multiplexing level increased. For the case M4386 and M1519 with relatively low fetal DNA fractions (5.9% and 8.1%, respectively), the probabilities of being smaller than 1.296% were 21.9% and 5.3%, respectively, with the 2-plex simulation. Such probabilities substantially increased to 36.8% and 21.1%, respectively, with the 8-plex simulation. In contrast, for the case M3228 and M3438 with relatively higher fetal DNA concentrations (15.8% and 26.1%, respectively), the chances were only 0.2% and 0% even with the 8-plex simulation. These results indicated that with a higher multiplexing level, the T21 cases containing lower fetal DNA concentrations were more susceptible to lower %chr21 and hence tended towards the false negative diagnostic results. Therefore, when multiplexed sequencing is used for fetal trisomy 21 detection, one may need to select the sequencing protocol for the maternal samples with varying fetal DNA concentrations.

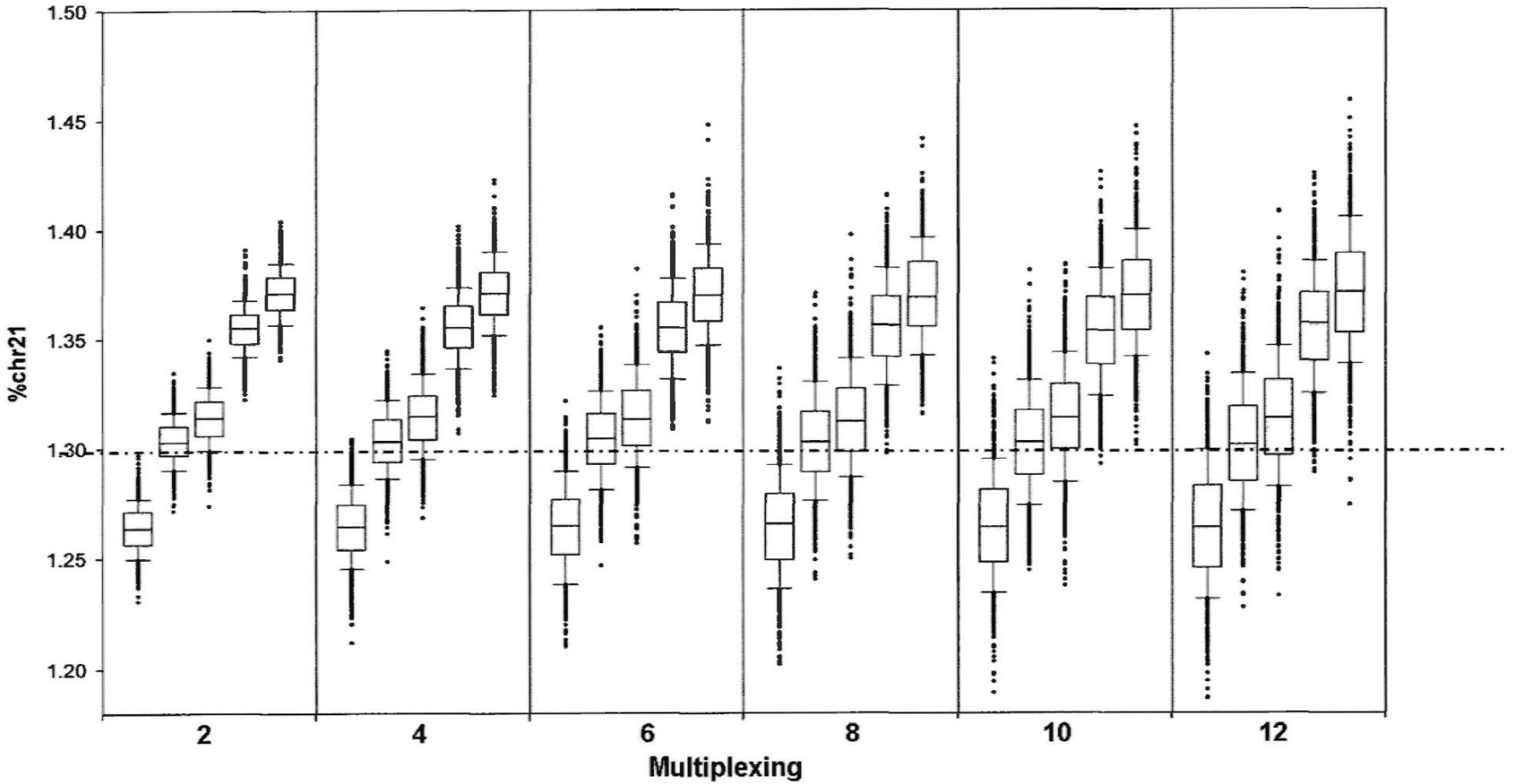


Figure 6.9 Effect of multiplexing level on fetal trisomy 21 detection.

The result is based on computer simulation with 1000 times' random sampling from the monoplex sequencing data (from left to right in each panel: euploid case U93619, T21 cases M4386, M1519, M3228 and M3438). The *lines inside the boxes* denote medians. The *boxes*

mark the interval between the 25th and 75th percentiles. The *whiskers* denote the interval between the 10th and 90th percentile. • indicate data points outside the 10th and 90th percentiles. The dash line indicates the %chr21 value of 1.296%.

6.4 Discussion

NGS-based maternal plasma DNA sequencing analysis permits the noninvasive prenatal diagnosis of fetal trisomy 21 with a high specificity and sensitivity (Chiu *et al.*, 2008a, Fan *et al.*, 2008). Although the proof-of-concept experiments were based on a small sample set, the maternal plasma DNA sequencing analysis have shown its bright future in clinical usage. In this chapter, two issues are investigated in an effort to move this new approach towards the future clinical implementation.

The first part of this chapter shows the feasibility of converting cell-free DNA from low volume of maternal plasma into an Illumina library, followed by cluster generation and sequencing. The results demonstrate that reducing the starting plasma input would not distort genomic representation of plasma DNA, which was further evidenced by the effectiveness in detecting T21 fetuses with $\leq$ 1.2 mL maternal plasma from T21 pregnancies. Starting from a low volume of plasma samples will simplify the laborious procedures before plasma DNA sequencing, thereby facilitating the clinical implementation.

The number of plasma DNA molecules that are present per unit volume of maternal plasma is limited (Lo *et al.* 1998b). Hence, material plasma input is a critical parameter for the locus-specific assays, such as digital PCR (Lo *et al.* 2007a; Lun *et al.* 2008b). For the maternal plasma sample containing 25% fetal DNA, 7,680 PCR analyses are required so as to reach the statistical confidence to determine the fetal status (Lo *et al.* 2007a). With the fetal DNA concentration halved, four times as many digital PCR analyses are needed (Lo *et al.* 2007a). Accordingly, tens of millilitres of maternal blood are needed to perform tens of thousands of digital PCRs. In contrast, for the maternal plasma DNA sequencing analysis by massively parallel

sequencing, the plasma volume would not be a concern. Maternal plasma DNA concentrations are typically hundreds to thousands of GE per milliliter (Lo *et al.* 1998b) and plasma DNA molecules are mainly short DNA fragments, with 86% being shorter than 201 bp (Chan *et al.* 2004). The hapoid human genome occupies a total of 3 billion base pairs. Assuming that each genome was evenly fragmented into 200-bp pieces, there would be billions of fragments in 1 mL of maternal plasma, of which ~1.2% were from chromosome 21. In this scenario, one copy of chromosome 21 would be sampled and counted many times in a locus-independent way, instead of just once in the locus-specific assay. Sequencing, therefore, is able to achieve much higher analytical precision without the need to increase the volume of input maternal plasma.

The second part of this chapter shows the feasibility of sequencing multiple maternal plasma samples concurrently by the use of the barcoding strategy. Multiplexed sequencing at two levels of sample throughput, namely, 8-plex and 4-plex sequencing protocols, are investigated. The 100% accuracy of the 4-plex protocol validates the effectiveness of this strategy. When 4-plex barcodes are employed, the reagent costs per sample (inclusive of library preparation and sequencing procedures) can be reduced by two thirds and the throughput can be quadrupled. However, when using 8-plex barcodes, the false classification occurred. This may relate to the enlarged fluctuation in genomic representation for a single sample as the multiplexing level increases.

Instead of sequencing at high fold coverage, the rational of maternal plasma sequencing is to simply sequence a random representative fraction of the human genome. In our previous and current monoplex datasets, usually ~2 million 36-bp U0-1-0-0 reads could be obtained for each plasma sample to generate a quantitative

profile of chromosomal distribution, which are equivalent to 2.4% of the human genome. It has been demonstrated that with such a sequencing depth, the frequency distribution of the chromosomal origin of the sequenced DNA fragments could well reflect the genomic representation of the original maternal plasma sample and thus an overrepresentation of chromosome 21 could be present for DNA in maternal plasma obtained from a trisomy 21 pregnancy (Chiu *et al.*, 2008a). However, with the increased multiplexing level, the sequenced DNA molecules per case (i.e., the sequencing depth per case) will decrease accordingly. By the 4-plex and 8-plex protocols, the obtained 36-bp U0-1-0-0 reads translate to only 0.6% and 0.3% of the human GE, respectively, thus potentially rendering the resultant chromosomal contribution less representative of the original sample. This point has been validated by the computer simulation. The increased variability in genomic representation for a single sample not only tends to result in an inaccurate reflection for the original sample but also probably enlarges the analytical imprecision for T21 detection if such a case is included in the euploid reference group. The latter one could probably explain why the CVs in the 8-plex data are generally larger than that in the 4-plex data in Figure 6.5C.

Fetal DNA concentration is a key factor for the noninvasive prenatal diagnosis of fetal chromosomal aneuploidy detection. Theoretically, with less fetal DNA presented in the original sample, more molecules would be required to reach the statistical confidence to detect an overrepresentation of chromosome 21 for a T21 pregnancy. Therefore, when using maternal plasma DNA sequencing analysis for fetal trisomy 21 detection, higher sequencing depth would be necessary for the maternal plasma samples with lower fetal DNA concentrations. The T21 female pregnancy that failed to be detected by the 8-plex protocol probably had a relatively

low fetal DNA concentration in maternal plasma. However, in the current analysis, the *SRY/HBB* QPCR assay and %chrY values are only able to determine the fetal DNA concentrations in the maternal plasma samples involving male fetuses; hence, there is no available information of fetal DNA concentration for this female case. Alternatively, one could measure the fetal DNA concentration in a maternal sample by the use of a gender-independent marker, such as the fetal epigenetic signature (Chan *et al.* 2006; Chiu *et al.* 2007). The samples with lower fetal DNA concentrations should be considered to be processed with a higher sequencing depth.

In conclusion, the studies in this chapter demonstrate that reducing plasma volume and barcoding multiple samples are effective measures to simplify the work process and reduce the cost per case as well as increase the sample throughput per run. A large-scale clinical trial is required to validate the diagnostic performance of massively parallel maternal plasma DNA sequencing for fetal trisomy 21 detection. The measures described here set the stage for the upcoming clinical trial and eventually push forward with the clinical implementation of maternal plasma DNA sequencing for noninvasive prenatal diagnosis.

SECTION V : MASSIVELY PARALLEL PAIRED-END SEQUENCING OF PLASMA DNA

CHAPTER 7: MASSIVELY PARALLEL PAIRED-END SEQUENCING OF DNA IN MATERNAL PLASMA FOR NONINVASIVE PRENATAL DIAGNOSIS

7.1 Introduction

Chiu *et al.* (Chiu *et al.* 2008) and Fan *et al.* (Fan *et al.* 2008) applied massively parallel short-read sequencing to analyze DNA in maternal plasma for fetal trisomy 21 detection. These studies were based on single-read (SR) sequencing where a short segment from one end of each plasma DNA molecule was sequenced. For SR sequencing, incidentally, reads aligned to the Y chromosome (< 0.01%), which should only be present in male individuals, were detected in the plasma of women conceived with female fetuses (Chiu *et al.* 2008). Comparisons between different alignment programs suggested that the observation was partly due to alignment errors (Chiu *et al.* 2008). With PE sequencing, whereby both ends of each short DNA molecule are sequenced, the alignment accuracy is expected to be improved.

PE sequencing is typically performed on libraries of DNA fragmented *in vitro* to hundreds of bases in length. Genome rearrangement or structural variations in the sequenced genome is suspected if the PE reads aligned to the reference genome spanning a region with a size or orientation deviating from that expected for the DNA library (Bashir *et al.* 2008; Campbell *et al.* 2008; Chen *et al.* 2009; Maher *et al.* 2009). PE sequencing also facilitates *de novo* sequence assembly using short reads (Farrer *et al.* 2009). Additionally, recent studies have demonstrated the use of PE RNA-sequencing for reliable detection of unannotated transcripts and spliced isoforms (Au *et al.* 2010; Wu *et al.* 2010).

Since DNA molecules in plasma exist *in vivo* as short fragments (Chan *et al.* 2004), direct PE sequencing of plasma DNA is possible and would allow one to obtain a comprehensive and high-resolution size profile of such molecules. Moreover, in maternal plasma, fetal DNA is shorter than maternal DNA (Chan *et al.* 2004). If the precise size profiles of fetal and maternal DNA molecules are known, one may devise strategies for fetal DNA enrichment based on preferential selection (Li *et al.* 2004b) or analysis of the shorter DNA fragments (Lun *et al.* 2008).

In this study, using the Illumina sequencing platform, I explore the use of PE sequencing for the analysis of plasma DNA. The sequence reads allegedly mapped to chrY are studied to determine whether PE sequencing is superior to SR sequencing with regard to the alignment accuracy. The high-resolution size profile of DNA molecules is generated using millions of sequence reads. On the basis of the size profile of maternal plasma DNA, I also investigate the effectiveness of preferential analysis of the shorter DNA fragments identified by PE sequencing for the noninvasive prenatal diagnosis of trisomy 21.

7.2 Methods

7.2.1 Subjects

Peripheral blood and tissue samples were collected according to the description in Chapter 3.1.1.

7.2.2 Sample preparation

Plasma was harvested from blood samples as described in Chapter 3.1.2. DNA was extracted from plasma samples according to procedures described in Chapter 3.2.1.

The extracted plasma DNA was subjected to QPCR as described in Chapter 3.3. Genomic DNA was extracted from placental tissue samples as described in Chapter 3.2.2 and then quantified by Nano-Drop.

7.2.3 Massively parallel paired-end sequencing

The massively parallel paired-end sequencing of plasma DNA was performed on the Illumina GA II system as described in Chapter 3.5.

7.2.4 Sequence alignment, filtering and BLAST validation

The first 32 bp from the 36 bp sequenced reads of each end were aligned to the repeat-masked human genome reference sequence (NCBI Build 36, version 48, downloaded from the Ensembl Genome Browser (http://www.ensembl.org) using the Efficient Large-Scale Alignment of Nucleotide Databases for PE sequencing (eland_pair) program in the GAPipeline-1.0 software package provided by Illumina. The program matches and suggests the most likely pairing of individual sequenced reads. A Perl script was compiled by Dr Nancy Tsui and Mr Peiyong Jiang to identify PE reads meeting the following criteria for subsequent analysis:

1) the individual members of each suggested pair were both sequenced on the same cluster position on the sequencing flow cell and could be aligned to the same chromosome with the correct orientation as expected for the reference human genome;

2) the sequenced reads of both members of the pair could be aligned to the repeat-masked reference human genome without any nucleotide mismatch;

3) the sequenced reads of each member of the pair had a uniqueness score > 4;

4) pairs demonstrated an insert size less than 600 bp.

PE reads meeting these four requirements are termed accepted PE reads. Afterwards, the accepted PE reads for each chromosome were sorted and recorded.

To validate the alignment accuracy of the eland_pair program, 150 accepted PE reads were randomly picked up to confirm if they were unique sequences by performing Basic Local Alignment Search Tool (BLAST) analysis against the human reference genome database in the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). During the BLAST analysis for each set of paired reads, the chromosomal origin, the alignment scores, the identity percentages of the two query reads, and the length of the chromosomal segment between the two aligned sequences were recorded and examined.

7.2.5 *Z*-score calculation

The number of accepted PE reads aligned to each chromosome was expressed as a proportion of all accepted PE reads generated for the sample. The mean and SD of the proportion of accepted PE reads for chromosomes 21 and X were established from the plasma samples of pregnancies with male euploid fetuses which were considered as the reference sample set. The *z*-score for each test sample was then calculated as previously described in Chapter 5.2.5. A *z*-score greater than 3 signifies a difference greater than the 99.9^{th} percentile of the proportion of accepted PE reads of the euploid reference sample set for the target chromosome, i.e., a *P* value of 0.001. The expected genomic representation of a repeat-masked reference genome from a haploid female was calculated as previously described in Chapter 5.2.6.

7.2.6 Fragment size analysis

The length of each DNA fragment was inferred from the data output of the eland_pair program by adding 32 bp to the absolute positional offset between the chromosomal positions at the start of each member of the paired sequence reads.

7.2.7 *In silico* size selection (ISSS)

A series of arbitrarily selected cutoff points, including 300 bp, 200 bp, 175 bp, 150 bp, 125 bp, 100 bp, 75 bp and 50 bp were used to study the effect of size selection on the fetal DNA enrichment and subsequent diagnostic performance. The awk utility of Linux was used to identify the paired reads with a size less than or equal to each of the analyzed size cutoffs. For each size cutoff, the proportion of accepted PE reads for each chromosome was recalculated and then subjected to the calculation of *z*-scores as described above.

7.3 Results

7.3.1 Validation of PE sequencing

Placental tissue DNA from two euploid fetuses and two T21 fetuses were sequenced. The clinical details and sequencing counts for each case are shown in Table 8.1. The proportion of accepted PE reads for each chromosome was close to that expected for the human genome (Figure 7.1A). 1.82% and 1.85% of PE reads from chr21 were obtained from the two T21 placental tissue samples, respectively, which were ~1.5-fold higher than the proportions for the two euploid samples (1.28% and 1.30%, respectively). These data suggested that the measurement of chromosome dosage using PE sequencing was feasible. Next, I checked the workability of PE sequencing of maternal plasma DNA. Three maternal plasma samples (one from a pregnancy with a female fetus and two from pregnancies each with a male fetus) were collected in the third trimester and subjected to PE sequencing (Table 8.1). The percentage of accepted PE reads mapped to each chromosome generally resembled the genomic representation expected for each chromosome in the human genome (Figure 7.1A). The absolute (fractional) accepted PE counts mapped to chrY for the two pregnancies with male fetuses were 710 (0.064%) and 829 (0.079%), respectively, indicating positive detection of fetal DNA by PE sequencing of maternal plasma DNA.

Interestingly, when plotting the measured ratio of the obtained %GR to the expected %GR against chromosomal GC content, the placental genomic DNA and plasma DNA showed reverse patterns. As shown in Figure 7.1B, the average ratios for the maternal plasma DNA had a positive correlation with chromosomal GC content (Pearson Product Moment Correlation, $r = 0.948$, $P = 2.09 \times 10^{-11}$), in line with the observations in chapter 5 and 6, whereas the average measured ratios for the

placental genomic DNA negatively correlated with chromosomal GC content (Pearson Product Moment Correlation, $r = -0.884$, $P = 4.99 \times 10^{-8}$).

Case No.	Sample type	Setting[a]	Gestational age (weeks + days)	Karyotype	Input DNA (ng) [b]	Total sequence count	Accepted PE reads	chrY PE reads	%chrY PE reads (%)
P1519	Placenta	Termination	20 +3	47XY +21	5,000	9,058,120	1,434,795	2,323	0.1619
P4022	Placenta	Termination	19+6	47XX +21	5,000	8,583,212	1,452,396	65	0.0045
P4800	Placenta	Cesarean	28+5	46XY	5,000	8,024,069	1,320,161	2,113	0.1601
P4801	Placenta	Cesarean	39+1	46XY	5,000	8,449,808	1,557,212	2,461	0.1580
M4800	Plasma	Predelivery	38+5	46XY	35	6,228,421	1,104,840	751	0.0680
M4801	Plasma	Predelivery	39+1	46XY	40	5,949,330	1,052,556	848	0.0806
M4814	Plasma	Predelivery	38+5	46XX	30	6,252,582	1,054,953	50	0.0047

a: the placental tissues from trisomy 21 pregnancies were collected after termination of pregnancy; the placental tissues from euploid pregnancies were collected after the delivery of the baby.

b: the placental genomic DNA was quantified by Nano-Drop, while the maternal plasma DNA was quantified by the *HBB* QPCR assay.

Table 7.1 Clinical details and sequencing counts of placental DNA and maternal plasma DNA samples.

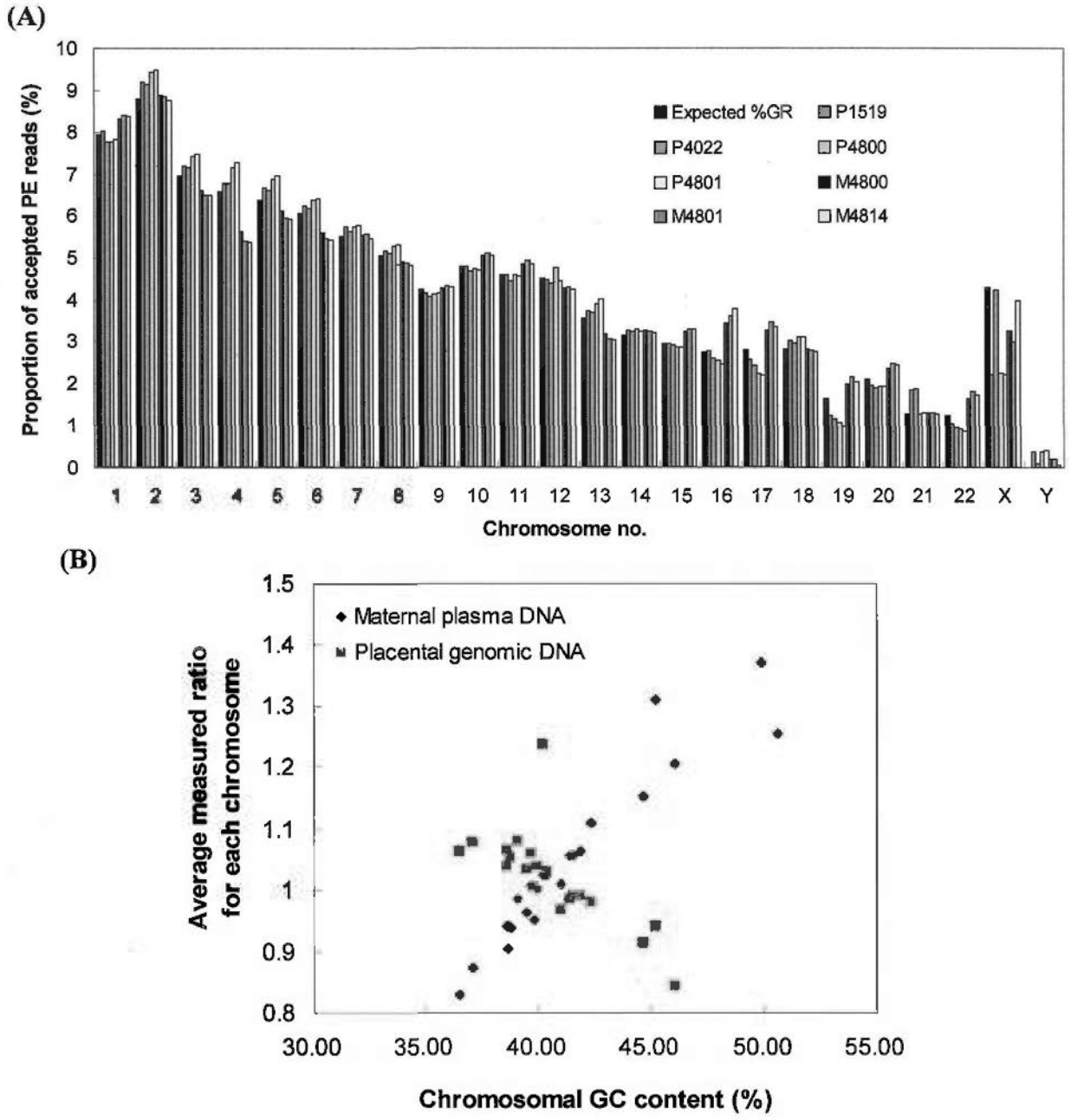


Figure 7.1 Distribution of PE reads among the human chromosomes in placental genomic DNA and maternal plasma DNA.

(A) Bar chart of proportion of accepted PE reads per chromosome for 4 placental genomic DNA samples and 3 maternal plasma samples from euploid pregnancies in the 3rd trimester. The percentage of genomic representation of each chromosome as expected for a repeat-masked reference haploid female genome is plotted for comparison (black bars). *(B)* Correlation between the average measured ratios and chromosomal GC content for the autosomes. The blue and pink dots represent the average ratios calculated from 3 maternal plasma samples and 4 placental genomic DNA samples, respectively.

7.3.2 Validation of alignment accuracy

A small number of accepted PE reads were mapped to chrY in both the maternal plasma sample involving a female fetus (50 reads, 0.0047%) and the female T21 placental tissue (64 reads, 0.0044%). Only 38% of these sequences were confirmed by BLAST analysis to be uniquely mapped to chrY. Similarly, 150 PE reads aligned to chrY were randomly picked from each of the two plasma samples of pregnancies with male fetuses. 90.4% (135 of 150) and 98.0% (147 of 150) of the paired sequences could be aligned uniquely to chrY by BLAST. Also, 150 PE sequences from each of the non-Y chromosomes were randomly selected from each of the three maternal plasma samples for BLAST analysis. Almost all (98.1% for chromosomes 4 and 5, 100% for all other chromosomes) accepted PE reads mapped to the non-Y chromosomes were validated by BLAST to align uniquely and perfectly to the corresponding chromosomes with exactly the same insert size as indicated by the eland_pair output.

In the current data, a small fraction of the reads with apparent mapping to chrY in the female DNA samples were observed (Table 7.1). We reported a similar observation in our previous study using SR sequencing and proved that those false-positive signals were due to non-specific bioinformatics alignment (Chiu *et al.* 2008). In this study, I compared if SR or PE sequencing was more prone to produce such an artifact. For PE sequencing, the reads from the two ends of each DNA fragments are generated independently as read1 and read2, respectively, and are paired by post-sequencing bioinformatics. Therefore, read1 from the PE sequencing run could be analyzed as if it was SR sequencing. When analyzed as SR sequencing, the absolute (and fractional) U0-1-0-0 sequence reads mapped to chrY for the two female DNA samples described above were 147 (0.0094%) and 171 (0.0072%),

respectively. This was almost doubled that of the corresponding accepted PE reads mapped to chrY shown in Table 7.1.

7.3.3 Identification of trisomy 21 fetus using PE sequencing

Nine women each pregnant with a euploid fetus and four women each pregnant with a T21 fetus were recruited in the first and second trimesters. Direct noninvasive detection of fetal trisomy 21 from maternal plasma was attempted based on similar methodological principles as described in Chapter 5 and 6 except that PE instead of SR sequencing was used, and accepted PE reads instead of U0-1-0-0 sequences were quantified. The clinical details and sequencing counts for each case are shown in Table 7.2. In total, 8.3−10.5 million DNA molecules were sequenced for each case, of which a median of 1.6 million pairs (17% of total) passed the criteria to be deemed as accepted PE reads.

Figure 7.2A shows that the percentages of accepted PE reads aligned to chr21 were higher for all T21 pregnancies than for euploid pregnancies and the values for chrX were higher and those for chrY were lower for all female pregnancies than male pregnancies. The ranges of proportions of accepted PE reads aligned to chromosome Y were 0.022-0.034% for the pregnancies with euploid male fetuses, 0.0048-0.0058% for pregnancies with euploid female fetuses and 0.029-0.038% for pregnancies with T21 male fetuses. Five maternal plasma samples each carrying a euploid male fetus were selected as the reference group for the calculation of the *z*-scores. To simulate the data obtained when SR sequencing was performed, the percentages of U0-1-0-0 reads from read1 of the PE sequencing run were used to calculate *z*-scores of the corresponding SR sequencing. *Z*-scores of chr21 for the four T21 fetuses ranged from 5.63−8.89 for SR sequencing and ranged from 8.07−12.00 for PE sequencing. *Z*-scores of chrX for the four female fetuses ranged from 5.04–7.69 for SR sequencing and ranged from 3.91−6.35 for PE sequencing. There were no statistically significant differences in the *z*-scores for chromosomes 21 or X

when comparing the PE and SR sequencing data (Wilcoxon signed-rank test, P = 0.125).

Case no.	Gestational age (weeks + days)	Karyotype	Input DNA (ng)	Total sequence count	U0-1-0-0 counts of read1[a]	Accepted PE reads
6	12 + 5	46XY	4.1	9,517,549	1,850,429	1,002,834
7	13 + 5	46XY	6.1	9,593,914	2,697,877	1,734,737
8	12 + 6	46XY	13	9,945,029	2,760,386	1,786,242
4467	14 + 4	47XY +21	4.3	10,368,770	2,615,616	1,686,055
4620	12 + 4	47XY +21	6.3	9,628,874	2,758,156	1,806,950
9	17 + 2	46XX	6.6	9,777,804	2,653,199	1,681,648
10	17 + 1	46XX	5.2	9,511,784	2,505,064	1,618,291
16	12 + 4	46XX	6.7	8,878,573	2,216,696	1,341,424
22	13 + 6	46XX	3.2	10,041,629	2,531,538	1,329,515
12	13	46XY	10.3	8,768,612	2,394,173	1,364,798
20	13 + 3	46XY	7.1	8,319,431	2,240,088	1,430,068
2849	14 + 3	47XY +21	5.9	10,104,453	2,414,339	1,570,738
4386	13 + 6	47XY +21	12.6	10,533,749	2,777,640	1,837,304

a: PE sequencing reads out the 36-bp information from both ends of DNA fragments. The first read could be regarded as an independent sequence read and subsequently used for U0-1-0-0 counting.

Table 7.2 Clinical details and sequencing counts of maternal plasma DNA samples.

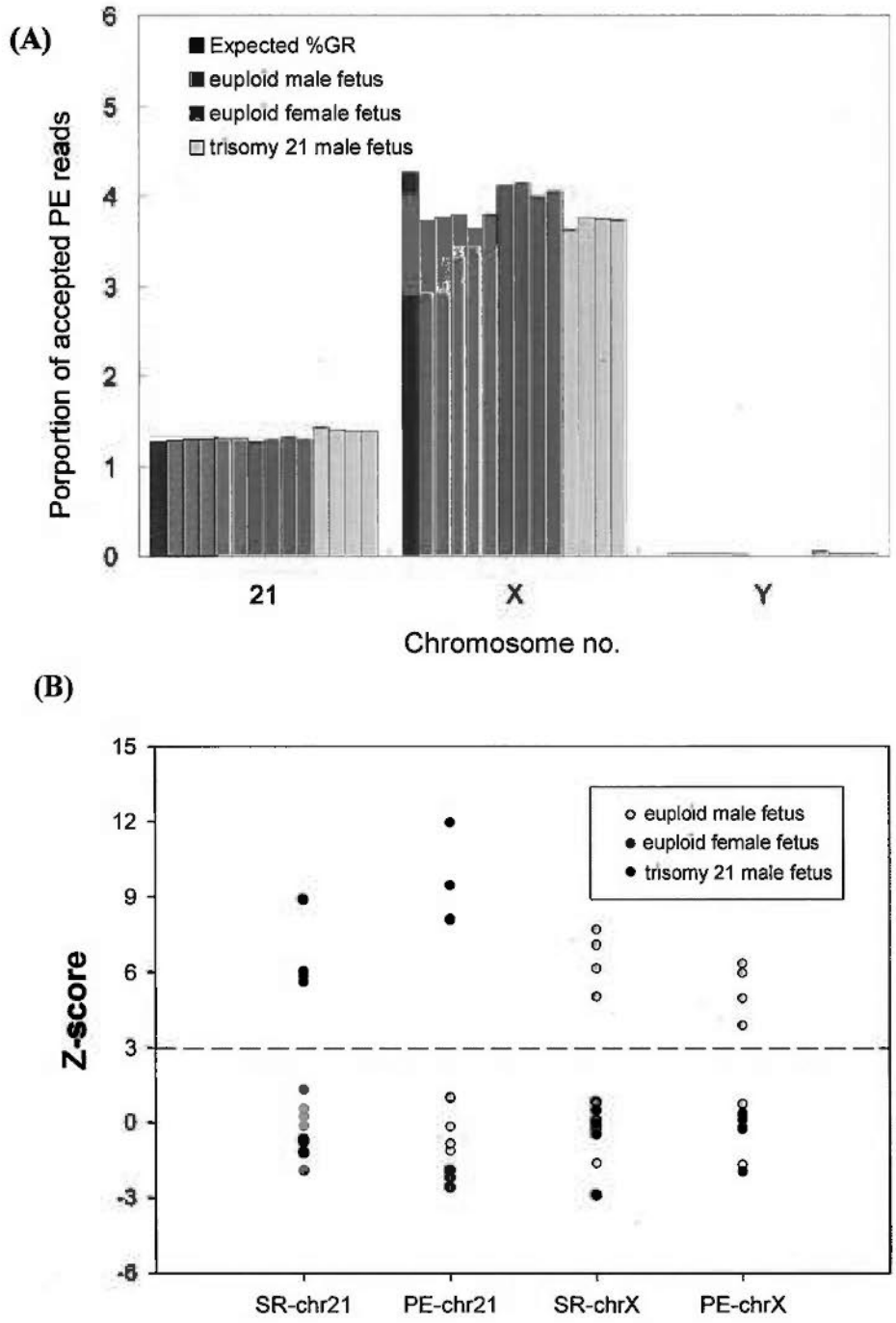


Figure 7.2 Fetal trisomy 21 detection.

(A) Bar chart of proportion of accepted PE reads for chromosomes 21, X and Y for the maternal plasma samples. The percentage of genomic representation as expected for a repeat-masked reference haploid female genome was plotted for comparison (black bars). *(B)* *Z*-scores of chromosome 21 and X for the test samples and reference samples using SR (read 1 of two reads) and PE sequencing data. The dashed line represents a *z*-score of 3 as a diagnostic cutoff.

7.3.4 Size distribution of DNA fragments in maternal plasma

One of the major advantages of PE sequencing approach is that it enables us to deduce the fragment size of each sequenced molecule, thus allowing a detailed size profile of DNA molecules in plasma at an unprecedented resolution. I proceeded to study the size profiles of plasma DNA from the nine pregnant women carrying male fetuses among the 13 pregnancies described above and plasma DNA from 4 adult males. After sequence alignment and filtering, the median accepted PE reads for the additional plasma samples from adult males were 1.58 million (range, 1.38 -1.78 million). For the maternal plasma samples, the reads mapped to chrY are of fetal origin while the reads for the other chromosomes are predominantly of maternal origin. I therefore analyzed the size profile of the reads aligned to the Y and non-Y chromosomes independently. There was no significant difference in size distribution between the Y and non-Y chromosomes for four adult male plasma samples (Table 7.3). A representative result for one adult male plasma sample is shown in Figure 7.3A. On the contrary, for the maternal plasma samples, there was a clear demarcation between the size distribution curves for the Y and non-Y chromosomes (Figure 7.3B). The fragments aligned to chrY were significantly shorter than those aligned to non-Y chromosomes for each maternal plasma sample (Table 7.3). Thereby, millions of PE reads demonstrated that the fetal-derived DNA molecules in maternal plasma were indeed shorter than maternal-derived ones, confirming the previous findings using QPCR (Chan *et al.* 2004).

Case no.	Sample type	Chromosome	Median	25th percentile	75th percentile	P value [a]
6	maternal plasma (euploid male fetus)	Non-Y	154	122	172	<0.001
		Y	135.5	107.5	162	
7	maternal plasma (euploid male fetus)	Non-Y	160	135	173	<0.001
		Y	141	113	163	
8	maternal plasma (euploid male fetus)	Non-Y	158	128	172	<0.001
		Y	127	90	155	
4467	maternal plasma (T21 male fetus)	Non-Y	149	113	169	<0.001
		Y	132.5	94	158	
4620	maternal plasma (T21 male fetus)	Non-Y	159	133	173	<0.001
		Y	144	120	164	
12	maternal plasma (euploid male fetus)	Non-Y	157	129	171	<0.001
		Y	134	103	162	
20	maternal plasma (euploid male fetus)	Non-Y	150	111	169	<0.001
		Y	133	100.25	161.75	
2849	maternal plasma (T21 male fetus)	Non-Y	157	129	171	<0.001
		Y	139	105.25	163	
4386	maternal plasma (T21 male fetus)	Non-Y	157	124	171	<0.001
		Y	131	94	156	
Male01	plasma from adult male	Non-Y	164	147	175	0.118
		Y	163	146	175	
Male02	plasma from adult male	Non-Y	165	149	175	0.134
		Y	164	147	175	
Male03	plasma from adult male	Non-Y	167	148	180	0.277
		Y	166	147	180	
Male04	plasma from adult male	Non-Y	164	144	175	0.262
		Y	163	143	175	

a: based on Mann-Whitney U test

Table 7.3 Summary statistics of fragment size of plasma DNA from the Y and non-Y chromosomes.

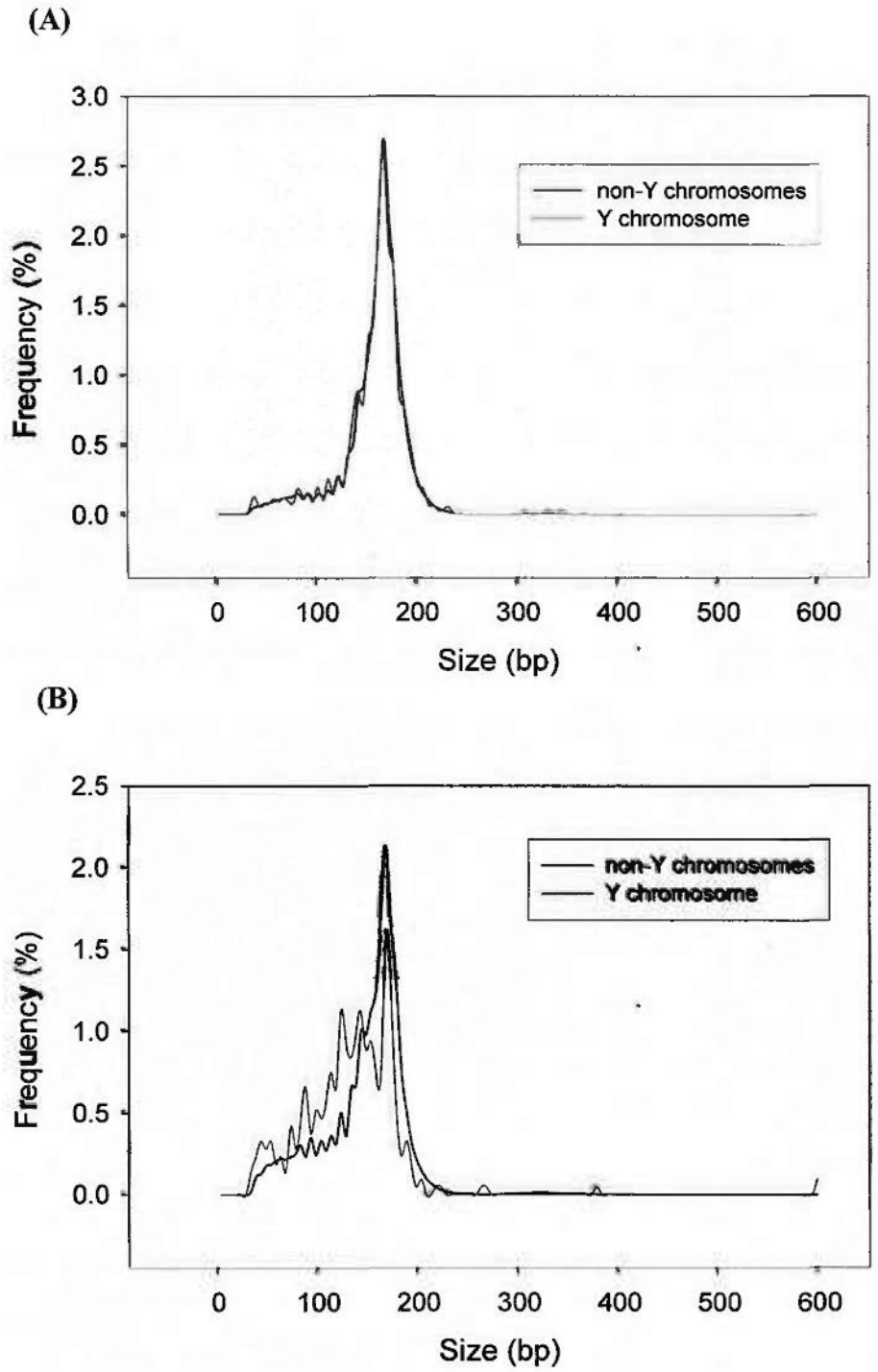


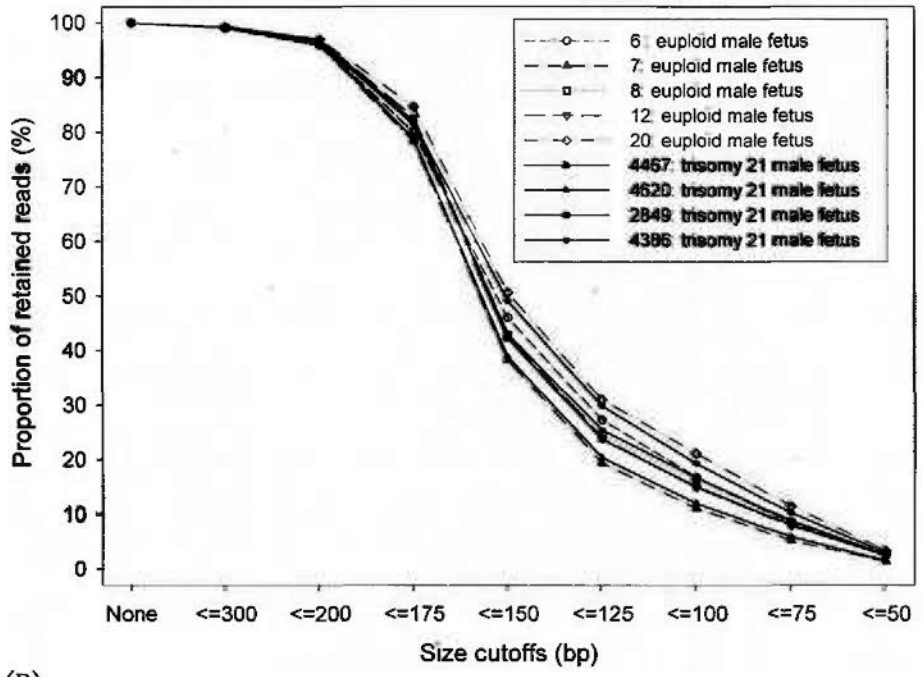
Figure 7.3 Representative size profiles of plasma DNA fragments.

Histograms (at 5-bp resolution) show the size distributions of accepted PE reads aligned to Y (red line) and non-Y (black line) chromosomes in the plasma from ***(A)*** an adult male and ***(B)*** a pregnant woman carrying a male fetus.

7.3.5 Enrichment of fetal DNA by ISSS

Since the fetal DNA molecules are significantly shorter than the maternal background ones, it is possible to enrich the fetal DNA at the post-sequencing stage by ISSS. Therefore, I next investigated whether any cutoff for DNA size could be used to achieve relative enrichment of fetal DNA in maternal plasma. I compared a series of arbitrarily selected cutoff points, including 300 bp, 200 bp, 175 bp, 150 bp, 125 bp, 100 bp, 75 bp and 50 bp. The proportions of retained reads at each size cutoff are shown in Figure 7.4A. I then determined the amount of retained reads from chrY as a proportion of all retained reads, termed retained %chrY. The optimal balance between the degrees of fetal DNA enrichment achieved with a reasonable retention of accepted PE reads seemed to be achieved at the cutoff points of 150 bp and 125 bp (Figure 7.4B).

(A)



(B)

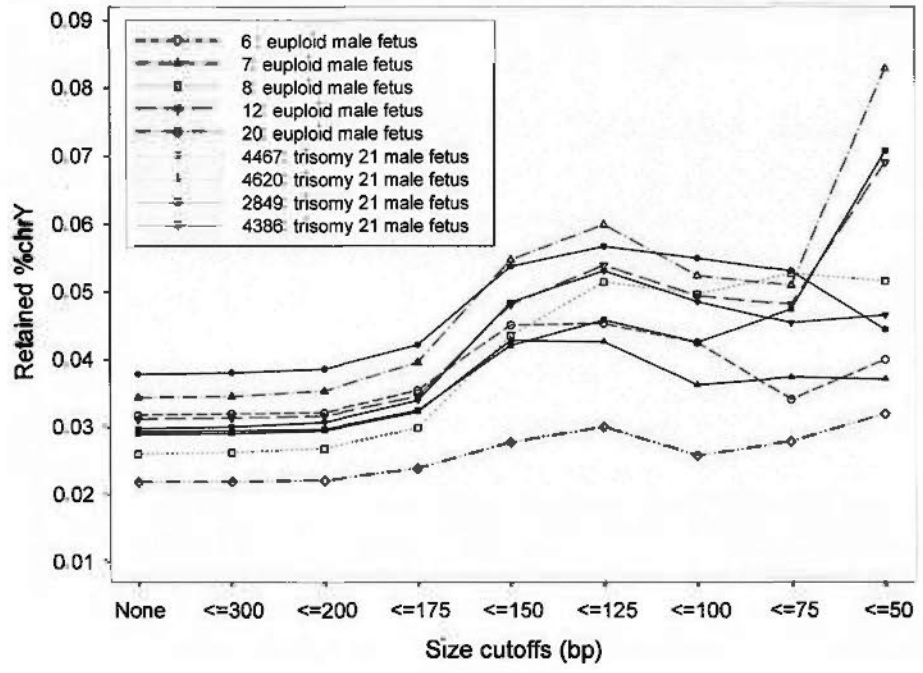


Figure 7.4 Effects of ISSS analysis.

Line charts show the effects of ISSS analysis on ***(A)*** the proportion of retained reads among all accepted PE reads and ***(B)*** the percentage of chromosome Y among the retained reads. Each line represents a maternal plasma sample.

7.3.6 Effect of ISSS on fetal trisomy 21 detection

Fetal chromosomal aneuploidy could be detected more readily by maternal plasma analysis in samples with higher fractional fetal DNA concentrations (Lo *et al.* 2007a) However, as demonstrated in Chapter 6, the detection of overrepresentation of the trisomic chromosome would be less precise when the absolute read number is reduced. To investigate the effect of ISSS strategy on the noninvasive prenatal detection of trisomy 21, I revisited the *z*-score of chr21 for each of the size cutoffs. A clearer demarcation in the *z*-scores of chr21 was achieved between the euploid and T21 cases when size cutoffs of 150 bp or above were used, but at 125 bp or less, the demarcation blurred (Figure 7.5A). The CVs (CV=SD/mean × 100%) were calculated for measuring the proportion of retained chr21 reads at each size cutoff using the euploid reference cases. The CV increased substantially when a size cutoff of 125 bp or less was used (Figure 7.5B).

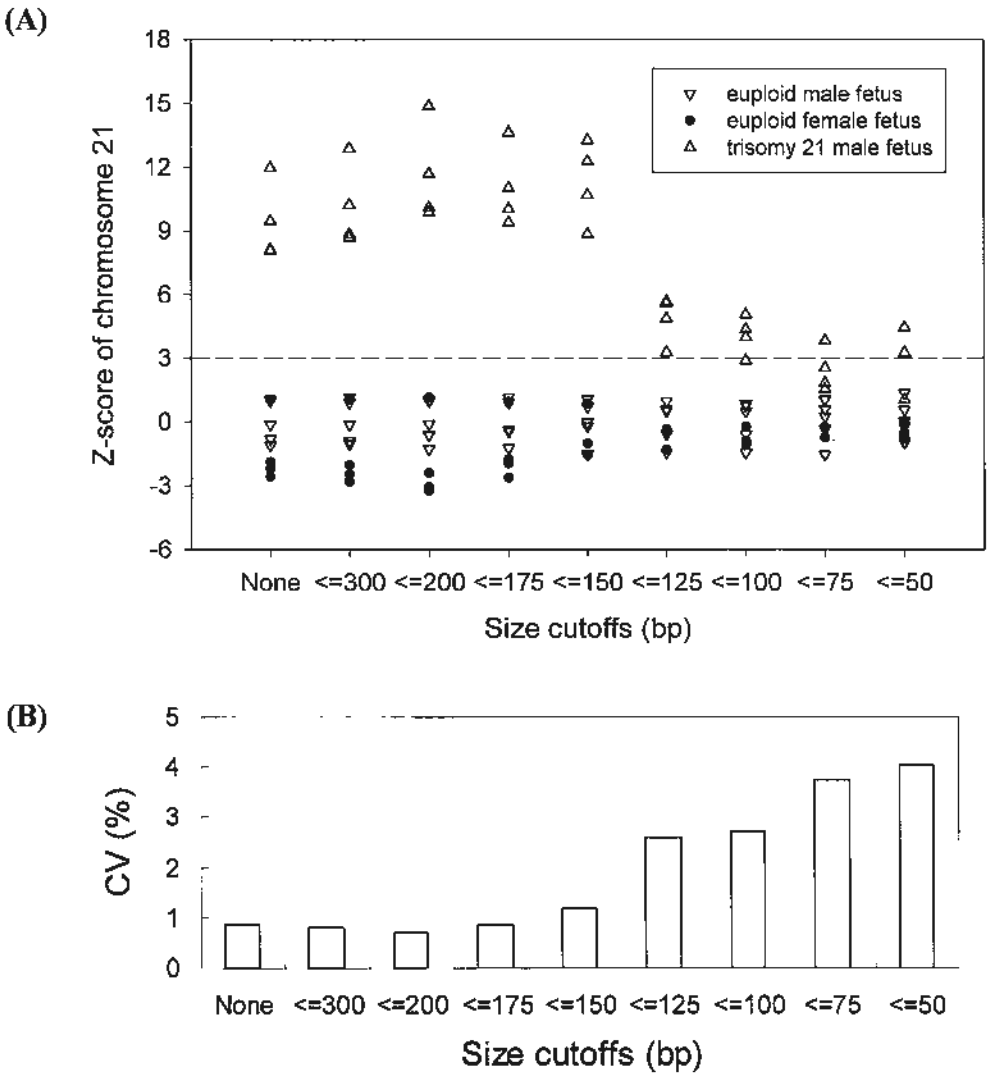


Figure 7.5 Application of ISSS analysis for fetal trisomy 21 detection.

(A) *Z*-scores of chromosome 21 at each DNA size cutoff. The dashed line represents a *z*-score of 3 as a diagnostic cutoff. ***(B)*** CV for measuring the genomic representation of chromosome 21 at each DNA size cutoff.

7.4 Discussion

In this chapter, I demonstrated the use of PE massively parallel sequencing for fetal DNA analysis and aneuploidy detection. PE massively parallel sequencing can be applied to the analysis of plasma DNA molecules which exists as short fragments by nature (Chan *et al.* 2004; Jahr *et al.* 2001). As evidenced by the chrY data from the female samples, PE sequencing can attain higher alignment accuracy than SR sequencing. This is possibly because the number of nucleotides sequenced and therefore available for alignment from each plasma DNA molecule is doubled in PE compared with SR sequencing, minimizing the chance of misalignment to other locations in the human genome. The positional requirement of not accepting pairs separated by too great a distance on the same chromosome is another potential reason why the chance of misalignment is reduced.

However, the median number of unique reads for PE sequencing, namely the accepted PE reads, was just 17% (~ 1.6 million reads) of the total sequenced reads while that for SR sequencing (U0-1-0-0 sequences of read1) of the same sample set was 26.4% (~ 2.5 million reads) of the total sequenced reads (Table 1). The difference was statistically significant (Wilcoxon signed-rank test, $P < 0.001$). The latter data were similar to those (23.3%, ~ 2.4 million reads) reported in an earlier study where 28 maternal plasma samples were analyzed using SR sequencing (Chiu *et al.* 2008). The reduced number of unique read counts for PE sequencing is possibly because of the more stringent definition of uniqueness whereby both reads in a pair, i.e. 64 bp, would need to align to the reference human genome without mismatches. Despite the reduced number of unique sequences, it was shown that PE sequencing of maternal plasma DNA allowed the detection of fetal DNA and the

assessment of fetal chromosome dosage in a similar manner as previously reported for SR sequencing (Chiu *et al.* 2008)..

The genomic representation of plasma DNA molecules originating from different chromosomes bore a significant relationship to the chromosomal GC content. It is interesting that the placental tissues fragmented by nebulization also showed a relationship between chromosomal GC content and genomic representation, but in a reverse direction. Though the reason for the observation is uncertain, one possibility is that the GC content of a DNA sequence has opposite effects on the natural fragmentation processes of plasma DNA compared with artificial processes such as nebulization. On the other hand, the difference in the processing between the two kinds of samples could be partially responsible for such an observation. For example, a gel-cutting step was involved during the tissue DNA library preparation, where a gel slice was selected and the DNA was extracted. It has been reported that melting the gel slice by heating may affect the representation of AT-rich sequences (Quail *et al.* 2008).

PE sequencing allows one to assess the size profile of plasma DNA at single molecule resolution. Previous investigations on the size of plasma nucleic acids were based on the comparison of locus-specific PCR amplicons of different lengths (Chan *et al.* 2003a; Chan *et al.* 2004; Diehl *et al.* 2005). Those locus-specific PCR assays will amplify any plasma DNA molecules that are larger than and contain the target amplicon. Thus, those approaches would determine the proportion of plasma DNA molecules that are at least beyond the size of the target amplicon and thus provide a crude estimation of differences in the size profile of DNA molecules across different plasma samples. In contrast to the locus-specific PCR approaches, massively parallel plasma DNA sequencing provides the size information for each sequenced DNA

molecule. Thus, a frequency distribution plot could be compiled. Fan *et al.* (Fan *et al.* 2008) used 454 sequencing to assess the DNA size profile of one maternal plasma sample by sequencing the full length (up to 250 bp) of each molecule. These authors found that the size distribution for the fetal derived chrY sequences was shorter than the non-Y chromosomes (Fan *et al.* 2008) but suggested that the analysis of more cases was required to confirm the finding. Though the PE sequencing was used instead, the present data from 9 male pregnancies and four adult male controls showed that fetal derived DNA sequences were indeed statistically significantly shorter than other DNA molecules in plasma.

By knowing the detailed size profile of DNA molecules in maternal plasma, one could objectively predict the effects of fetal DNA enrichment based on size selection of the shorter sequences. The size selection could be done by physical means such as gel electrophoresis(Li *et al.* 2004b). Alternatively, one could selectively analyze the shorter sequences at the post-sequencing stage (Lun *et al.* 2008). Selective analysis of plasma DNA sequences shorter than a specified size cutoff would indeed increase the proportion of fetal derived sequences but at a reduction in the absolute number of retained sequences. The overrepresentation of chr21 sequences for T21 pregnancies should be more apparent in dataset with enriched fetal DNA proportion. However, the reduced total number of sequenced reads would render the measurement of the representation of chr21 less precise. There is therefore a tradeoff between the extent of fetal DNA enrichment and reduction in overall retained reads when any particular size cutoff value is used. The effects of the chosen size cutoff would be reflected in the CV for the measurement of the representation of chr21. Less precise measurements, reflected by a larger CV, would result in larger SDs and thus reduce the *z*-score demarcation between the aneuploid and euploid cases.

In conclusion, PE sequencing shows the comparable diagnostic performance with SR sequencing. Although relatively lower throughput and higher reagent costs than SR sequencing, PE sequencing has the distinct advantages over SR sequencing such as higher alignment accuracy and the provision of the size information on the sequenced DNA molecules. The latter one would be important for further studying the biological implications of plasma nucleic acids, which would be discussed in the next chapter.

CHAPTER 8: BIOLOGICAL IMPLICATIONS FROM PAIRED-END SEQUENCING OF PLASMA DNA

8.1 Introduction

Many promising diagnostic applications of circulating nucleic acids have been demonstrated for noninvasive prenatal diagnosis and cancer detection/monitoring, yet much remains to be learnt regarding their cellular origin, release and clearance mechanisms. Different hypotheses of the release mechanisms have been proposed, including DNA release after cellular apoptosis and/or necrosis (Jahr *et al.* 2001; Lui *et al.* 2002) and active cellular release (van der Vaart *et al.* 2007). Apart from production, the clearance of circulating DNA is also poorly understood. Previous data have suggested that circulating fetal DNA is cleared very rapidly from maternal plasma (Lo *et al.* 1999b). Potential mechanisms for circulating DNA clearance include plasma nucleases and hepatic and renal clearance (Lo 2001).

Better characterization of the size distribution of plasma DNA in different clinical settings, e.g., in healthy subjects, cancer patients and pregnant women, could improve our understanding of the release and elimination mechanisms of plasma DNA. Previous investigations into the size distribution of circulating DNA have provided valuable information. Jahr *et al.* demonstrated that cell-free cancer DNA exists in lengths which are multiples of 180 bp in the circulation by gel electrophoresis, corresponding to the DNA fragmentation during cell apoptosis (Jahr *et al.* 2001). Using PCR-based approaches, Chan *et al.* showed that circulating EBV DNA (with 87% being shorter than 180 bp) and fetal DNA (with 86% being shorter than 201 bp) molecules in the plasma are both relatively small in size (Chan *et al.* 2003a; Chan *et al.* 2004).

NGS technologies enable one to achieve a complete analysis of circulating DNA at single molecule resolution. Recently, using the 454 Genome Sequencer FLX system, Beck *et al.* have profiled the circulating DNA in apparently healthy individuals (Beck *et al.* 2009). However, in their study, main focus was on the genomic sequence representation of plasma DNA and little information regarding the molecular size of plasma DNA was provided (Beck *et al.* 2009). In this chapter, by the use of the Illumina sequencing platform, I investigate the size profiles of plasma DNA from both pregnant women and healthy individuals. The inter-individual and intra-individual comparisons of the size distribution of plasma DNA are performed to investigate the dynamic change in fragment size during pregnancy and the epigenetic effect on fragment size of plasma DNA.

8.2 Methods

8.2.1 Subjects

In addition to the plasma samples from 4 healthy adult males and 16 pregnant women described in the last chapter, plasma samples were collected from 3 healthy nonpregnant adult female volunteers.

8.2.2 Sample preparation

Plasma was harvested from blood samples as described in Chapter 3.1.2. DNA was extracted from maternal plasma according to procedures described in Chapter 3.2.1. The extracted DNA was subjected to QPCR as described in Chapter 3.3.

8.2.3 Massively parallel paired-end sequencing of plasma DNA

The massively parallel paired-end sequencing of plasma DNA was performed on the Illumina GA II system as described in Chapter 3.5.

8.2.4 Sequence and size analyses

The alignment and selection of PE sequence reads were executed according to the same criteria described in Chapter 7.2.4. The fragment size of each sequenced plasma DNA molecule was deduced as described in Chapter 7.2.6.

8.2.5 Size cutoff analysis

A program was compiled by Mr Peiyong Jiang to identify the paired reads with fragment size less than or equal to a defined size cutoff. A series of arbitrarily selected cutoff points, including 300 bp, 200 bp, 175 bp, 150 bp, 125 bp, 100 bp, 75 bp and 50 bp, were used to calculate the proportion of DNA fragments shorter than or equal to each size cutoff inside each size distribution (e.g., size distribution of plasma DNA fragments derived from chromosome Y (chrY) and plasma DNA fragments from non-Y chromosomes).

8.2.6 Size ranking analysis

For each case, 22 autosomes and chrX were compared in terms of the fragment size of sequences aligned to them. Fragments from all 23 chromosomes were ranked according to their size in an ascending order (i.e., the shortest fragment ranks 1 and the longest fragment with the highest ranking). Then, the rankings for all fragments mapped to the same chromosome were added together. The sum of the rankings was then divided by the number of fragments aligned to the particular chromosome to arrive at the average ranking of fragment size for that chromosome. The chromosome

with the largest average ranking of fragment size would be the longest (i.e., ranks 1) and the chromosome with the smallest one would be the shortest (i.e., ranks 23).

8.2.7 Methylation analysis

Fragment size analysis of plasma DNA within differentially methylated CpG islands (CGIs) was performed by Prof Sun Hao. This analysis was on the basis of a recent work published by Zeschnigk *et al.*, who studied the methylation status of CGIs in female blood DNA by massively parallel bisulfite sequencing of CG-rich DNA fragments and identified a number of fully methylated and differentially methylated CGIs (Zeschnigk *et al.* 2009). Their sequencing output file (Blood.bed), containing the chromosomal coordinates and the discovered methylation status for each record, was accessible on the website (http://hmg.oxfordjournals.org/cgi/content/full/ddp054/DC1). The chromosomal coordinates of accepted PE reads in our sequencing data were cross-referred to those of their records. Once overlapped, the fragment size of sequence reads from our data was recorded along with the corresponding methylation status of the CGIs from their data for downstream comparison.

8.3 Results

8.3.1 Massively parallel paired-end sequencing of plasma DNA

Plasma DNA samples from 3 healthy nonpregnant adult females were sequenced. The accepted PE reads obtained from these three samples were 1,585,067, 1,621,704 and 1,776,377, respectively. The downstream analyses were performed by gathering all accumulated data from PE sequencing of plasma DNA. Table 1 summarizes the samples involved in the current study.

Case no.	Sample group	Gestational age (weeks + days)	Fetal sex	Fetal karyotype
P1519	Placental tissue	20 +3	male	47XY +21
P4022		19+6	female	47XX +21
P4800		28+5	male	46XY
P4801		39+1	male	46XY
M4800	Maternal plasma each involving a male fetus (late pregnancy)	38+5	male	46XY
M4801		39+1	male	46XY
M4814	Maternal plasma each involving a female fetus (late pregnancy)	38+5	female	46XX
6	Maternal plasma each involving a male fetus (early pregnancy)	12 + 5	male	46XY
7		13 + 5	male	46XY
8		12 + 6	male	46XY
12		13	male	46XY
20		13 + 3	male	46XY
9	Maternal plasma each involving a female fetus (early pregnancy)	17 + 2	female	46XX
10		17 + 1	female	46XX
16		12 + 4	female	46XX
22		13 + 6	female	46XX
2849	Maternal plasma each involving a male trisomy 21 fetus (early pregnancy)	14 + 3	male	47XY +21
4386		13 + 6	male	47XY +21
4467		14 + 4	male	47XY +21
4620		12 + 4	male	47XY +21
Female01	Plasma from nonpregnant adult females	n/a	n/a	n/a
Female02		n/a	n/a	n/a
Female03		n/a	n/a	n/a
Male01	Plasma from adult males	n/a	n/a	n/a
Male02		n/a	n/a	n/a
Male03		n/a	n/a	n/a
Male04		n/a	n/a	n/a

Table 8.1 Clinical information of all samples in the current study.

8.3.2 Size distribution of plasma DNA

Overall size distribution pattern

For each plasma sample, the histogram of fragment size of all sequenced molecules was plotted to obtain an overall picture of size distribution of plasma DNA molecules. Representative results are shown in Figure 8.1. The sequenced DNA molecules in plasma were found to be short fragments, with > 90% of which being shorter than 200 bp and ~1% being from 300~400 bp. This overall size distribution determined by PE sequencing is concordant with the previous observations from quantitative PCR (Chan *et al.* 2004) and conventional cloning and sequencing (Suzuki *et al.* 2008). These data support the hypothesis that the circulating cell-free DNA in plasma is mainly derived from apoptotic cells (Jahr *et al.* 2001; Lui *et al.* 2002). Strikingly, despite the different groups of plasma samples, each histogram showed a sharp peak at 166 bp, which is thought to be a standard size of DNA component in a chromatosome (a nucleosome with one bound linker histone) (Widom 1992). Interestingly, the sequenced plasma DNA molecules from the adult females were observed to possess a distinctive pattern of size distribution, where a kind of bimodal pattern arose up with a lower peak showing up at around 50 bp besides the much higher peak at 166 bp.

To quantitatively evaluate the difference in fragment size among sample groups, I selectively used several size cutoffs and calculated the proportion of fragments shorter than or equal to each size cutoff in the respective plasma samples. The intra-individual and inter-individual differences in fragment size were then compared based on this size cutoff analysis.

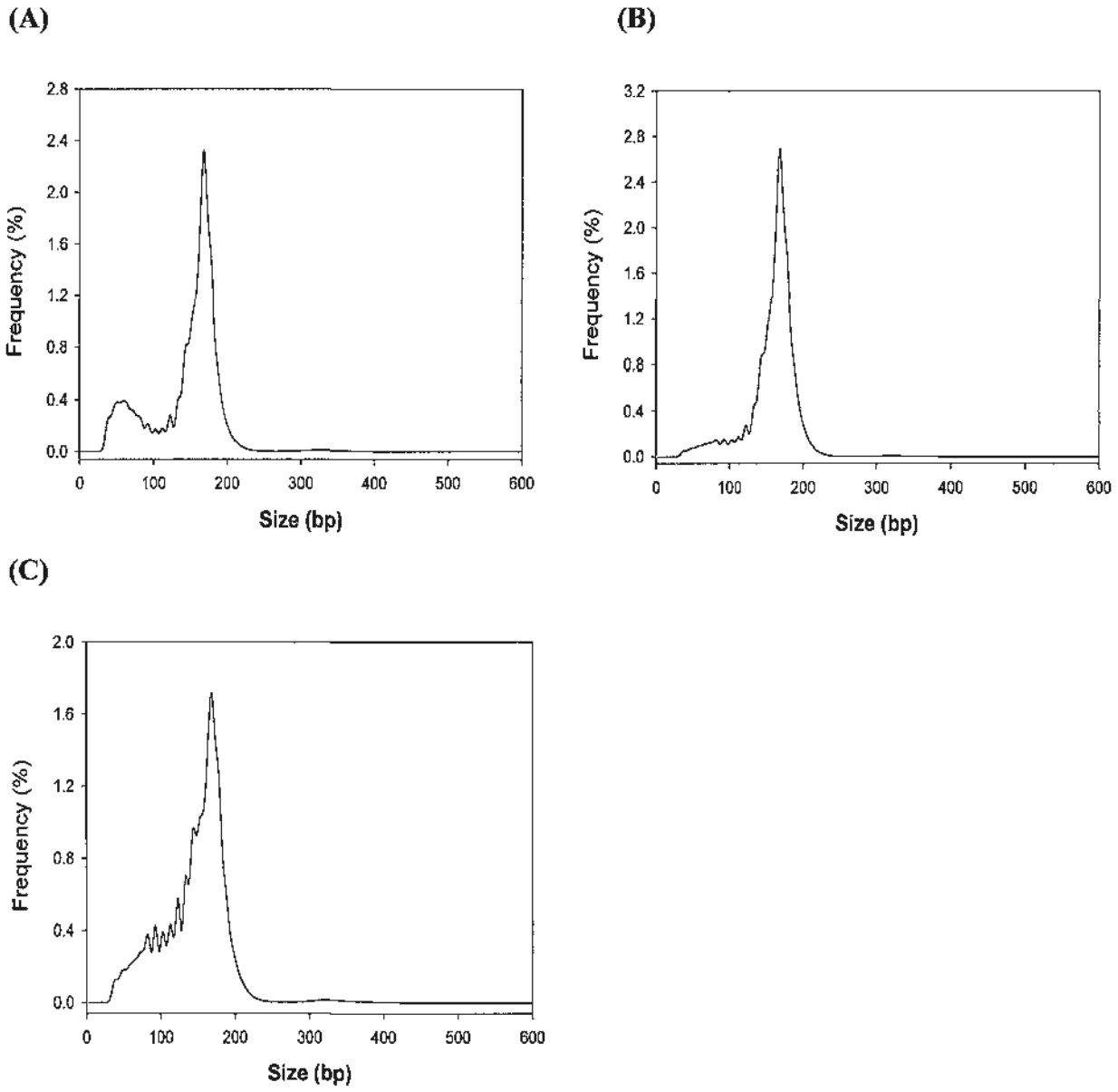


Figure 8.1 Histograms of fragment size of sequenced plasma DNA molecules.

Histograms (at 5-bp resolution) show the overall size distributions of all accepted PE reads in the plasma samples of ***(A)*** an adult female (Female 01), ***(B)*** an adult male (Male 01) and ***(C)*** a pregnant woman bearing a male fetus (6).

Size cutoff analysis of plasma DNA in adult females and males

I first applied the cutoff analysis to compare the plasma DNA from non-Y chromosomes between the adult female and the adult male groups. The results are shown in Figure 8.2A. Despite not reaching statistical significance (Mann–Whitney rank-sum test, $P = 0.057$), the proportions at the size cutoffs of 50 bp, 75 bp, 100 bp, 125 bp, 150 bp and 175 bp in adult females tended to be larger than adult males, indicative of shorter plasma DNA in adult females than adult males.

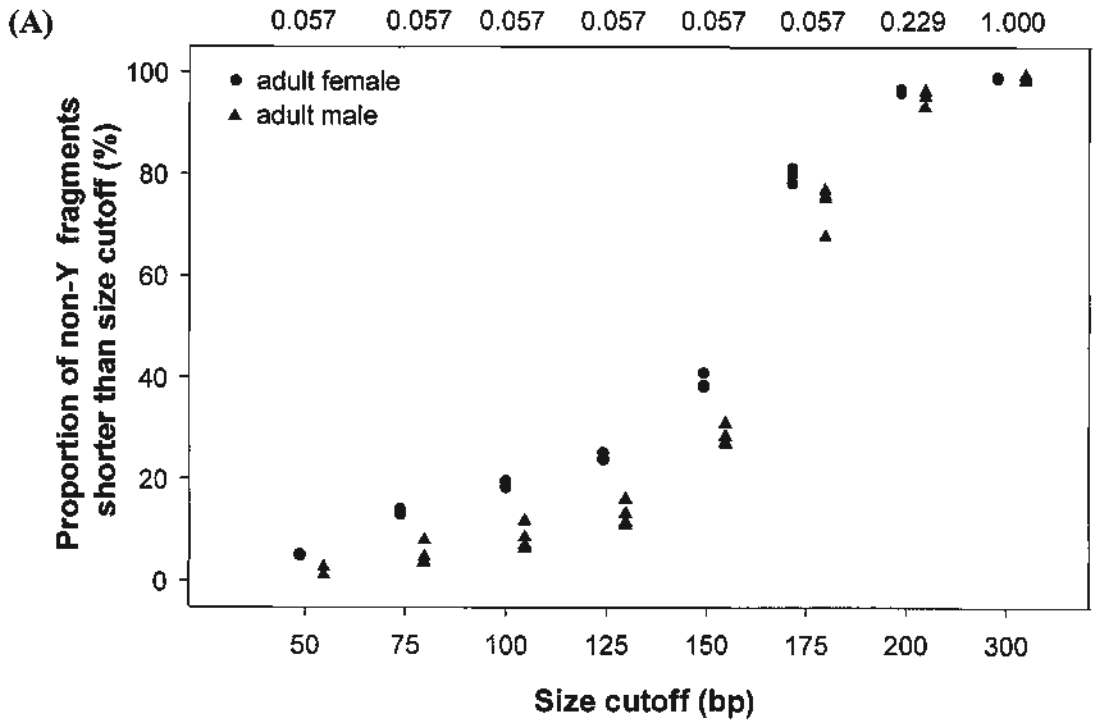
Size cutoff analysis of fetal-derived DNA in maternal plasma

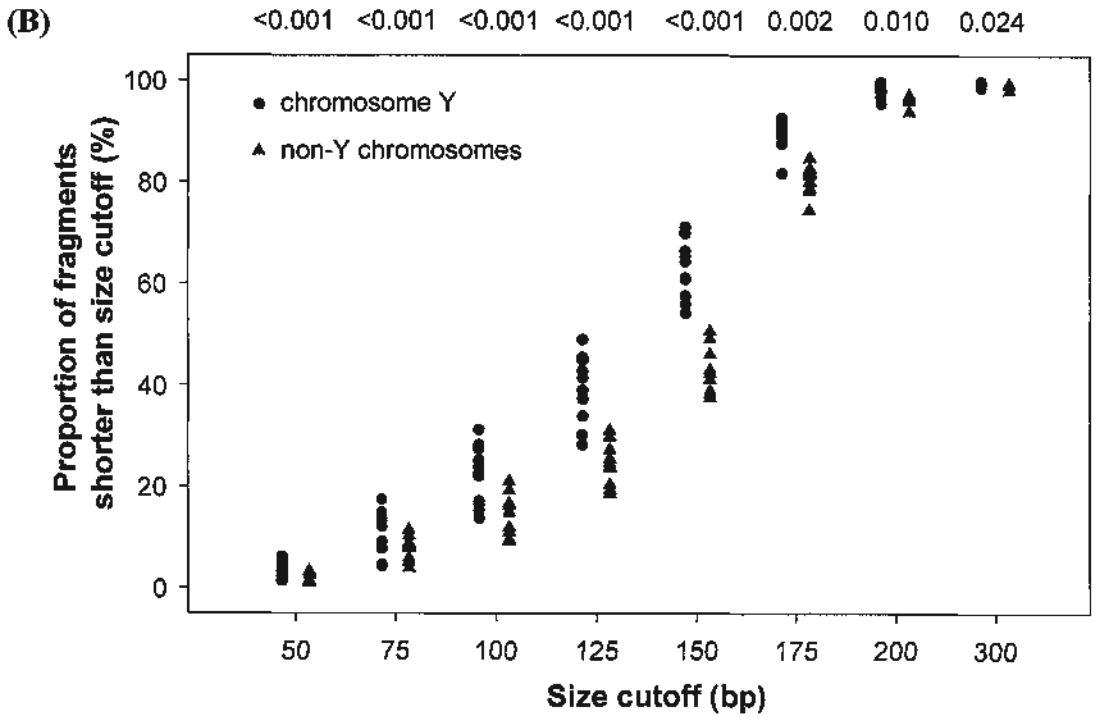
In the last chapter, I have demonstrated that the fetal-derived DNA molecules are significantly shorter than the maternally-derived ones by comparing the overall size distributions for Y and non-Y chromosomes. Here, the same conclusion could be drawn by the size cutoff analysis of plasma DNA from the Y and non-Y chromosomes for all 11 male pregnancies. As shown in Figure 8.2B, larger proportion of short DNA fragments from chrY than the non-Y chromosomes were observed at each size cutoff (Wilcoxon signed-rank test, $P < 0.05$ for each of the size cutoffs).

On the other hand, focusing on the chrY fragments for male pregnancies, one could observe a discrepancy of chrY fragments in maternal plasma between early (< 18 weeks) and late pregnancies (> 38 weeks). The proportions of chrY fragments in maternal plasma obtained from late pregnancy were found significantly lower than those from early pregnancy at the cutoffs of 50 bp, 75 bp and 150 bp (Mann–Whitney rank-sum test, $P < 0.05$) (Figure 8.2C), suggesting the fetal-derived DNA molecules might shift towards longer fragments in late pregnancy.

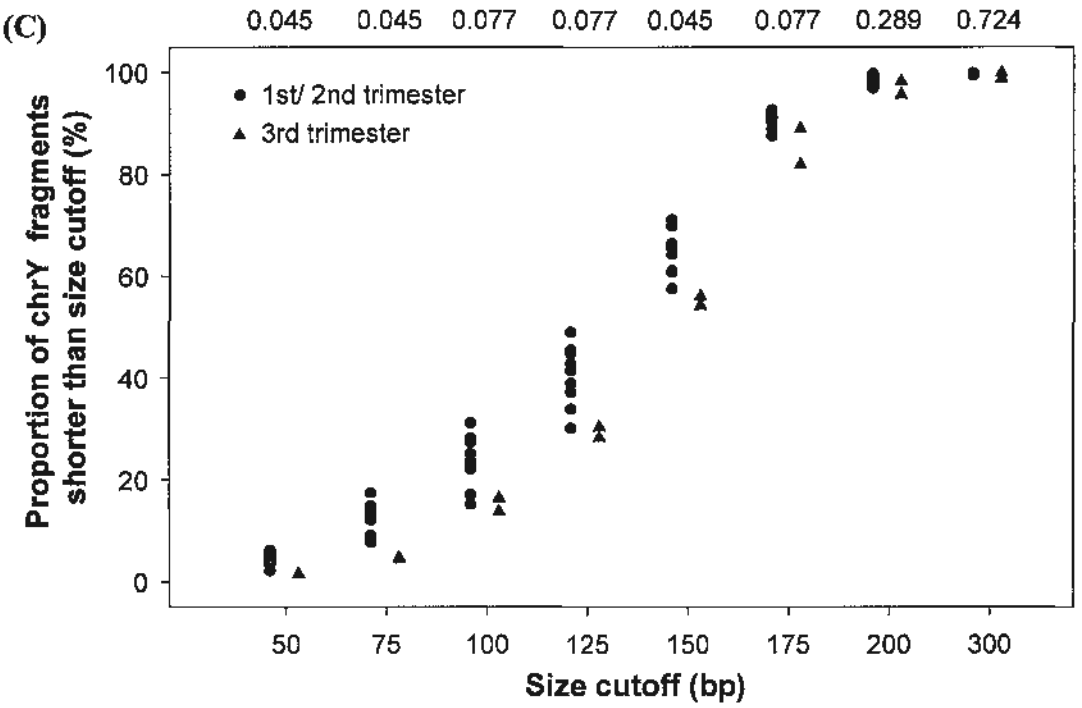
Size cutoff analysis of plasma DNA in pregnant and nonpregnant women

Because fetal DNA constitutes a minority of total DNA maternal plasma (Lo *et al.* 1998a), the plasma DNA fragments from non-Y chromosomes predominantly represent the maternal DNA molecules. Hence, I analyzed the plasma DNA from non-Y chromosomes between 12 pregnant women carrying euploid fetuses and 3 nonpregnant women. As shown in Figure 8.2D, the pregnant women had lower proportions at the cutoffs of 50 bp and 75 bp than nonpregnant women (Mann–Whitney rank-sum test, $P < 0.05$). Among the 12 euploid pregnancies, 9 were collected in early pregnancy (< 18 weeks) and 3 in late pregnancy (> 38 weeks). The maternal plasma samples obtained from late pregnancy showed lower proportions than those from early pregnancy at the cutoffs of 50b bp, 75 bp, 100 bp and 125 bp (Mann–Whitney rank-sum test, $P < 0.05$) (Figure 8.2E). Such difference between early and late pregnancies suggested a dynamic change in fragment size of maternal plasma DNA as pregnancy progressed.

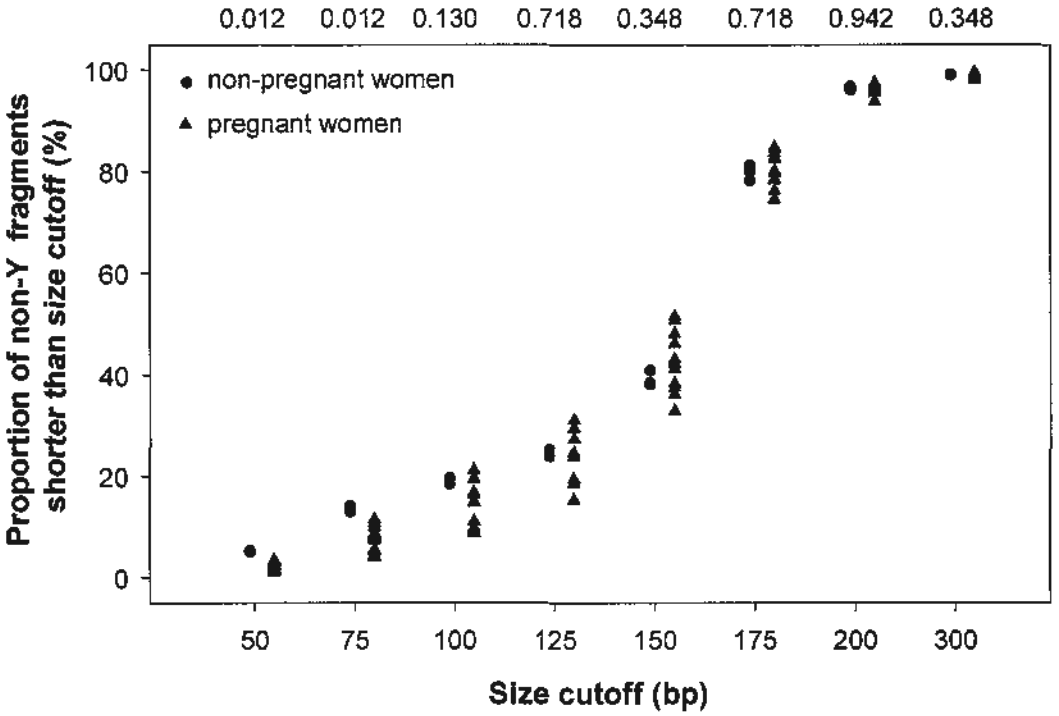

(A)
0.057 0.057 0.057 0.057 0.057 0.057 0.229 1.000
adult female
adult male
Proportion of non-Y fragments shorter than size cutoff (%)
0
20
40
60
80
100
50 75 100 125 150 175 200 300
Size cutoff (bp)


(B)
<0.001 <0.001 <0.001 <0.001 <0.001 0.002 0.010 0.024
chromosome Y
non-Y chromosomes
Proportion of fragments shorter than size cutoff (%)
0
20
40
60
80
100
50 75 100 125 150 175 200 300
Size cutoff (bp)

(C)


0.045
0.045
0.077
0.077
0.045
0.077
0.289
0.724
1st/ 2nd trimester
3rd trimester
Proportion of chrY fragments shorter than size cutoff (%)
100
80
60
40
20
0
50
75
100
125
150
175
200
300
Size cutoff (bp)

(D)


0.012
0.012
0.130
0.718
0.348
0.718
0.942
0.348
non-pregnant women
pregnant women
Proportion of non-Y fragments shorter than size cutoff (%)
100
80
60
40
20
0
50
75
100
125
150
175
200
300
Size cutoff (bp)

(E)

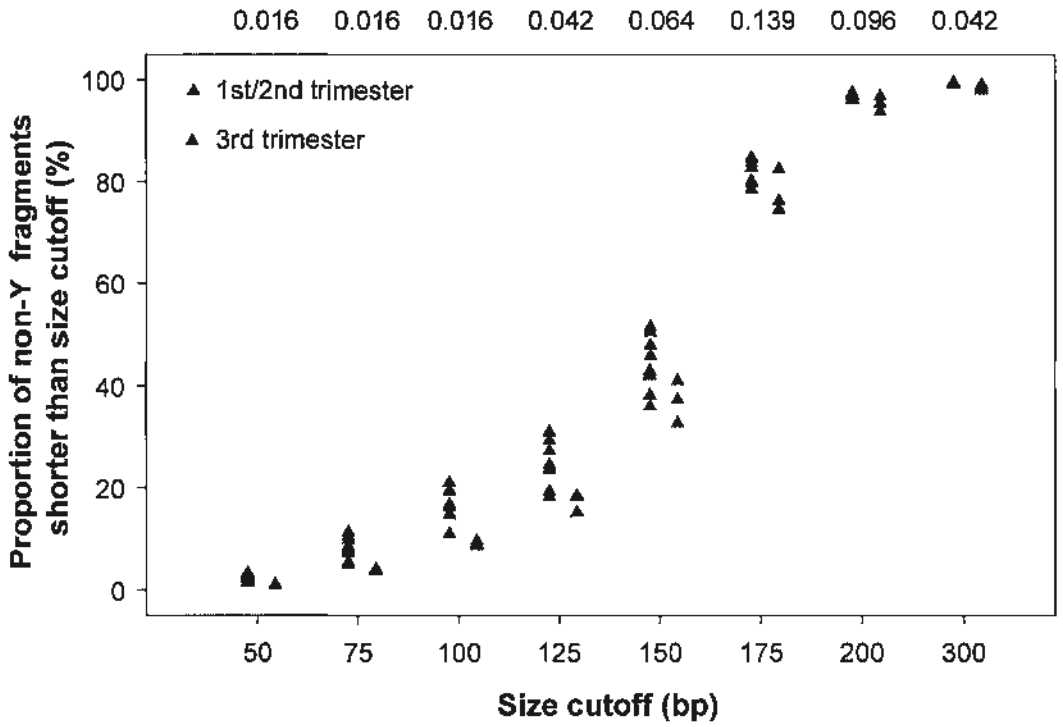


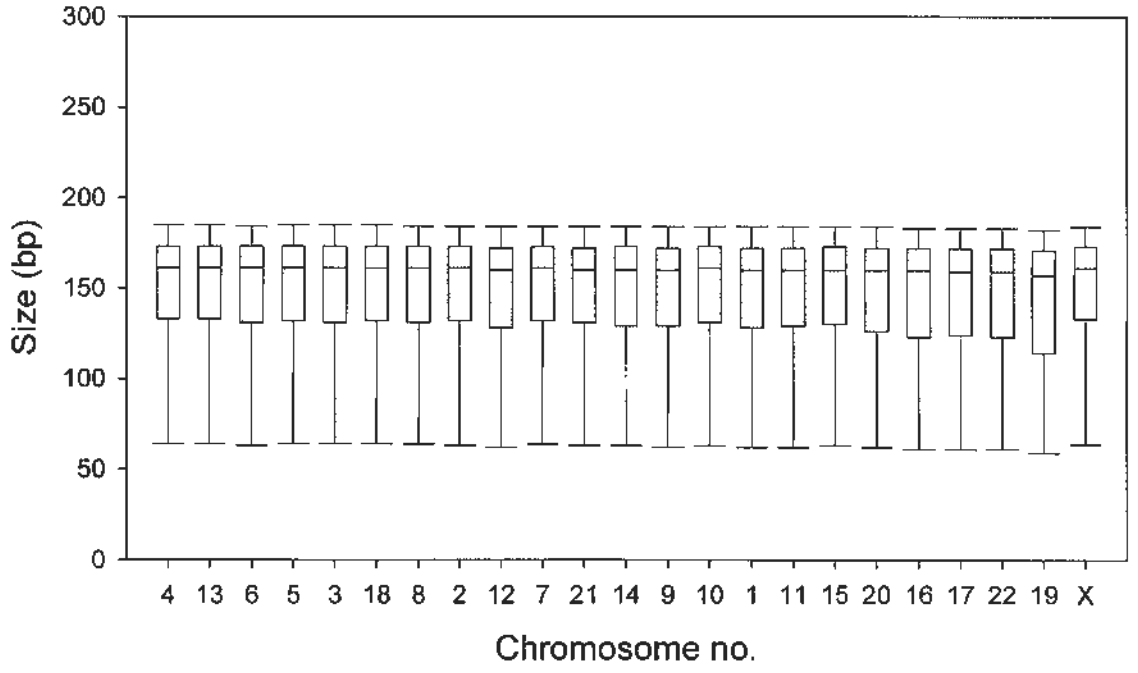
Figure 8.2 Results from the size cutoff analysis of plasma DNA.

(A) The proportion of plasma DNA from non-Y chromosomes at each cutoff is compared between adult males and females. Numbers above the plots represent the P values obtained with the Mann–Whitney rank-sum test for the respective cutoffs. ***(B)*** The proportion of plasma DNA at each cutoff is compared between the Y and non-Y chromosomes for all 11 male pregnancies. Numbers above the plots represent the P values obtained with the Wilcoxon signed-rank test for the respective cutoffs. ***(C)*** The proportion of chrY DNA fragments at each cutoff is compared between early (1^{st} or 2^{nd} trimester, gestational age < 18 weeks) and late (3^{rd} trimester, gestational age > 38 weeks) pregnancies. Numbers above the plots represent the P values obtained with the Mann–Whitney rank-sum test for the respective cutoffs. ***(D)*** The proportion of plasma DNA from the non-Y chromosomes at each cutoff is compared between nonpregnant women and pregnant women. Numbers above the plots represent the P values obtained with the Mann–Whitney rank-sum test for the respective cutoffs. ***(E)*** The proportion of plasma DNA from non-Y chromosomes at each cutoff is compared between early and late pregnancies. Numbers above the plots represent the P values obtained with the Mann–Whitney rank-sum test for the respective cutoffs.

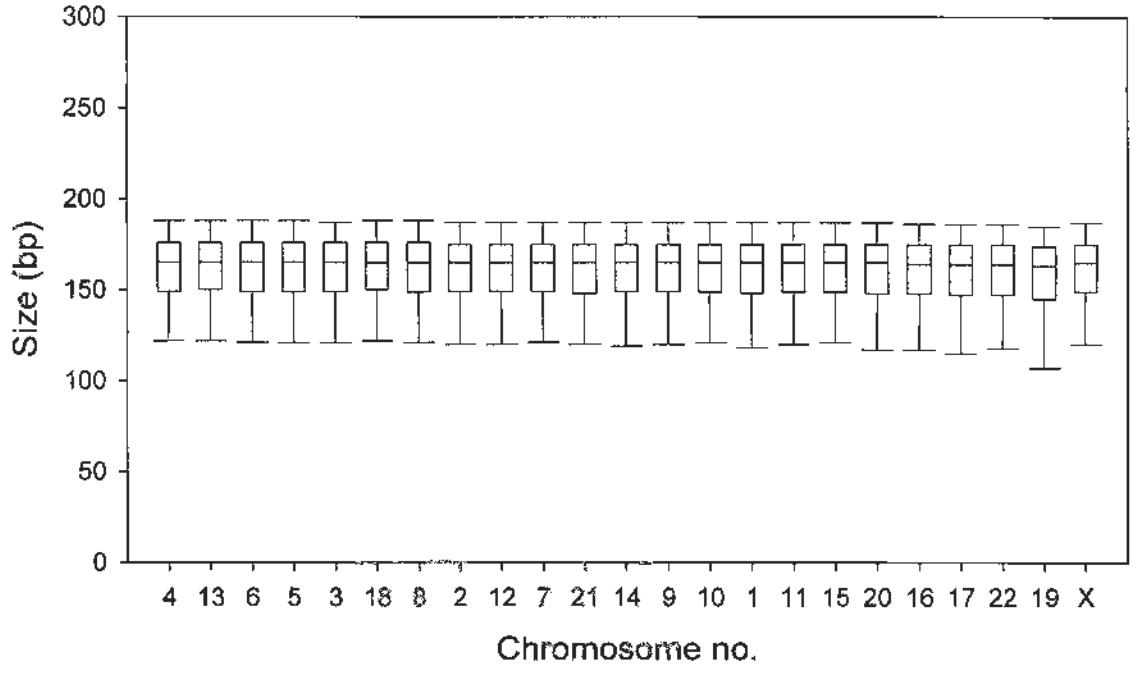
8.3.3 Fragment size of plasma DNA and GC content

Figure 8.3 shows the size distributions of plasma DNA fragments from each chromosome, corresponding to the cases in Figure 8.1. An inter-chromosomal variation in size distribution of plasma DNA fragments was observed and a tendency towards shorter size distribution with increasing chromosomal GC content was found. As mentioned in Chapter 5, GC content varies among human chromosomes which can be broadly categorized into five groups, referred as groups I to V for chromosomes from the lowest to the highest GC contents (Kel-Margoulis *et al.* 2003). Kruskal-Wallis One Way Analysis of Variance on Ranks followed by Bonferroni-corrected pairwise comparisons revealed that there were statistically significant differences in size distribution of plasma DNA fragments between chromosomes from groups V and I, groups V and II, and groups V and III for each plasma sample ($P < 0.001$ for all plasma samples). To enlarge the tiny magnitude of size difference, I ranked (from longest to shortest) chromosomes in terms of the size distribution of DNA fragments aligned to them for each plasma sample. An intriguing relationship between the size ranking and the chromosomal GC content was observed (Figure 8.4A). Interestingly, the reverse pattern was observed for the four mechanically sheared placental tissue genomic DNA (Figure 8.4B).

(A)


300
250
200
150
100
50
0
Size (bp)
4 13 6 5 3 18 8 2 12 7 21 14 9 10 1 11 15 20 16 17 22 19 X
Chromosome no.

(B)


300
250
200
150
100
50
0
Size (bp)
4 13 6 5 3 18 8 2 12 7 21 14 9 10 1 11 15 20 16 17 22 19 X
Chromosome no.

(C)

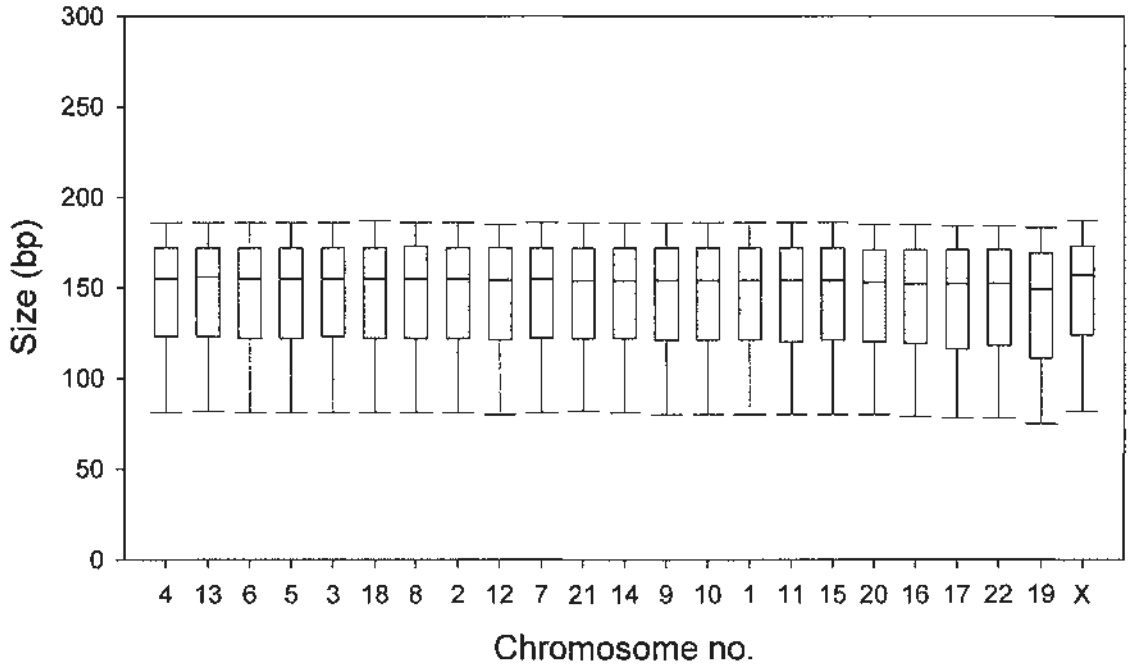


Figure 8.3 Size distributions of plasma DNA fragments for each chromosome.

Box-plots show the size distributions for each chromosome in the plasma sample from ***(A)*** the same adult female as Figure 8.1A, ***(B)*** the same adult male as Figure 8.1B and ***(C)*** the same pregnancy with a male fetus as Figure 8.1C. The *lines inside the boxes* denote medians. The *boxes* mark the interval between the 25th and 75th percentiles. The *whiskers* denote the interval between the 10th and 90th percentile. Outliers beyond the 10th and 90th percentiles are not plotted. On the X-axis, the autosomes are arranged in an ascending order of chromosomal GC content, followed by chrX.

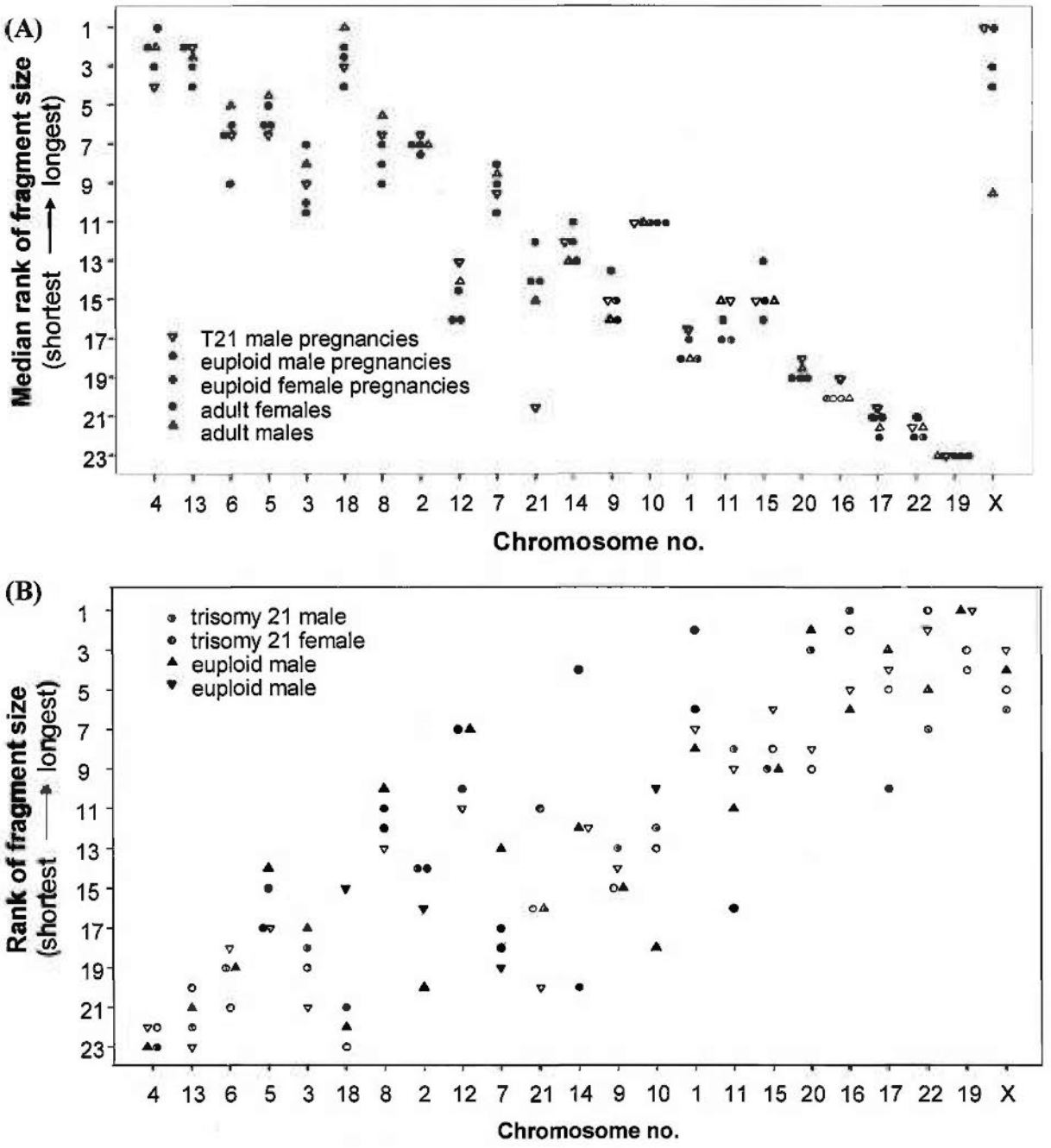


Figure 8.4 Chromosomal GC content and the size rankings of DNA fragments from plasma DNA and randomly sheared placental tissue DNA.

(A) The median ranks of fragment size for each chromosome in the plasma samples from euploid pregnancies with male or female fetuses, trisomy 21 pregnancies with male fetuses and adult males as well as adult females are shown. ***(B)*** The ranks of fragment size for each chromosome in 4 placental tissue DNA samples. On the X-axis, the autosomes are arranged in an ascending order of chromosomal GC content. ChrX has been placed on the right of the diagram because its ranking is governed by the sex of the fetus. On the Y-axis, the rankings of fragment size are arranged in a discending order, i.e., the longest ranks 1 while the shortest ranks 23.

8.3.4 Fragment size of plasma DNA from chromosome X

From our data, the size ranking of chrX fragments varied across different sample groups. The chrX DNA fragments for the adult male plasma samples ranked lower than both nonpregnant and pregnant women (Figure 8.4A). As most of the chrX DNA molecules in maternal plasma are derived from the pregnant woman, these data suggested that chrX DNA molecules from an adult female were longer than those of an adult male. The female genome has two doses of chrX, one of which is active and the other is inactive, whereas the male genome has one dose of chrX. In female mammals, most genes on one X chromosome are silenced as a result of X-chromosome inactivation (Lyon 1961; Plath *et al.* 2002). We therefore suspected that the size difference of chrX DNA fragments in plasma between females and males might relate to the distinct characteristics of chrX in the two sample groups.

8.3.5 Methylation effect on fragment size of plasma DNA

In the study performed by Zeschnigk *et al.*, the methylation status of CGIs in female blood DNA was examined by the massive parallel bisulfite sequencing (Zeschnigk *et al.* 2009). Since previous work has demonstrated that plasma DNA molecules are mainly derived from the hematopoietic system (Lui *et al.* 2002), it would be of particular interest to investigate the fragment size of plasma DNA from female subjects in relation to the methylation status of CGIs in female blood DNA. Among 10,303 CGI regions studied in that paper, 861 (8.4%) CGI regions were highly methylated (with methylation percentage being $\geq$ 60%) and 9,442 (91.6%) GCI regions were lowly methylated (with methylation percentage being $<$ 60%). To accumulate the number of DNA fragments that could be overlapped with the studied CGI regions, the accepted PE reads within the same sample group were pooled

together for data analysis. Four sample groups, i.e., pregnancies with euploid female fetuses, pregnancies with euploid male fetuses, pregnancies with T21 male fetuses and nonpregnant women, were analyzed, respectively. The statistics of fragment size of plasma DNA within DMRs are summarized in Table 8.2. Remarkably, the DNA fragments from highly methylated regions were statistically significantly longer than those from lowly methylated regions for all analyzed groups (Mann–Whitney rank-sum test, $P < 0.001$ for all groups).

Sample group	DMR[a]	Read no.	Median	25th percentile	75th percentile	P value [b]
maternal plasma (euploid female fetus)	high	515	163	141	174	<0.001
	low	3,589	159	131	173	
maternal plasma (euploid male fetus)	high	1,423	160	137	174	<0.001
	low	10,330	153	119	171	
maternal plasma (T21 male fetus)	high	815	158	131	172	<0.001
	low	6,172	149	109	170	
Plasma from nonpregnant adult females	high	1,017	163	142	174	<0.001
	low	7,845	152	85	171	

a: "high" indicates CGI regions with methylation percentage being ≥ 60%, whereas "low" indicates CGI regions with methylation percentage being < 60%.
b: based on Mann-Whitney U test

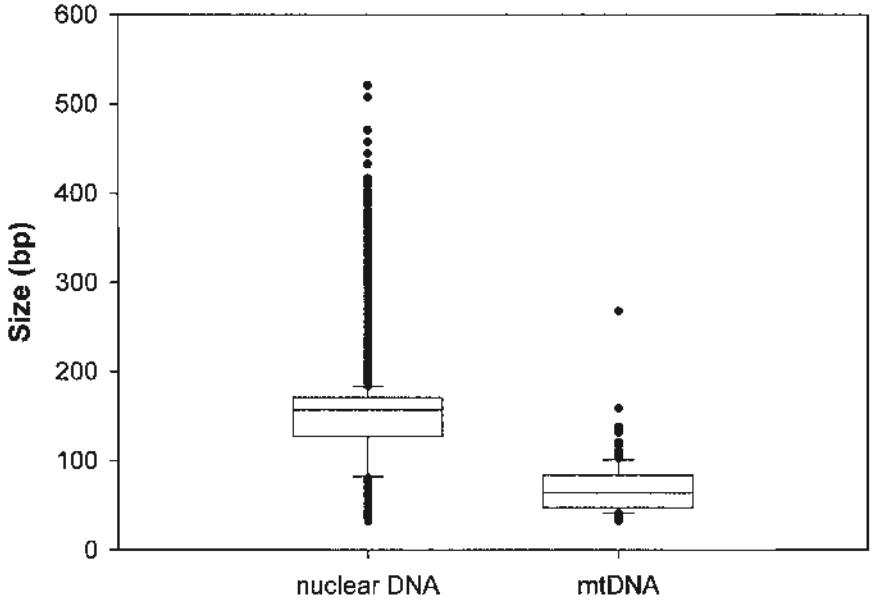
Table 8.2 Summary statistics of fragment size of plasma DNA within DMRs.

8.3.6 Fragment size of plasma DNA derived from the mitochondrial genome

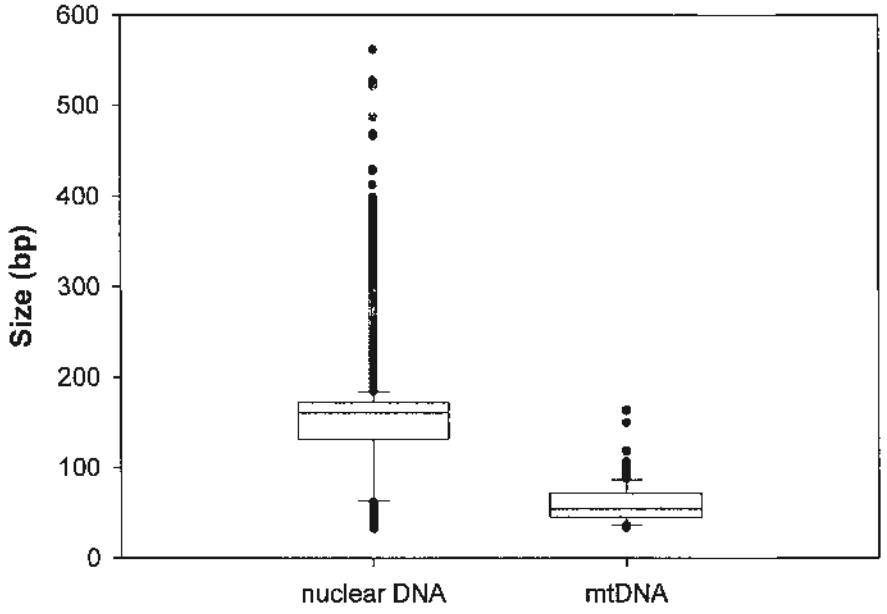
The circulating mitochondrial DNA (mtDNA) is detectable in plasma and serum (Chiu *et al.* 2003; Hibi *et al.* 2001; Okochi *et al.* 2002). In our sequencing data, a median (range) of 181 (42-341) accepted PE reads, which could be uniquely and perfectly aligned to the mitochondrial genome in the context of both nuclear and mitochondrial reference genomes, were obtained from plasma samples, representing 0.0132% (range, 0.0027%-0.0199%) of total accepted PE reads. The fragment size of plasma DNA from mitochondrial genome was analyzed and the comparison of fragment size between nuclear DNA (inclusive of autosomes and sex chromosomes) and mtDNA was performed. Surprisingly, the mtDNA fragments in plasma were quite short, being less than half of those of the nuclear DNA in plasma. Representative results in the plasma samples from a pregnant woman, a nonpregnant woman and an adult male are shown in Figure 8.5. The length of circulating mtDNA was significantly shorter than that of the circulating nuclear DNA for all plasma samples (Mann–Whitney rank-sum test, $P < 0.001$ for all cases). To ascertain that this observation is specific for plasma samples, I next checked the length of nuclear DNA and mtDNA in the placental tissue DNA. From our sequencing data, there were 2561, 2844, 1325 and 2813 accepted PE reads from mitochondrial genome for the 4 placental tissue DNA samples, respectively. The median of fragment size of mtDNA in the placental tissue samples was comparable to or even above that of nuclear DNA. For the two placental tissue samples from euploid fetuses (P4800 and P4801), there was no significant difference between mitochondrial and nuclear DNA (Mann–Whitney rank-sum test, $P = 0.802$ and 0.459, respectively); while for the other two placental tissue samples from T21 fetuses (P1519 and P4022), the length

of mtDNA fragments was significantly longer than that of the nuclear DNA (Mann–Whitney rank-sum test, $P < 0.001$ for both cases).

(A)


600
500
400
300
200
100
0
Size (bp)
nuclear DNA
mtDNA

(B)


600
500
400
300
200
100
0
Size (bp)
nuclear DNA
mtDNA

(C)

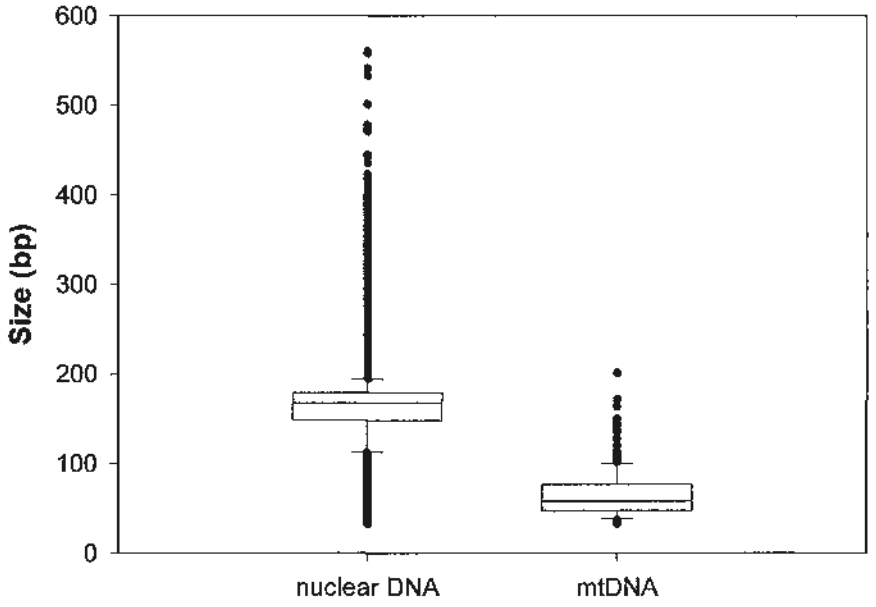


Figure 8.5 Fragment size of plasma DNA from the nuclear and mitochondrial genomes.

Box-plots show the difference in size distribution between nuclear DNA and mitochondrial DNA (indicated as "mtDNA") in the plasma samples from ***(A)*** a euploid male pregnancy (12), ***(B)*** an adult female (Female 01) and ***(C)*** an adult male (Male 03). The *lines inside the boxes* denote medians. The *boxes* mark the interval between the 25th and 75th percentiles. The *whiskers* denote the interval between the 10th and 90th percentile. •indicate data points outside the 10th and 90th percentiles.

8.3.7 Fragment size of plasma DNA for prenatal diagnosis

As described above, size distributions of fragments from chrX and 21 could be represented by a size ranking when compared with other chromosome. Figure 8.6A shows that the size rankings of chr21 DNA fragments in plasma for the five T21 pregnancies (ranging from 19 to 22) are lower than those for the euploid pregnancies (ranging from 10 to 19). Since the DNA fragments released by a fetus are shorter than those from the mother, a trisomy 21 fetus will reduce the overall measured size of fragments derived from chr21 in maternal plasma by releasing three doses of chr21 per fetal cell. Conversely, a euploid fetus would only be able to release two doses of chr21 per fetal cell. On the chrX side, it ranked 1 in all male pregnancies but 4 in all female pregnancies (Figure 8.6B). Similarly, the lowered size rankings of chrX DNA molecules for the female pregnancies compared with the male pregnancies could be once again traced back to the fact that the DNA fragments released by a fetus are shorter than those from the mother. Thus, in contrast to a male fetus that would only be able to release a single dose of chrX, a female fetus will reduce the overall measured size of fragments derived from chrX in maternal plasma by releasing a double dose of chrX. These observations indicated that the difference in size distribution would be of diagnostic value in the noninvasive prenatal diagnosis of fetal triosmy 21 and fetal sex determination.

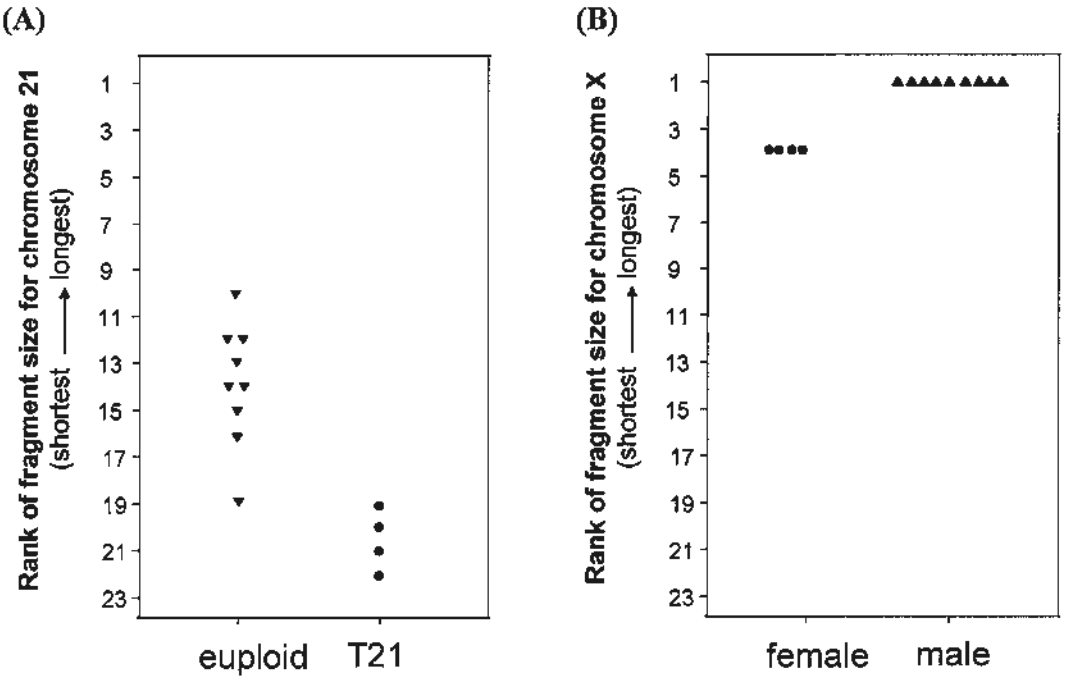


Figure 8.6 Rank of fragment size for chromosome 21 and X for prenatal diagnosis.

(A) Rank of fragment size for chromosome 21 in the maternal plasma samples from euploid and trisomy 21 (indicated as "T21") pregnancies. ***(B)*** Rank of fragment size for chromosome X in maternal plasma samples from female and male pregnancies.

8.4 Discussion

In this chapter, the size characteristics in pregnant women and healthy individuals are studied. The high-resolution size profile provided by massively parallel PE sequencing enables us to reveal the kinetics of both maternal and fetal DNA molecules in maternal plasma and helps us to better understand the underlying biological events of plasma DNA.

Based on QPCR, Chan *et al.* have demonstrated that the plasma DNA molecules are longer in pregnant women than in nonpregnant women (Chan *et al.* 2004). Our observation is supportive of this finding at single molecule resolution and in a locus-independent manner. Intriguingly, the significant size differences of both fetal and total DNA fragments in maternal plasma between early and late pregnancies were observed, implying that there would be dynamic changes in fragment size of circulating fetal and total DNA molecules at different gestational ages. In combination with the previous finding that both fetal and total plasma DNA (predominantly maternally-contributed) concentrations increased in late pregnancy, especially during the last few weeks of normal pregnancies (Chan *et al.* 2003b; Lo *et al.* 1998a), it is possible that the high turnover of fetomaternal interface would lead to much more cell-free DNA into maternal circulation, thus overloading the capacity of clearance mechanisms. The suggested mechanisms for circulating DNA clearance include plasma nucleases and hepatic and renal clearance (Lo 2001). It has been demonstrated that once the liver saturated, it cannot accommodate more DNA (Liu *et al.* 2007). The continuously increased DNA would therefore further overburden other clearance mechanisms, e.g., plasma nucleases which can degrade DNA gradually, subsequently resulting in longer plasma DNA fragments. There is another possibility that other physiologic changes during pregnancy would enhance the release of DNA

of larger molecular sizes (Chan *et al.* 2004). To further explore the biological basis for these observations, future work could be conducted by serially collecting maternal plasma samples during pregnancy and documenting the dynamic changes in both fetal and maternal DNA as pregnancy advances.

The size ranking of the plasma DNA molecules originating from different chromosomes bore some relationship to the chromosomal GC content (Figure 8.4). It is interesting that the placental tissues fragmented by nebulization also showed a relationship between DNA GC content and molecular size but in a reversed direction. Though the reason for the observation is uncertain, one possibility is that the GC content of a DNA sequence has opposite effects on the natural fragmentation processes of plasma DNA compared with artificial processes such as nebulization. Besides, the relationship between fragment size and chromosomal GC content may partially explain the observed GC bias in quantitative analyses of chromosomal representation for both tissue DNA and plasma DNA (Chapter 7). Assuming that the GC-rich chromosome (for plasma DNA/placental tissue DNA) had relatively shorter/longer fragments, one copy of such chromosome would produce more/less DNA fragments in plasma, causing a corresponding increase/decrease of sequence reads for that chromosome for tissue DNA and plasma, respectively. However, as there is only 1 bp to 3 bp difference in the median of fragment size among chromosomes, such perturbation would contribute a little, if any, to the observed GC bias for quantitative analysis. The GC bias is more likely to stem from the internal bias of Illumina sequencing system as discussed in Chapter 5.

Nucleosomal complexes consist of 2 copies each of the core histones H2A, H2B, H3, and H4 with DNA on the outside and are joined by a stretch of free DNA termed "linker DNA" (Luger 2003). The 166-bp size feature is reminiscent of the

monochromatosome, which consists of nucleosomal DNA wrapping around the histone core (146 bp) and a linker unit (20 bp) (Jiang *et al.* 2009). Increasing lines of evidence indicate that the plasma DNA is mainly derived from cell apoptosis, during which DNA is degraded into nucleosomal units in most cases (Wyllie 1980). Hence, we reasoned that plasma DNA was released during apoptosis of the respective cells, giving rise to the abundance of 166-bp fragments that resembles the monochromatosome. In favor of this hypothesis is the high concordance between the direct immunologic detection of nucleosomal DNA and the quantification of *HBB* gene sequences in plasma samples (Holdenrieder *et al.* 2005).

The stability of plasma DNA might be attributable to the arrangement of DNA in nucleosomes, which shields them from proteolytic digestion in blood (Holdenrieder *et al.* 2005). DNA methylation and histone modification are believed to be interdependent processes that can regulate higher-order chromatin structures and the accessibility of chromatin to various factors (Bartova *et al.* 2008; Li 2002). Previous works suggest that the change in chromatin structure may allow easy accessibility of nuclear DNase to chromosomal DNA and may be one of the molecular mechanisms of internucleosomal DNA fragmentation (Enomoto *et al.* 2002; Enomoto *et al.* 2003). The inter-chromosomal variation in fragment size of plasma DNA associated with chromosomal GC content prompted us to investigate the hypothesis that the epigenetic modification that regulates chromatin structures may influence the fragment size of plasma DNA. Interestingly, a significant difference in fragment size between plasma DNA within lowly and highly methylated regions was observed, indicating that DNA methylation may be associated with plasma DNA fragmentation. An additional piece of evidence is from the difference in size rankings between chrX DNA fragments in females and males, because it has been known that

the X-inactivation process in female converts an X chromosome from active euchromatin into transcriptionally silent and highly condensed heterochromatin through a series of events that include DNA methylation and histone modification (Li 2002). Moreover, *Alu* repeat sequences, which are highly methylated and correlate with a closed chromatin structure, have been reported with higher abundance in plasma and serum compared to unique genes (Beck *et al.* 2009; Stroun *et al.* 2001), providing a supportive piece of evidence of our hypothesis.

However, how exactly the epigenetic modification influences the fragments size of plasma DNA and whether such effect is direct or indirect are unclear and require further investigation. Regarding the liberation mechanism of DNA into circulation, one possible mechanism is that the condensed chromatin structures corresponding to the highly methylated regions would be less vulnerable to degradation by nuclear DNase, resulting in relatively longer fragments released into circulation. Once released, the DNA fragments associated with a highly condensed chromatin configuration may persist as more compacted nucleosomes in the circulation, thus maybe protecting the DNA from degradation of plasma nuclease (Amoura *et al.* 1997; Deligezer *et al.* 2003; Holdenrieder *et al.* 2001).

Strikingly, the circulating mtDNA is much shorter than the circulating nuclear DNA in plasma. The sharp shortening of mtDNA in plasma could be attributable to its unique structural features. MtDNA is a double-stranded, circular molecule of 16 5 kb in length (Wallace 1999), with up to several thousand copies of this genome found in a mammalian cell (Cavelier *et al.* 2000; Satoh *et al.* 1991). It is considered to be a more susceptible target for various damaging agents than nuclear DNA (Mandavilli *et al.* 2002; May *et al.* 2000), which has been supposed to result from the absence of histones (Caron *et al.* 1979; DeFrancesco *et al.* 1981; Guliaeva *et al.* 2006).

Moreover, previous work demonstrates mtDNA is methylated at a very low level (Maekawa *et al.* 2004; Shmookler Reis *et al.* 1983). Therefore, unlike circulating nuclear DNA which is probably complexed with histone, circulating mtDNA may be much more vulnerable to the clearance system of plasma DNA, resulting in shorter fragment size distribution.

From our result, distinctive size distribution patterns of plasma DNA from females and males were observed, with shorter fragments (< 100 bp) in the plasma samples from females than males. Further study is needed to confirm this finding. However, to some extent, this observation could also be linked to the differential methylation status between females and males, as El-Maarri *et al.* have demonstrated the gender specific differences in levels of DNA methylation, i.e., a tendency toward higher methylation levels in males, by analyzing the selected loci from human total blood (El-Maarri *et al.* 2007).

Apart from the biological implications, the size distribution of target chromosomes may be useful for the noninvasive prenatal diagnosis of fetal chromosomal aneuploidies. The clear separation of size rankings for chrX between female and male fetuses shows its usefulness for fetal sex determination, while the difference in size ranking for chr21 between euploid and T21 pregnancies indicates an alternative way for fetal trisomy 21 detection. Large sample size is required to testify the diagnostic performance of this analytical method in the future.

In summary, the implications from the size profiles afforded by PE sequencing can improve our understanding of biological events of circulating DNA in plasma. PE sequencing, combining with other *in vivo* or *in vitro* studies, will provide us with a new avenue for studying the underlying mechanisms of plasma DNA.

CHAPTER 9: MASSIVELY PARALLEL PAIRED-END SEQUENCING OF PLASMA DNA IN HEMATOPOIETIC STEM CELL TRANSPLANT (HSCT) RECIPIENTS

9.1 Introduction

In a hematopoietic stem cell transplant (HSCT) model, the hematopoietic system of the transplant recipient is predominantly of donor origin, while the nonhematopoietic tissues are recipient in origin. The sex-mismatched HSCT model has been employed for the investigation of the relevant biological indications, such as the demonstration of the origin of plasma DNA (Lui *et al.* 2002), the validation of potential circulating mRNA marker (Chan *et al.* 2007) and the study of the occurrence of nonhost DNA in urine (Hung *et al.* 2009).

Lui *et al.* used the sex-mismatched HSCT model to study the relative contribution of hematopoietic and nonhematopoietic cells to circulating DNA by QPCR and demonstrated that the DNA in plasma and serum was predominantly hematopoietic in origin, with a median of 59.5% of the DNA in the plasma of HSCT recipients (Lui *et al.* 2002). Apart from the quantitative contribution, however, the size information of the DNA molecules from the two cellular sources is not currently available. Massively parallel paired-end sequencing of plasma DNA obtained from HSCT recipients would provide an overall picture of chromosomal distribution and a high-resolution size profile of the circulating DNA in these patients, sheding some lights on the characterization of the plasma DNA originating from hematopoietic and nonhematopoietic cells.

In this study, I attempt to investigate both of the chromosomal representation and the size distribution of the plasma DNA obtained from patients after transplantation by PE sequencing on the Illumina sequencing platform. Because in sex-mismatched HSCT patients either the donor or recipient is male, the sequence reads aligned to chromosome Y are markers of male DNA and would allow us to investigate the proportion and size characteristics of DNA fragments from donor or recipient origin in the patients' plasma. Two main aspects of the sequencing data are mined. One is to quantify the contributions of hematopoietic and nonhematopoietic cells to the circulating DNA by analyzing the proportions of accepted PE reads for chromosome X (%chrX) and Y (%chrY) and the other is to characterize the fragment size of plasma DNA from two cellular sources by comparing the size distributions of plasma DNA from the Y and non-Y chromosomes in the sex-mismatched cases. Additionally, the apparently distinctive size patterns of plasma DNA for adult females and males observed in Chapter 8 are examined in the sex-matched HSCT recipients.

9.2 Methods

9.2.1 Subjects

Prof Emily Hung kindly helped to recruit 8 HSCT patients from the Bone Marrow Transplant Clinic of the Department of Paediatrics, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong. All patients were in remission with respect to their primary conditions. Informed consent was obtained from patients or their parents.

9.2.2 Sample preparation

Plasma was harvested from blood samples as described in Chapter 3.1.2. DNA was extracted from maternal plasma according to procedures described in Chapter 3.2.1. The extracted DNA was subjected to QPCR as described in Chapter 3.3.

9.2.3 Fluorescence *in situ* hybridization and DNA short tandem repeat analyses for peripheral blood chimerism

In fluorescence *in situ* hybridization (FISH) analysis for peripheral blood chimerism, peripheral blood samples were subjected to density gradient centrifugation in Ficoll-Hypaque of specific density 1.077. The cell pellet was re-suspended in phosphate buffered saline supplemented with 1% bovine serum albumin. One hundred microliters of the cell suspension of approximately 1×10^5 mononuclear cells were cytospun onto glass slides and stored at -80°C until analysis. Cytospun slides were fixed in Carnoy's fixative (absolute methanol/glacial acetic acid, 3:1) at room temperature for 2 min, air dried and then hybridized with X and Y probes in an automated denaturation/hybridization system (HYBrite, Vysis, Downers Grove, IL, USA). The X and Y probes were commercially available and were directly labeled with SpectrumGreen by the manufacturer (Vysis). The X probe hybridized to the alpha satellite repeats at *DXZ1* (Xp11.1 - q11.1) and the Y probe hybridized to the satellite III at *DYZ1* (Yq12). Male and female cells were run in parallel to control for the hybridization efficiency. Hybridization signals of at least 600 interphase cells were scored separately by two examiners.

For DNA short tandem repeats (*STR*), DNA was extracted from peripheral blood samples of pre-transplant recipients and donors and amplified with commercially available *STR* primers (Applied Biosystems, Foster City, CA, USA). The products were capillary electrophoresed in an ABI PRISM 310 DNA sequencer (Applied

Biosystems). Results were analyzed using the GeneScan 2.1 software (Applied Biosystems). Alleles which could distinguish the donor from the recipient were selected as informative alleles for further analysis. Artificial mixtures of pre-HSCT recipient and donor DNA in different ratios were used to establish a standard curve for subsequent quantification. To study the chimerism status, post-HSCT recipient DNA was amplified by *STR* primers as described. The percentage peak areas of the informative alleles were then extrapolated from the standard curve and reported as the percentage of donor-derived cells in peripheral blood.

9.2.4 Massively parallel paired-end sequencing of plasma DNA

The massively parallel paired-end sequencing of plasma DNA was performed on the Illumina GA II system as described in Chapter 3.5.

9.2.5 Sequence and size analyses

The sequence alignment and selection were done based on the same criteria described in Chapter 7.2.4. The size of each sequenced plasma DNA molecule was deduced as described in Chapter 7.2.6.

9.2.6 Calculation of the percentage of male DNA

SRY/HBB assay

The calculation based on the *SRY/HBB* assay has been described in the previous study (Lui *et al.* 2002). Because the *SRY* gene is found in all nucleated cells of males only, whereas the *HBB* gene is present in all nucleated cells of both males and females (Lo *et al.* 1998a), the percentage of male DNA (P) in a particular plasma sample could be calculated using the following equation: P = *SRY/HBB* × 100%,

where *SRY* and *HBB* denote the quantities of the *SRY* and *HBB* sequences measured by the respective QPCR assays described in Chapter 3.3.

Genomic representation of chromosome X and Y

The number of accepted PE reads aligned to each human chromosome was expressed as a proportion of total accepted PE reads obtained for the sample. The proportions of chromosome X (%chrX) and chromosome Y (%chrY) were calculated for each sequenced plasma sample.

Our previously published data (Chiu *et al.* 2008) and the data in Chapter 7 showed that a small number of sequences would be falsely aligned to chromosome Y, even for female cases. Hence, the %chrY value in the plasma DNA from the post-HSCT patient is a composite of the amount of chrY sequences contributed by the male side (containing 100% male DNA) and the female side (false alignment). In the cumulative PE sequencing data, the means of %chrY were 0.0051% for the 5 female cases (inclusive of 3 healthy females and 2 female patients receiving HSCT from female donors) and 0.147% for the 6 male cases (inclusive of 4 healthy males and 2 male patients receiving HSCT from male donors). Thus, the percentage of male DNA (P) can be derived from the equation: $\%chrY = 0.147P + 0.0051(1-P)$.

On the chrX side, the %chrX value in the plasma DNA from the post-HSCT patient is a composite of the amount of chrX sequences contributed by the male side (only one dose of chrX) and the female side (two doses of chrX). The means of %chrX were 4.024% for the 5 female cases mentioned above (containing 100% female DNA) and 2.035% for the 6 male cases mentioned above (containing 100% male DNA). Thus, the percentage of male DNA (P) can be deduced from the equation: $\%chrX = 2.035P + 4.024 (1-P)$.

9.3 Results

9.3.1 Massively parallel paired-end sequencing of plasma DNA in HSCT recipients

Plasma DNA from 8 post-HSCT patients was subjected to PE sequencing, 4 of which were sex-mismatched HSCT patients. The characteristics of the HSCT recipients involved in the current study are shown in Table 9.1. The 4 sex-matched HSCT recipients and the 4 sex-mismatched HSCT recipients were in complete remission with respect to their hematologic conditions. The analysis of chimerism status, i.e., the presence of lymphohematopoietic cells of nonhost origin measured with FISH and DNA *STR* analyses (Antin *et al.* 2001) as described above, revealed that all patients had > 99% donor lymphohematopoietic cells in the peripheral blood, fulfiling the criterion for full chimerism with complete lymphohematopoietic replacement (Antin *et al.* 2001).

A median of 1.41×10^6 reads/sample (range, $1.23\text{-}1.52 \times 10^6$), representing 19% (range, 17.9%-22.4%) of the raw reads, could be retrieved as accepted PE reads meeting the criteria described in Chapter 7.2.4. I compared the %GR of each chromosome observed for each case to that expected for the respective repeat-masked diploid genomes. Given that the majority of plasma DNA molecules stem from hematopoietic cells which had been almost entirely converted to the donor hematopoietic cells in these HSCT recipients (Lui *et al.* 2002), two sets of reference genomes, namely, the female repeat-masked diploid genome and the male repeat-masked diploid genome, were used as the respective references for the cases involving female and male donors. The chromosomal representations for all 8 cases are shown in Figure 9.1A. Basically, the obtained %GR of each chromosome was

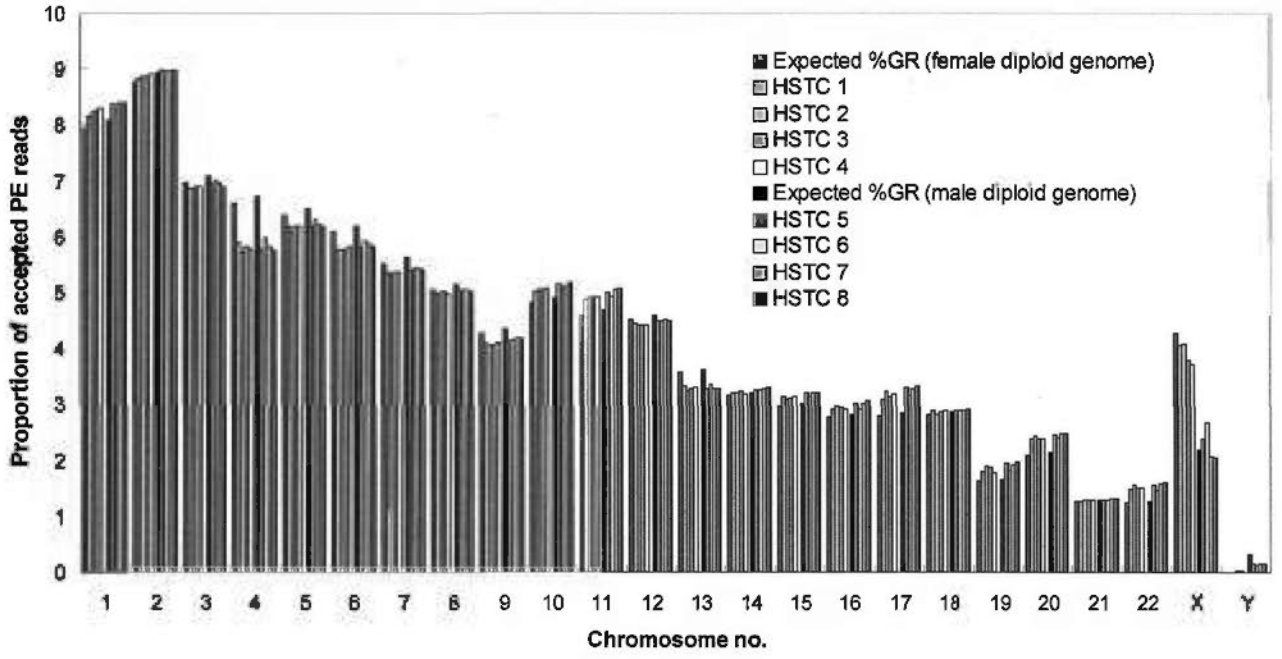
close to the expected value. Similar to the previous data, there was a GC bias observed for the current sequencing data (Figure 9.1B and C). The average ratio of observed to expected %GR strongly correlated with chromosomal GC content for autosomes in both groups (Pearson Product Moment Correlation, $r = 0.904$, $P = 8.4 \times 10^{-9}$ for the female donor group; $r = 0.925$, $P = 7.36 \times 10^{-10}$ for the male donor group).

Case no.	Sex of Donor	Sex of Subject	Diagnosis	Chimerism (%)	Accepted PE reads	Reads of chrX	Reads of chrY
HSCT 1	F	F	SAA	100	1,233,069	50,002	61
HSCT 2	F	F	SAA	100	1,311,485	53,372	74
HSCT 3	F	M	WAS	99.3	1,418,655	53,800	364
HSCT 4	F	M	ALL	99.8	1,368,548	50,869	422
HSCT 5	M	F	ALL	99.8	1,497,673	35,517	2,002
HSCT 6	M	F	ALL	99.8	1,414,200	37,742	1,621
HSCT 7	M	M	βTM	100	1,523,686	31,260	2,245
HSCT 8	M	M	SAA	100	1,399,930	28,694	2,111

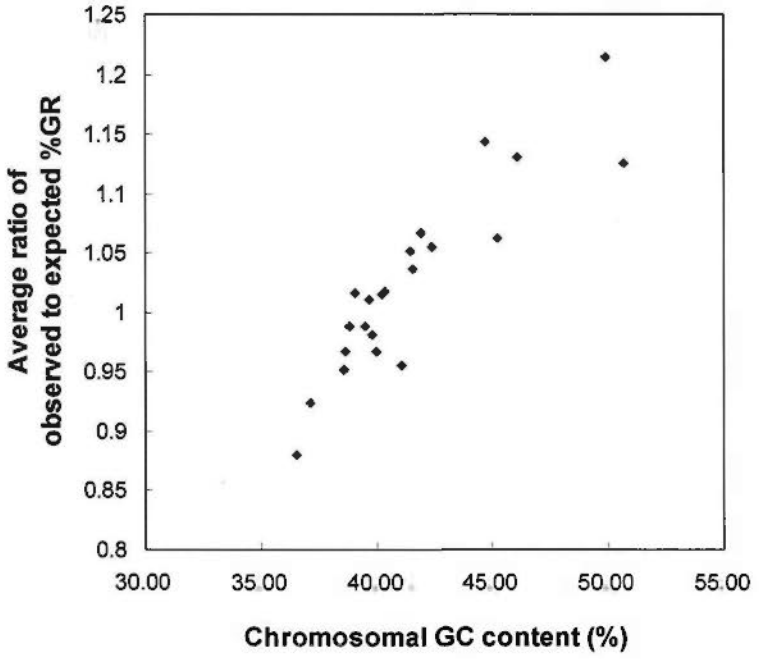
SAA, Severe aplastic anaemia; WAS, Wiskott-Aldrich Syndrome; ALL, acute lymphoblastic

Table 9.1 Clinical details and sequencing counts of the plasma samples from HSCT recipients.

(A)


Proportion of accepted PE reads
Chromosome no.
Expected %GR (female diploid genome)
HSTC 1
HSTC 2
HSTC 3
HSTC 4
Expected %GR (male diploid genome)
HSTC 5
HSTC 6
HSTC 7
HSTC 8

(B)


Average ratio of observed to expected %GR
Chromosomal GC content (%)

(C)

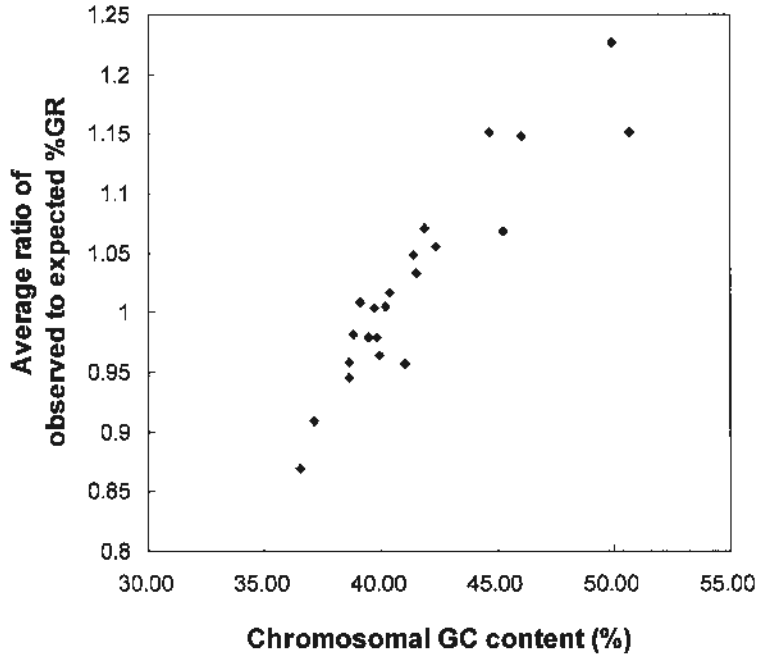


Figure 9.1 Distribution of PE reads among chromosomes for the plasma samples from the HSCT recipients.

(A) Bar chart of proportion of accepted PE reads per chromosome for 4 sex-matched HSCT patients and 4 sex-mismatched HSCT patients. The percentage of genomic representation of each chromosome as expected for the repeat-masked reference diploid female (red bars) and male (black bars) genomes is plotted for comparison. ***(B)*** Correlation between the average measured ratios against chromosomal GC content for autosomes for the patients receiving HSCT from female donors. ***(C)*** Correlation between the average measured ratios against chromosomal GC content for autosomes for the patients receiving HSCT from male donors.

9.3.2 Quantification of hematopoietic contribution by the *SRY/HBB* assays

Prior to sequencing, the *SRY/HBB* QPCR assays were performed for evaluating the percentage of male DNA. In the two female sex-mismatched recipients, the percentages of male DNA (i.e., male donor-derived) were 93.04% and 70.49%, respectively. In the two male sex-mismatched recipients, the percentages of male DNA (i.e., male recipient-derived) were only 7.02% and 12.63%, respectively; thus, the proportions of donor-derived DNA were 92.98% and 87.37%, respectively. Taken together, these 4 sex-mismatched recipients showed very high contributions of donor-derived DNA with a median of hematopoietic contribution of 90.18%, in keeping with the data presented in previous study (Lui *et al.* 2002).

9.3.3 Quantification of hematopoietic contribution by %chrX and %chrY

The number of accepted PE reads aligned to chrX and chrY are shown in Table 9.1. The %chrX and %chrY for the post-HSCT patients and the healthy individuals are presented in Figure 9.2. There was an overrepresentation of chrX in the male patients receiving HSCT from female donors compared with adult males whereas an underrepresentation of chrX in the female patients receiving HSCT from male donors compared with adult females (Figure 9.2A). Opposite status was observed for %chrY (Figure 9.2A). On the other hand, for the sex-matched cases, the %chrX and %chrY values were comparable with the healthy females or males. The fluctuation in %chrX or %chrY values between the sex-mismatched recipients and the healthy individuals is informative for deducing the contribution of hematopoietic and nonhematopoietic cells respectively as describe above. The results are shown in Figure 9.2B. The quantification results from sequencing data correlated well with those determined by the *SRY/HBB* QPCR assays (Pearson Product Moment Correlation, r = 0.992, P =

1.08×10^{-6} for the *SRY/HBB* assay and %chrX-based calculation; $r = 0.995$, $P = 3.13 \times 10^{-7}$ for the *SRY/HBB* assay and %chrY-based calculation).

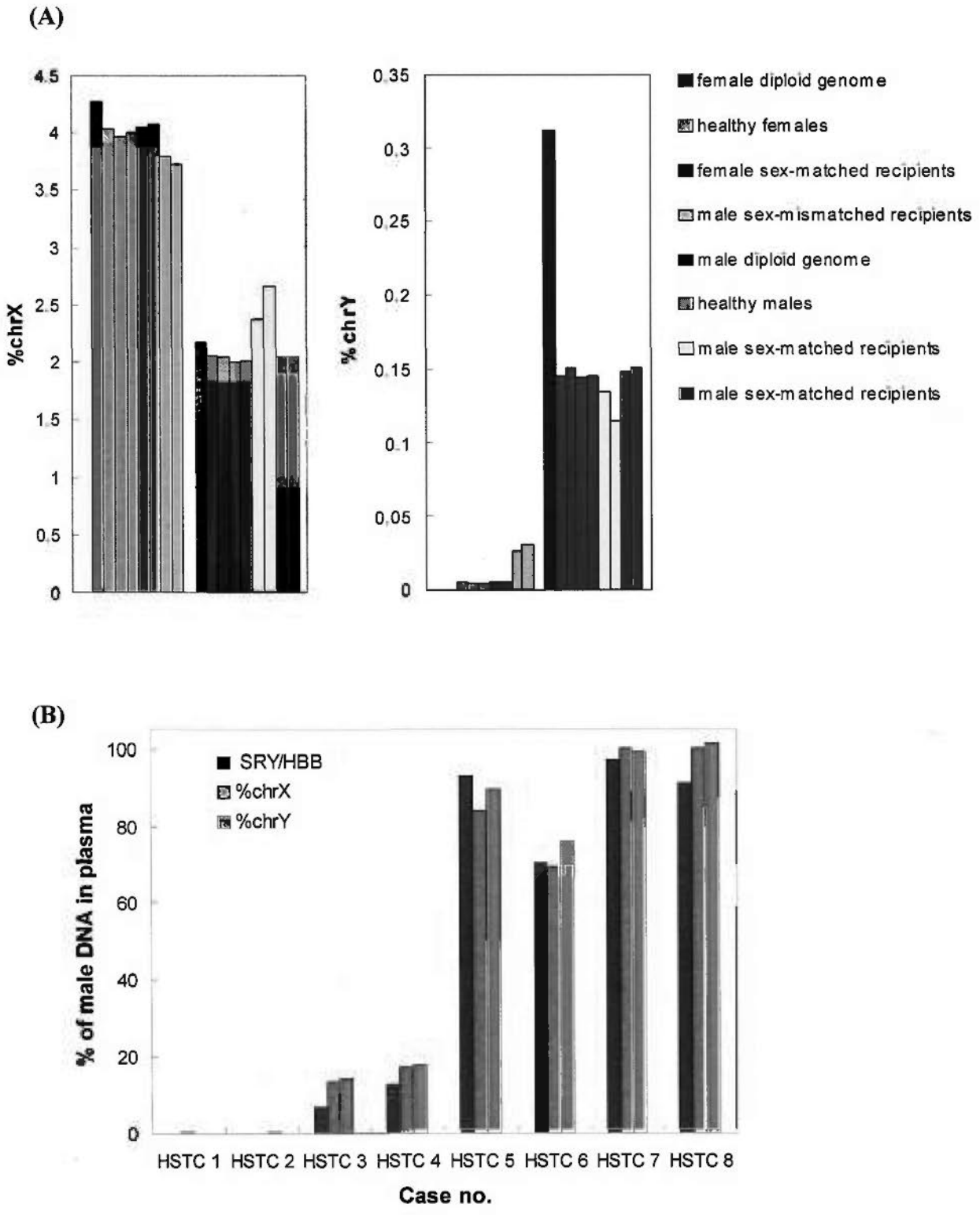


Figure 9.2 Percentage of male DNA in the post-HSCT patients.

(A) %chrX and %chrY in the plasma samples of the post-HSCT patients and healthy individuals. ***(B)*** The percentage of male DNA calculated by the readouts from the *SRY/HBB* assays and %chrX and %chrY values from PE sequencing, respectively.

9.3.4 Size distribution of plasma DNA in HSCT recipients

Size distribution of plasma DNA in male sex-mismatched HSCT recipients

In the male patients receiving HSCT from female donors (patients HSCT 3 and HSCT 4), both the nonhematopoietic cells and the remaining hematopoietic cells of the male recipients could account for the chrY fragments. However, in these male patients, the hematopoietic system was close to 100% converted into female after transplantation (Table 9.1). Therefore, the free chrY fragments in the plasma of these cases mainly originated from the nonhematopoietic cells of the male recipients, serving as an indicator of the free plasma DNA originating from nonhematopoietic systems. On the other hand, the DNA fragments from non-Y chromosomes were contributed by the female donor and the male recipient. However, in view of the major contribution of donor-derived DNA in plasma (86.58% and 82.83% for HSCT 3 and HSCT 4, respectively, estimated by %chrX), the non-Y fragments would represent the majority of DNA molecules originating from hematopoietic systems. Taking both sides together, it is reasonable to infer the difference in fragment size between hematopoietically- and nonhematopoietically-derived DNA molecules from the comparison of size distributions for Y and non-Y chromosomes. The medians (interquartile range) of length of DNA fragments from Y and non-Y chromosomes in the plasma of patient HSCT 3 were 148 bp (113 bp-168 bp) and 158 bp (126 bp-171 bp), respectively (Figure 9.3A) while those in the plasma of patient HSCT 4 were 159 bp (127 bp-172 bp) and 165 bp (148 bp-175 bp), respectively (Figure 9.3B). The DNA fragments from non-Y chromosomes were significantly longer than chrY fragments in each case (Mann-Whitney rank-sum test, $P < 0.001$ for both cases), revealing that the hematopoietically-derived DNA molecules were significantly longer than nonhematopoietically-derived ones.

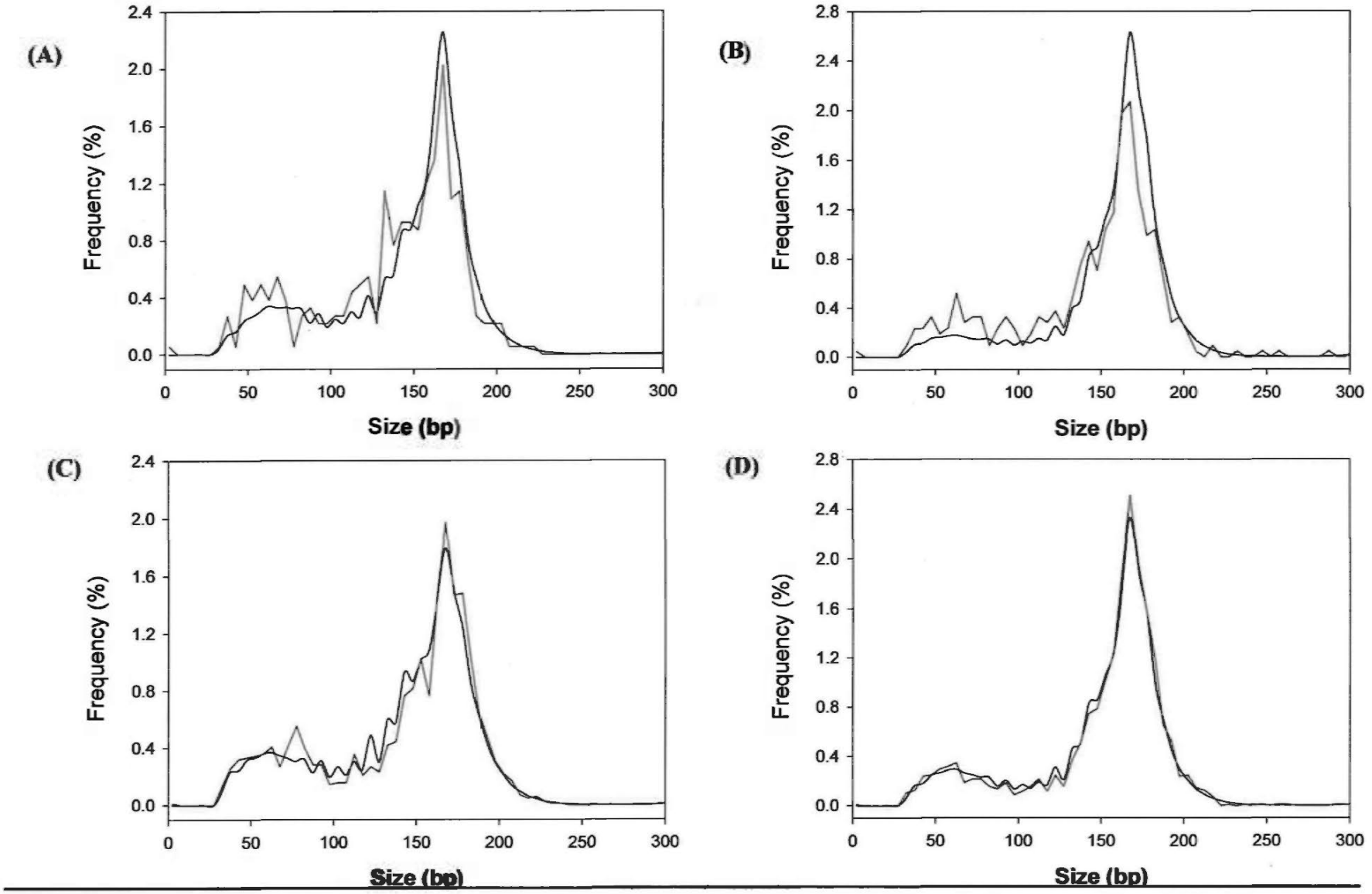
Size distribution of plasma DNA in female sex-mismatched HSCT recipients

For the cases involving female recipients and male donors (patients HSCT 5 and HSCT 6), the chrY fragments were derived from the hematopoietic cells of the male donors, thus representing the plasma DNA originating from hematopoietic system. In these plasma samples, the DNA fragments from non-Y chromosomes were also predominantly from the hematopoietic cells of the male donors, due to the major contribution of donor-derived DNA in plasma (84.05% and 69.27% for HSCT 5 and HSCT 6, respectively, estimated by %chrX). The medians (interquartile range) of length of DNA fragments from Y and non-Y chromosomes in the plasma of patient HSCT 5 were 164 bp (141 bp-175 bp) and 162 bp (138 bp-174 bp), respectively (Figure 9.3C) while those in the plasma of patient HSCT 6 were 159 bp (112 bp-174 bp) and 156 bp (119 bp-172 bp), respectively (Figure 9.3D). There was no significant difference between the size distributions for the Y and non-Y chromosomes (Mann-Whitney rank-sum test, P = 0.063 and 0.189 for HSCT 5 and HSCT 6, respectively).

Size distribution of plasma DNA in male sex-matched HSCT recipients

For the male sex-matched recipients (patients HSCT 7 and HSCT 8), both chrY fragments and DNA fragments from non-Y chromosomes proportionally consisted of the male donor-derived DNA molecules and the male recipient-derived DNA molecules; hence, the size distribution of chrY fragments should fall into that of DNA fragments from non-Y chromosomes. The medians (interquartile range) of length of DNA fragments from Y and non-Y chromosomes for patient HSCT 7 were 162 bp (137 bp-175 bp) and 162 bp (137 bp-174 bp), respectively (Figure 9.3E) while those for patient HSCT 8 were 162 bp (137 bp-174 bp) and 162 bp (138 bp-174 bp), respectively (Figure 9.3F). As expected, the results showed no statistical

difference between the size distributions for the Y and non-Y chromosomes (Mann-Whitney rank-sum test, P =0.290 and 0.873 for HSCT 7 and HSCT 8, respectively).


(A)
Frequency (%)
Size (bp)
(B)
Frequency (%)
Size (bp)
(C)
Frequency (%)
Size (bp)
(D)
Frequency (%)
Size (bp)

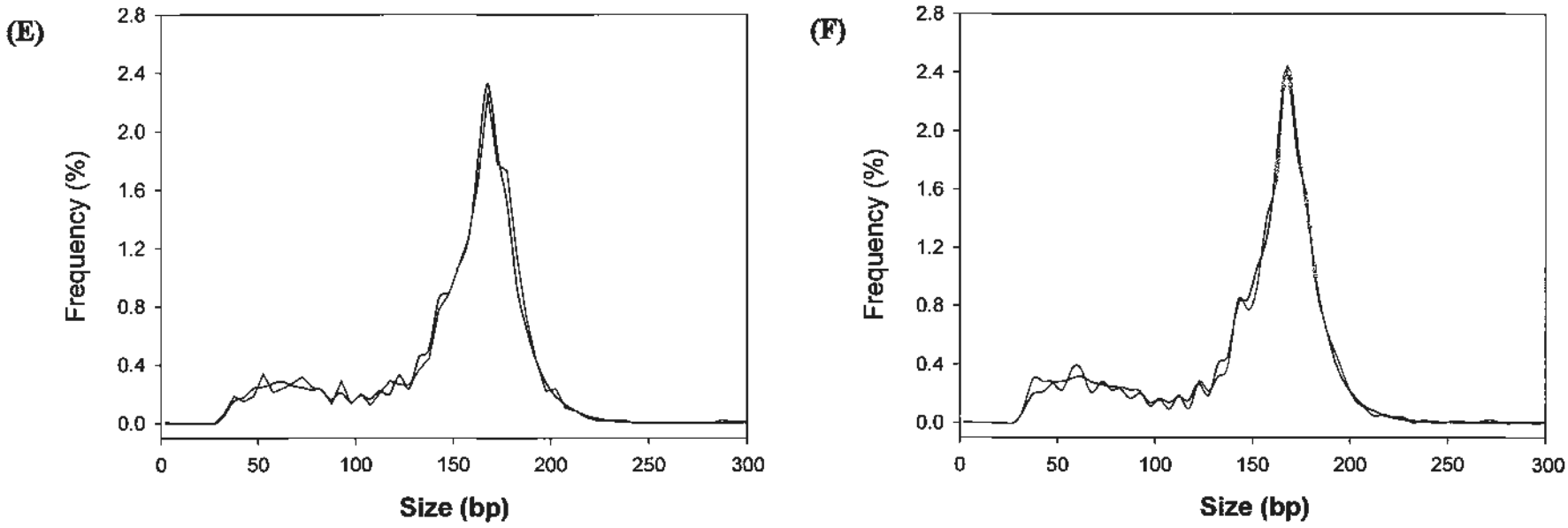


Figure 9.3 Size distributions of DNA fragments derived from the Y and non-Y chromosomes in plasma DNA from the post-HSCT patients.

Histograms (at 5-bp resolution) show the size distributions of accepted PE reads aligned to Y (red line) and non-Y (black line) chromosomes in the plasma from ***(A)*** patient HSCT 3, ***(B)*** patient HSCT 4, ***(C)*** patient HSCT 5, ***(D)*** patient HSCT 6, ***(E)*** patient HSCT 7 and ***(F)*** patient HSCT 8. Only the plasma DNA molecules with fragment size ≦ 300 bp are shown.

9.3.5 Specific size distribution pattern among females and males

In the last chapter, there seemed to be a specific pattern of size distribution of plasma DNA for healthy adult females and males, i.e., there was a bimodal pattern appearing for the plasma DNA from adult females, but no such pattern was observed for the plasma DNA from adult males. To determine whether this phenomenon would exist in the plasma of the HSCT patients, I compared the overall size distributions of plasma DNA from the sex-matched recipients with those from healthy individuals. Plasma DNA from the female sex-matched recipient showed the similar size distribution pattern to those from healthy adult females. In particular, the proportions of DNA fragments ≤ 100 bp in the two female sex-matched recipients were 19.03% and 17.92%, respectively (Figure 9.4A and 9.4B), being concordant with those in the healthy adult females (median, 19.30%). On the contrary, there was a discrepancy in size distribution of plasma DNA fragments between the male sex-matched recipients and the healthy adult males. The proportions of DNA fragments ≤ 100 bp in the two male sex-matched recipients were 14.89% and 16.15%, respectively (Figure 9.4C and 9.4D), in contrast to those in the healthy adult males (median, 7.79%).

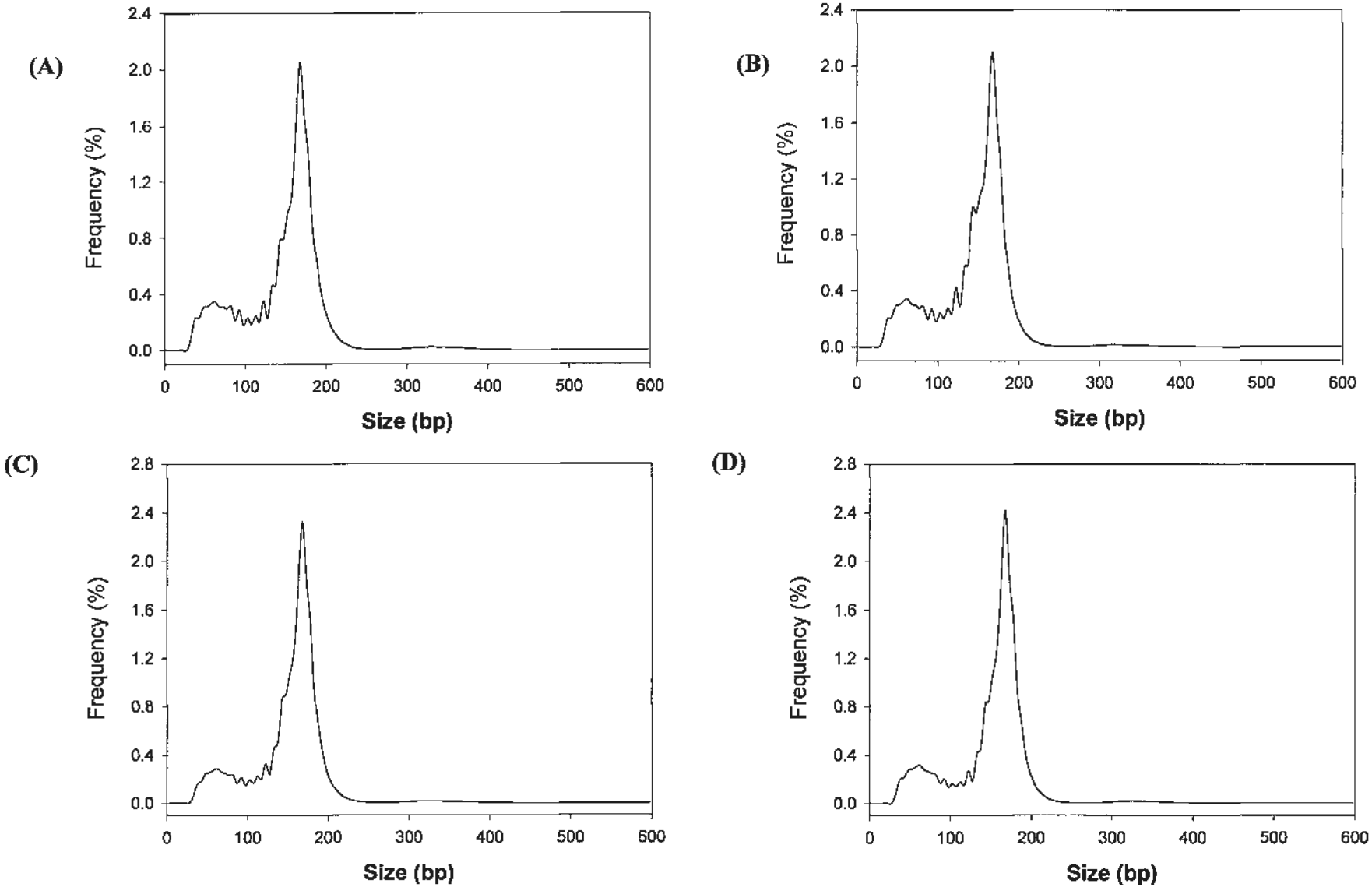

(A)
(B)
(C)
(D)
Frequency (%)
Size (bp)
0
100
200
300
400
500
600
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8

Figure 9.4 Overall size profiles of DNA fragments in the plasma of the sex-matched recipients.

Histograms (at 5-bp resolution) show the size distributions of DNA fragments in the plasma from ***(A)*** patient HSCT 1, ***(B)*** patient HSCT 2, ***(C)*** patient HSCT 7 and ***(D)*** patient HSCT 8.

9.4 Discussion

In this study, massively parallel PE sequencing approach was applied to sequence the plasma DNA obtained from patients receiving HSCT in an effort to characterize the nature of plasma DNA with respect to the cellular origin and the fragment size. The results presented here substantiate that most of the plasma DNA is hematopoietically-derived in these transplant recipients and demonstrate that the plasma DNA molecules of hematopoietic origin are longer than those of nonhematopoietic origin.

The quantitative contributions of two cellular sources calculated from %chrX and %chrY are slightly different from those calculated from the *SRY/HBB* assays. As discussed in Chapter 7, this may be caused by the use of different analytical platforms. The former one quantifies the molecular number in a locus-independent manner whereas the latter one only targets the specific gene loci, probably leading to a less precise measurement. Instead of the whole blood samples (Antin *et al.* 2001; Lo *et al.* 1995; Lo *et al.* 1993), plasma was proposed to serve as an alternative sample type for the determination of HSCT chimerism (Lui *et al.* 2002). The quantification approach described here could be used to determine HSCT chimerism by the plasma DNA analysis.

One interesting finding from our results is that the hematopoietically-derived DNA molecules are longer than other tissue-derived DNA molecules. These observations suggest that the nonhematopoietically-derived DNA molecules may undergo more extensive degradation in plasma than the hematopoietically-derived ones. One explanation of this may be that such nonhematopoietic DNA molecules might have a longer distance to travel between the tissue of origin and the point of sampling (e.g.

the antecubital vein). The longer resident time may make such molecules more susceptible to attack by plasma nucleases. Conversely, for hematopoietic DNA, the tissue of origin and the point of sampling can be regarded to be within the same system. An alternative explanation is that the nonhematopoietic tissue-derived DNA has been degraded before reaching blood stream due to the barrier of the organs.

The size difference between hematopoietically- and nonhematopoietically-derived DNA in plasma is reminiscent of the size difference between maternal and fetal-derived DNA fragment in maternal plasma (Chan *et al.* 2004). Fetal-derived DNA in maternal plasma is believed to originate from placental cells (Lo *et al.* 2007a) while the background maternal DNA is mainly derived from maternal hematopoietic cells (Lui *et al.* 2002). The placenta-derived fetal DNA may undergo the same regulation as other nonhematopoietic tissue-derived DNA in plasma, thus being shorter than the background maternal DNA that is predominantly of maternal hematopoietic origin. Hence, I hypothesize that there may be shared mechanistic steps regulating the release and/or degradation of fetus- and nonhematopoietic tissue-derived DNA. This hypothesis could be further validated by characterizing the molecular size of the plasma DNA from sex-mismatched solid organ transplant patients (Lo *et al.* 1998b).

From the size histograms, a discrepancy in proportion of short fragments was observed between the male post-HSCT patients and the healthy adult males. It is possible that hematopoietically-derived plasma DNA (though longer than those from other tissue sources) in post-HSCT patients is slightly more actively degraded than that of healthy controls. In regard to the liberation of plasma DNA, the altered cell turnover after HSCT (Alpdogan *et al.* 2008; Brugnoni *et al.* 1999; Wekerle *et al.*

2001) may also attribute to such discrepancy. Further work could be done with more cases involved to perform a detailed inter-individual comparison.

In conclusion, PE sequencing of circulating DNA in the sex-mismatched HSCT model not only confirms the previous finding of the major hematopoietic contribution to circulating DNA, but also reveals the difference in fragment size between DNA molecules from two cellular origins. The underlying mechanisms regulating such difference in molecular size remain to be explored. However, the size distribution of plasma DNA may be clinically useful for the evaluation of relapse after HSCT or other tissue transplantation. Because plasma DNA has been associated with cell death, donor-derived DNA may be released as a result of graft rejection or other sources of tissue damage in the transplanted organ (Lui *et al.* 2003). In this scenario, both the DNA concentration and the size distribution of the overall or specific organ-derived DNA molecules may change, which could potentially act as useful markers for these processes. It would be particularly informative to document serial data on the variation in DNA concentration and fragment size of plasma DNA during and after the transplantation procedure, especially those who are suffering from graft rejection episodes.

SECTION VI : CONCLUDING REMARKS

CHAPTER 10: Conclusions and future perspectives

10.1 Fetal DNA molecules exist in maternal plasma with diverse fragment species.

Since 1997, the discovery of fetal DNA in maternal plasma has stimulated many applications using maternal plasma as starting material on the purpose of noninvasive prenatal diagnosis. Nonetheless, the relevant biological study is relatively limited. The knowledge of biological properties of the fetal DNA molecules in maternal plasma, in turn, would be conducive to the development of diagnostic assays. For example, Chan *et al.* have demonstrated the fetal DNA molecules in maternal plasma are significantly shorter than maternally-derived ones (Chan *et al.* 2004). This finding later led to a number of improvements on the diagnostic performances of the maternal plasma DNA assessments, such as improving the prenatal detection of fetal mutations using the size-fractionated circulatory DNA in maternal plasma (Li *et al.* 2005; Li *et al.* 2004a; Li *et al.* 2004b) and relatively enriching the fetal DNA with shorter PCR amplicons (Lun *et al.* 2008; Sikora *et al.* 2009). Therefore, in this thesis, efforts are put to study the characteristics of fetal DNA in maternal plasma.

As demonstrated in Chapter 4, cell-free fetal DNA molecules in maternal plasma exist in the form of diverse fragment species with non-uniform cleavage sites, suggesting the difficulties in detecting fetal chromosomal aneuploidies when using the conventional QPCR that only quantifies a fixed amplicon. This led us to turn our group's focus to the recently developed NGS platforms which can perform the DNA sequence analysis in a locus-independent way without the restriction of fragment species.

10.2 Massively parallel sequencing of maternal plasma DNA for fetal trisomy 13 and 18 detection

Our group have demonstrated the use of massively parallel sequencing of maternal plasma DNA for the noninvasive prenatal detection of fetal trisomy 21 with 100% accuracy (Chiu *et al.* 2008). In clinical application, noninvasive prenatal detection of fetal trisomy 13 and 18 is also of great importance; hence, in Chapter 5, I explore the efficacy of the same approach for the noninvasive prenatal diagnosis of trisomy 13 and 18. In contrast to chromosome 21 which has moderate GC content, chromosome 13 and 18 are relatively GC-low and possess a number of gene deserts within their chromosomes (Dunham *et al.* 2004; Hattori *et al.* 2000; Nusbaum *et al.* 2005). When encountering a NGS platform with substantial GC bias (Chiu *et al.* 2008; Dohm *et al.* 2008; Fan *et al.* 2008; Hillier *et al.* 2008), the compositional features may contribute to the imprecise measurements of the DNA fragments originating from the two chromosomes, thereby resulting in suboptimal diagnostic performances for noninvasive prenatal detection of fetal trisomy 13 and 18. Besides, this internal GC bias persists across multiple sequencing modes, i.e., SR, PE and multiplexing sequencing modes (Chapter 5, 6, 7 and 9).

GC bias introduced by the NGS platforms places a practical limit on the sensitivity of aneuploidy detection. Generally, there are several ways to address such GC bias and hence improve the diagnostic performance. One is to normalize the GC bias in the combination of the sequencing data and the compositional feature of the target chromosome. In chapter 5, I demonstrate the feasibility of this method in terms of the reduction in the sample-to-sample variation by the use of a manual size-selection method. Bioinformatically, this analysis can be carried out thoroughly. For instance, establishing a statistical model by taking all potential factors into account (Chu *et al.*

2009) and correcting the read number by the local genomic GC content in each narrow bin (Fan *et al.* 2010) can achieve the GC normalization.

Another way is to develop other NGS-based methods for the noninvasive prenatal diagnosis of fetal chromosomal aneuploidies instead of the current methods relying on single molecule counting. As demonstrated in Chapter 8, the size ranking of chromosome 21 is lowered in the maternal plasma samples from trisomy 21 pregnancies, indicating that fragment size for the at-risk chromosome would be an alternative diagnostic parameter other than the genomic representation of this chromosome.

The third way is to employ the third-generation sequencing platforms for massively parallel maternal plasma DNA sequencing. The current NGS platforms achieve the cloning-free DNA amplification, but still require a PCR for library enrichment, potentially creating mutations in clonally-amplified templates and introducing amplification bias towards AT-rich and GC-rich target sequences which results in under- or overrepresentations of chromosomes (Metzker 2009). The intrinsic GC bias of Illumina sequencing platform that is employed for sequencing maternal plasma DNA in this thesis, has been reported, and believed to stem from the amplification steps (Chiu *et al.* 2008; Dohm *et al.* 2008; Fan *et al.* 2008; Hillier *et al.* 2008). In contrast, the third-generation sequencing platforms, which adopt amplification-free, single-molecule sequencing technologies, are supposed to mitigate the underlying GC bias with each DNA molecule equally sequenced.

10.3 Massively parallel sequencing of maternal plasma DNA for fetal trisomy 21 detection

Trisomy 21 occurs in 1 in 800 live births (Driscoll *et al.* 2009). Chances of a woman carrying a trisomy 21 fetus increase with advancing maternal age (Heffner 2004). The noninvasive prenatal diagnosis of fetal trisomy 21 is a long-sought goal for obstetricians. Highly precise fetal DNA quantitative analysis by massively parallel sequencing of maternal plasma DNA permits the accurate detection of fetal trisomy 21 (Chiu *et al.* 2009a). However, before becoming an easy-to-use and inexpensive routine test in clinical practice, this approach needs further refinements. Hence, two measures for promoting its clinical implementation, namely, lowering the starting volume of maternal plasma and barcoding multiple maternal plasma samples, are investigated in Chapter 6 for the purposes of easing sample handling and reducing the cost per case, respectively.

As demonstrated in Chapter 6, as low as one milliliter of maternal plasma would be adequate for massively parallel sequencing of maternal plasma DNA to achieve the successful detection of fetal trisomy 21. On the barcoding side, 4-plex protocol with around 0.6 million of unique and perfect sequence reads per case shows 100% sensitivity and specificity for fetal trisomy 21 detection. Both studies augment the practical feasibility of maternal plasma DNA sequencing for the noninvasive prenatal diagnosis of fetal trisomy 21.

Sequencing costs will fall dramatically in the coming years. In the meantime, the sequencing throughput of a number of newly launched sequencers is being expanded. For example, the SOLiD sequencing platform (Applied Biosystems by Life Technologies) and the HiSeq 2000 (Illumina) can generate 5 times (Chiu *et al.*

2009b) and 6 times more reads (http://www.illumina.com/) per run comparing to the current Illumina GA II platform, respectively. At an equivalent sequencing depth per case, the capacity of these sequencers will convert the current 4-plex level to 24-plex and 28-plex per sequencing chamber, respectively; hence, 192 and 224 cases can be analyzed simultaneously per run, enabling a much more cost-effective access in practice and consequently bringing this approach into clinical usage in the near future.

10.4 Massively parallel paired-end sequencing of maternal plasma DNA for fetal chromosomal aneuploidy detection and fragment size analysis

PE sequencing, in which both ends of a fragment are sequenced to provide more information about the fragment, have the potential to enhance the utility of short reads for maternal plasma DNA sequence analysis. In spite of prolonged duration (doubling the duration of SR sequencing) and higher cost (requiring additional reagents for the second read generation and another round of cyclic sequencing), PE sequencing permits additional information of plasma DNA fragment, i.e., molecular size.

As SR sequencing, PE sequencing is able to achieve the fetal chromosomal aneuploidy detection by identifying the chromosomal origin of each sequenced molecule and quantifying the genomic representation based on the single molecule counting strategy. More importantly, as described in Chapter 7, the obtained fragment size information enables us to gain the size characteristics of maternal plasma DNA at an unprecedented resolution and demonstrate the significant difference in molecular size between maternally- and fetal-derived DNA molecules in maternal plasma. The fetal DNA fractional concentration is consequently enriched

at the post-sequencing stage with the ISSS method. Although the effectiveness of size-enriched fetal DNA for trisomy 21 detection is attenuated by the decreased number of retrieved sequence reads, the molecular size by itself show potential to be an independent marker for the noninvasive diagnosis of fetal chromosomal aneuploidies.

10.5 Biological properties of plasma DNA revealed by massively parallel paired-end sequencing of plasma DNA

High-resolution size profiles of plasma DNA offered by PE sequencing help to exhibit the biological properties of plasma DNA and shed light on the underlying release and clearance mechanisms for circulating DNA in plasma. In Chapter 8 and 9, a series of detailed analyses are performed and several important issues are revealed. First, the size distribution of plasma DNA from different types of individuals and patients hold the same sharp peak at ~166 bp, a standard size of human chromotosome (Widom 1992), thus leading to the hypothesis that the circulating nuclear DNA in plasma may be largely contributed by nucleosomal DNA release during cell apoptosis. This notion is further supported by the other observations, including the methylation effect on fragment size of plasma DNA, the difference in chrX fragment size between females and males, and the extremely short DNA fragment originating from mtDNA. Secondly, there is a dynamic change in fragment size of both maternally- and fetal-derived DNA molecules in maternal plasma from pregnant women, suggesting the clearance mechanisms for plasma DNA is conducted continuously and maybe overloaded during late pregnancy. Thirdly, by analyzing the fragment size of plasma DNA from post-HSCT patients, the difference in molecular size between hematopoietically- and nonhematopoietically-derived DNA molecules is uncovered. Although much work is

still required to completely unravel the enigmatic mechanisms of circulating plasma DNA, the results in this thesis have made a step forward in this regard.

These data are not only of academic interest but also of practical importance for the future developments of molecular markers for prenatal diagnosis. For instance, the placental epigenetic signature in maternal plasma is a gender- and polymorphism-independent fetal DNA marker for prenatal diagnosis (Chan *et al.* 2006; Chim *et al.* 2005), and investigators are systematically searching for the placenta-specific epigenetic signatures on target chromosome for the noninvasive prenatal detection of fetal aneuploidies (Chim *et al.* 2008; Papageorgiou *et al.* 2009). The difference in fragment size between highly and lowly methylated regions discovered in Chapter 8 suggests that the hypermethylated DNA molecules in plasma are relatively longer and hence readier to be detected in a PCR system. As a result, a locus hypermethylated in the fetus/placenta and hypomethylated in maternal blood cells would be a more promising marker for prenatal diagnosis by maternal plasma analysis.

10.6 Prospects for future work

Large-scale clinical trial for validation and further improvements

For the noninvasive prenatal detection of trisomy 21 by the maternal plasma DNA sequence analysis, both the experimental protocol and analytical method have been refined and are ready for clinical practice; subsequently, a large-scale clinical trial can be performed to validate its diagnostic performance and feasibility for clinical use. Nevertheless, for the NGS-based noninvasive detection of fetal trisomy 13 and 18, either further optimization of the analytical method based on molecule counting or alternative analytical method other than molecule counting is needed.

On the currently available NGS platforms, several improvements can be introduced to the massively parallel sequencing of maternal plasma DNA to meet the requirement of enlarging sample throughput but without the loss of sequencing depth. One is to couple sequencing with a recently developed microsystem which can integratively isolate the plasma DNA according to a defined fragment size (Hahn *et al.* 2009). As demonstrated in Chapter 7, the fetal DNA is enriched mostly at the size cutoffs of 125 bp and 150 bp, one therefore can only sequence the size-fractionated maternal plasma DNA (e.g., shorter than 150 bp) produced by the microsystem for the noninvasive prenatal diagnosis of chromosomal aneuploidies. A more efficient way is to exploit the target-enrichment strategy (Mamanova *et al.* 2010). With the availability of tools to selectively enrich target sequences of interest (Mamanova *et al.* 2010), one can only focus on an at-risk chromosome relative to one of the non-trisomic chromosomes, rather than at the whole genome level, and assess whether there is a chromosomal imbalance between two chromosomes in a pregnancy carrying a trisomic fetus, so that the increased sample throughput per run can be attained with the sufficient sequencing depth per case.

The repeat-masked genome is used throughout the whole thesis. However, after sequence alignment, there are still sequence reads that could be aligned to multiple locations, e.g., sequence reads mapped to multi-copy genes or gene clusters. For clinical utility, only the uniquely matched sequence reads should be selected for the subsequent analysis. As an alternative to such post-alignment processing, one can use a specifically masked version of the genome, e.g., a reference genome with the multi-copy genes and gene clusters masked, for alignment.

Further fragment size analysis of maternal plasma DNA

Many biological issues concerning the circulating DNA in maternal plasma could be revisited in terms of fragment size offered by PE sequencing. The biological basis for the observed changes in fragment size of fetal and total DNA molecules in maternal plasma requires further investigation. It is of great interest to document serial data on the variation in fragment size of plasma DNA during pregnancy, so as to validate whether the short DNA would be diluted as pregnancy progresses. It is also of relevance to see whether the bimodal distribution pattern would return following delivery by sequencing the plasma DNA from pre- and post-delivery maternal plasma samples. Moreover, it is worth investigating how the size profiles look like in the conditions of the pregnancy-associated complications, e.g., preeclampsia, in which a five-fold increase in fetal DNA concentration in serum compared with controls (Lo *et al.* 1999a) and the impaired clearance of fetal DNA from maternal plasma (Lau *et al.* 2002) have been observed.

Fragment size analysis in other clinical scenarios

Apart from the pregnancy-associated diagnosis, it is envisioned that the fragment size analysis on plasma DNA molecules has great potential to be applied to other clinical scenarios, such as oncology and solid transplantation. Taking cancer for example, circulating DNA size has promising implications in cancer management (Chan *et al.* 2008; Jiang *et al.* 2006; Umetani *et al.* 2006a; Umetani *et al.* 2006b; Wang *et al.* 2003), and the presence of long circulating DNA strands, i.e. increased DNA integrity, has been suggested to be indicative of neoplastic disease (Jiang *et al.* 2006; Umetani *et al.* 2006a; Umetani *et al.* 2006b; Wang *et al.* 2003). Previous studies have typically used quantitative PCR, in which the ratio of copy number of long DNA fragments to that of short DNA fragments within the same gene or amplification locus is calculated to size DNA (Jiang *et al.* 2006; Umetani *et al.*

2006a; Umetani *et al.* 2006b; Wang *et al.* 2003). This quantification method, however, is not so accurate and tends towards a variety of results, in view of the fact that the plasma DNA in cancer patient is also fragmented in nature (Giacona *et al.* 1998). Predictably, one can precisely define single molecule DNA integrity based on the detailed size profile of plasma DNA provided by PE sequencing to discriminate the cancer patients from healthy individuals and nonmalignant patients or monitor the dynamics in fragment size as the progression of diseases.

Further biological study of plasma DNA

In this thesis, it is hypothesized that nucleosomal DNA released from apoptotic cells is probably the predominant source of plasma DNA. However, conclusive evidence for the support of this hypothesis requires further analysis. One important future work is to explore the biological properties of plasma DNA beyond the size profile, such as cleavage site (Widlak *et al.* 2000) and nucleotide content/pattern (Segal *et al.* 2006) of each sequenced fragment species. With the comparative analysis between the resultant observations and the features of the sophisticated biological procedures (e.g., apoptosis and necrosis), we may be able to trace the biological release/clearance mechanisms of plasma DNA with high certainty. Besides, *in vitro* and *in vivo* experiments (Jimenez *et al.* 2003; Tjoa *et al.* 2006) might provide other solid evidence for such hypothesis.

In addition, a detailed sequence annotation of plasma DNA from healthy individuals, with respect to sequences attributable to repeats, genes, RNA, protein-coding DNA sequences and so on, is necessary for biological study and will benefit other clinical settings where the plasma DNA analysis is informative for disease diagnosis or monitoring, e.g. cancer and infectious diseases. Recently, Beck *et al.* have attempted

to generate the sequence profile of the circulating DNA in healthy individuals (Beck *et al.* 2009). However, it has been noted that only thousands of sequence read per case were obtained in their analysis. Although substantially more than previous sequences obtained from cloning and sequencing method (Suzuki *et al.* 2008), the sequencing depth of plasma DNA in their work was relatively insufficient and would potentially result in some uninterpretable sequence bias. It is quite viable to use our readily available datasets to establish a normative sequence distribution in the plasma DNA from healthy individuals. As a result, future comparison between the sequence profiles of plasma DNA for patients with malignant diseases and normal controls will be evidence-based and effort-saving.

Comprehensive analysis based on gender-independent markers

In the analytical method in Chapter 9, the DNA fragments from chrY are used to represent the specific DNA from the male donor/recipient. On the non-Y side, although predominantly of hematological origin, the DNA fragments from non-Y chromosomes in the plasma of HSCT recipients are indeed intermixed with a minor but definite proportion of nonhematologically-contributed DNA. Moreover, the comparative analysis between fragment size for Y and non-Y chromosomes is only achievable for the sex-mismatched cases. Hence, a gender-independent marker would be ideal for a comprehensive and complete analysis. One of such markers is SNP. In practice, the blood samples from donors and recipients are easily accessible. Using the commercial SNP microarrays, the SNP information from both donors and recipients could be obtained. In combination with PE sequencing of plasma DNA, one could separately analyze millions of reads according to the SNPs and then attain the fragment patterns of purely donor- and recipient-derived DNA. Such a strategy could thus show us a more valid and in-depth picture of plasma DNA from both sides

and could be applicable for both sex-matched and sex-mismatched HSCT patients. HSCT model is instructive for the biological study of plasma DNA. Better characterization of plasma DNA from post-HSCT patients will provide the insight into the biological events of plasma DNA and also facilitate the clinical use of molecular size information, e.g., measuring dynamic changes in fragment size for post-transplantation monitoring.

Appendix I Clinical details and sequencing counts of ten maternal plasma samples involving euploid fetuses sequenced in previous sequencing runs.

Case no.	Fetal sex	GA (weeks + days)	Input DNA (ng)	Karyotype	Total sequenced counts	Total U0-1-0-0 counts	Chr13 U0-1-0-0 counts	Chr18 U0-1-0-0 counts	ChrY U0-1-0-0 counts
M2972	M	13+2	16	46XY	12,183,087	2,269,126	81,263	66,938	975
M3245	M	17+2	14	46XY	11,895,455	2,385,658	87,727	71,459	777
M4181	M	17+2	18	46XY	11,231,617	2,233,436	78,269	66,362	964
M2791	M	13+5	13	46XY	10,679,744	2,374,594	83,239	70,012	972
M4402	M	12+4	20	46XY	11,782,969	2,769,857	97,492	81,243	645
M4404	M	17+5	20	46XY	11,203,712	2,743,024	98,017	80,412	708
M4420	M	13+5	20	46XY	9,820,505	2,590,205	89,070	75,753	464
M4421	M	12+3	20	46XY	10,206,122	2,427,082	81,356	69,976	893
M4422	M	17	20	46XY	10,776,947	2,610,307	88,314	74,996	868
M4443	M	11+6	20	46XY	9,661,057	2,396,222	79,996	69,602	854

M, male; GA, gestational age;

Appendix II Clinical details and sequencing counts of maternal plasma samples analyzed by massively parallel multiplexed sequencing.

Case no.	GA (weeks + days)	Input DNA (ng)[a]	Fetal%[b]	Karyotype	8-plex		4-plex	
					Total sequenced counts	Total U0-1-0-0 counts	Total sequenced counts	Total U0-1-0-0 counts
U92797	12+2	9.5	3.0	46XY	1,140,919	357,060	2,130,540	695,369
U93156	12+6	5.8	0.9	46XY	865,710	254,048	1,768,081	563,163
U94103	12+2	11.1	2.6	46XY	966,943	256,570	2,139,721	680,437
U94306	12+4	5.5	4.8	46XY	940,427	294,925	1,921,148	629,393
U94427	13+4	7.8	5.9	46XY	1,029,072	295,418	2,210,795	706,889
U94837	13+2	6.5	7.4	46XY	953,951	295,115	1,872,674	605,584
U94884	12+3	5.8	5.8	46XY	1,304,357	381,617	2,649,312	834,462
U94892	14+3	7.3	5.0	46XY	1,058,862	302,222	2,240,156	719,105
U2797	12+3	5.5	1.9	46XY	754,515	237,658	1,583,184	510,234
U52950	13+3	5.9	0.3	46XY	738,871	200,133	834,295	277,751
U80172	13	8.1	5.3	46XY	794,096	215,479	897,513	298,849
U93022	14+1	6.0	10.3	46XY	820,039	256,525	1,728,935	570,652
U93543	12+1	2.8	6.4	46XY	824,587	266,777	1,725,641	568,674
U93897	12	3.3	13.0	46XY	528,788	174,347	1,378,453	463,058
U94084	13+5	4.2	1.3	46XY	797,257	229,803	1,078,018	307,438
U94171	13+1	4.5	3.6	46XY	780,344	241,169	1,523,069	511,160
U94178	13+5	6.3	3.3	46XY	757,687	216,785	1,674,777	545,023
U94294	12+4	2.3	9.6	46XY	843,377	260,490	2,201,952	714,317
U94394	16+3	2.8	14.4	46XY	754,886	222,693	853,438	250,353
U94407	13+2	4.4	4.7	46XY	817,085	265,009	1,781,217	588,393
U94410	14+6	3.4	8.2	46XY	743,365	237,020	1,449,414	462,273
U94570	11+3	2.9	11.4	46XY	867,794	255,800	1,805,013	559,626
U94628	11+6	4.9	1.5	46XY	822,140	254,185	1,976,182	619,651
U94813	14+2	10.2	5.6	46XY	871,876	259,536	1,607,763	515,859

U94818	14+3	6.5	15.0	46XY	1,060,194	306,389	1,954,820	626,878
U15518	14	16.0	na	46XX	858,552	263,825	1,397,189	477,053
U37106	13+1	8.9	na	46XX	946,097	288,180	1,640,805	546,231
U39505	12+5	25.7	na	46XX	556,386	167,663	1,220,373	403,396
U52385	12+1	4.2	na	46XX	689,800	191,287	1,759,657	572,593
U8610	12+6	5.5	na	46XX	911,323	279,861	2,070,466	650,014
U93173	14+4	4.1	na	46XX	692,318	219,450	1,681,035	556,120
U93243	12+1	12.3	0.4	46XX	812,243	259,808	1,767,632	594,307
U93375	12+6	5.3	na	46XX	640,677	206,964	1,590,574	531,134
U93619	13+1	4.5	na	46XX	742,981	232,596	1,350,500	430,729
U94059	14	8.9	na	46XX	792,952	245,214	2,003,503	662,756
U94132	14+5	12.4	na	46XX	846,506	255,003	1,526,102	510,019
U94149	13+2	4.0	na	46XX	851,216	278,292	1,916,027	645,760
U94244	14	4.5	na	46XX	763,558	247,003	1,861,219	614,936
U94255	12+1	7.2	na	46XX	874,274	279,122	2,254,830	735,531
U94316	14+1	7.1	na	46XX	928,331	277,860	2,155,768	711,384
U94326	13+2	5.8	na	46XX	891,851	246,039	2,284,262	730,122
U94488	13+5	9.6	na	46XX	863,270	268,495	1,773,858	575,933
U94573	11+6	3.0	na	46XX	678,511	215,392	1,175,578	375,851
U94626	13+5	6.0	na	46XX	775,754	218,972	1,394,929	444,014
U94648	14+1	8.2	na	46XX	947,412	285,205	2,025,893	634,746
U32137	13	5.3	9.0	21+47XY	853,587	280,783	1,852,537	624,781
U94099	14+2	8.7	2.8	21+47XY	716,906	231,391	1,851,165	622,462
U94946	13+1	9.9	1.5	21+47XY	1,031,553	323,321	1,944,088	615,586
U93853	12+4	6.5	na	21+47XX	888,606	254,418	1,141,505	328,086
U94112	13	19.9	na	21+47XX	882,269	287,874	1,789,404	602,362
U94376	14	6.6	na	21+47XX	983,863	306,234	2,111,056	690,654
U94870	13+1	7.1	na	21+47XX	823,750	219,251	956,018	319,138

a: the DNA amount was estimated by the *HBB* QPCR.

b: the fetal DNA% was estimated based on the results from the *SRY/HBB* QPCR assays.

Appendix III Clinical details and sequencing counts of ten maternal plasma samples involving T21 fetuses sequenced in previous sequencing runs.

Case no.	Fetal sex	GA (weeks + days)	Input DNA (ng)	Karyotype	Total sequenced counts	Total U0-1-0-0 counts	%chrY
M4386	M	13+6	20	21+47XY	9,626,221	2,516,289	0.016
M1519	M	20+3	29	21+47XY	10,403,129	2,331,208	0.019
M3228	M	14+5	19	21+47XY	10,625,019	2,271,819	0.031
M3438	M	13+1	23	21+47XY	10,417,954	2,416,779	0.046

M, male; GA, gestational age;

Reference

Adachi, T., M. Nakanishi, Y. Otsuka, K. Nishimura, G. Hirokawa, Y. Goto, H. Nonogi, and N. Iwai. 2010. Plasma MicroRNA 499 as a Biomarker of Acute Myocardial Infarction. *Clin Chem.*

Adessi, C., G. Matton, G. Ayala, G. Turcatti, J.J. Mermod, P. Mayer, and E. Kawashima. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* **28:** E87.

Ahn, S.M., T.H. Kim, S. Lee, D. Kim, H. Ghang, D.S. Kim, B.C. Kim, S.Y. Kim, W.Y. Kim, C. Kim et al. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19:** 1622-1629.

Alberry, M., D. Maddocks, M. Jones, M. Abdel Hadi, S. Abdel-Fattah, N. Avent, and P.W. Soothill. 2007. Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. *Prenat Diagn* **27:** 415-418.

Ali, S. and S.E. Hasnain. 2003. Genomics of the human Y-chromosome. 1. Association with male infertility. *Gene* **321:** 25-37.

Alpdogan, S.O., S.X. Lu, N. Patel, S. McGoldrick, D. Suh, T. Budak-Alpdogan, O.M. Smith, J. Grubin, C. King, G.L. Goldberg et al. 2008. Rapidly proliferating CD44hi peripheral T cells undergo apoptosis and delay posttransplantation T-cell reconstitution after allogeneic bone marrow transplantation. *Blood* **112:** 4755-4764.

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350-355.

Amicucci, P., M. Gennarelli, G. Novelli, and B. Dallapiccola. 2000. Prenatal diagnosis of myotonic dystrophy using fetal DNA obtained from maternal plasma. *Clin Chem* **46:** 301-302.

Amoura, Z., J.C. Piette, H. Chabre, P. Cacoub, T. Papo, B. Wechsler, J.F. Bach, and S. Koutouzov. 1997. Circulating plasma levels of nucleosomes in patients with systemic lupus erythematosus: correlation with serum antinucleosome antibody titers and absence of clear association with disease activity. *Arthritis Rheum* **40:** 2217-2225.

Anker, P., M. Stroun, and P.A. Maurice. 1975. Spontaneous release of DNA by human blood lymphocytes as shown in an in vitro system. *Cancer Res* **35:** 2375-2382.

Anker, P., M. Stroun, and P.A. Maurice. 1976. Spontaneous extracellular synthesis of DNA released by human blood lymphocytes. *Cancer Res* **36:** 2832-2839.

Antin, J.H., R. Childs, A.H. Filipovich, S. Giralt, S. Mackinnon, T. Spitzer, and D. Weisdorf. 2001. Establishment of complete and mixed donor chimerism after allogeneic lymphohematopoietic transplantation: recommendations from a workshop at the 2001 Tandem Meetings of the International Bone Marrow Transplant Registry and the American Society of Blood and Marrow Transplantation. *Biol Blood Marrow Transplant* **7:** 473-485.

Ariga, H., H. Ohto, M.P. Busch, S. Imamura, R. Watson, W. Reed, and T.H. Lee. 2001. Kinetics of fetal cellular and cell-free DNA in the maternal circulation during and after pregnancy: implications for noninvasive prenatal diagnosis. *Transfusion* **41:** 1524-1530.

Au, K.F., H. Jiang, L. Lin, Y. Xing, and W.H. Wong. 2010. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*

Bartova, E., J. Krejci, A. Harnicarova, G. Galiova, and S. Kozubek. 2008. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem* **56:** 711-721.

Bashir, A., S. Volik, C. Collins, V. Bafna, and B.J. Raphael. 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol* **4:** e1000051.

Beck, J., H.B. Urnovitz, W.M. Mitchell, and E. Schutz. 2010. Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. *Mol Cancer Res* **8:** 335-342.

Beck, J., H.B. Urnovitz, J. Riggert, M. Clerici, and E. Schutz. 2009. Profile of the circulating DNA in apparently healthy individuals. *Clin Chem* **55:** 730-738.

Benachi, A., A. Yamgnane, M. Olivi, Y. Dumez, E. Gautier, and J.M. Costa. 2005. Impact of formaldehyde on the in vitro proportion of fetal DNA in maternal plasma and serum. *Clin Chem* **51:** 242-244.

Benita, Y., R.S. Oosting, M.C. Lok, M.J. Wise, and I. Humphery-Smith. 2003. Regionalized GC content of template DNA as a predictor of PCR success. *Nucleic Acids Res* **31:** e99.

Bentley, D.R. 2006. Whole-genome re-sequencing. *Curr Opin Genet Dev* **16:** 545-552.

Bentley, D.R. S. Balasubramanian H.P. Swerdlow G.P. Smith J. Milton C.G. Brown K.P. Hall D.J. Evers C.L. Barnes H.R. Bignell et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Bianchi, D.W. 1997. Progress in the genetic analysis of fetal cells circulating in maternal blood. *Curr Opin Obstet Gynecol* **9:** 121-125.

Bianchi, D.W., A.F. Flint, M.F. Pizzimenti, J.H. Knoll, and S.A. Latt. 1990. Isolation of fetal DNA from nucleated erythrocytes in maternal blood. *Proc Natl Acad Sci U S A* **87:** 3279-3283.

Bianchi, D.W., J.M. Williams, L.M. Sullivan, F.W. Hanson, K.W. Klinger, and A.P. Shuber. 1997. PCR quantitation of fetal cells in maternal blood in normal and aneuploid pregnancies. *Am J Hum Genet* **61:** 822-829.

Bianchi, D.W., G.K. Zickwolf, G.J. Weil, S. Sylvester, and M.A. DeMaria. 1996. Male fetal progenitor cells persist in maternal blood for as long as 27 years postpartum. *Proc Natl Acad Sci U S A* **93:** 705-708.

Birch, L., C.A. English, K. O'Donoghue, O. Barigye, N.M. Fisk, and J.T. Keer. 2005. Accurate and robust quantification of circulating fetal and total DNA in maternal plasma from 5 to 41 weeks of gestation. *Clin Chem* **51:** 312-320.

Bischoff, F.Z., D.E. Lewis, and J.L. Simpson. 2005. Cell-free fetal DNA in maternal blood: kinetics, source and structure. *Hum Reprod Update* **11:** 59-67.

Botezatu, I., O. Serdyuk, G. Potapova, V. Shelepov, R. Alechina, Y. Molyaka, V. Ananev, I. Bazin, A. Garin, M. Narimanov et al. 2000. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. *Clin Chem* **46:** 1078-1084.

Branton, D., D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang et al. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* **26:** 1146-1153.

Braslavsky, I., B. Hebert, E. Kartalov, and S.R. Quake. 2003. Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* **100:** 3960-3964.

Brugnoni, D., P. Airo, M. Pennacchio, G. Carella, A. Malagoli, A.G. Ugazio, F. Porta, and R. Cattaneo. 1999. Immune reconstitution after bone marrow transplantation for combined immunodeficiencies: down-modulation of Bcl-2 and high expression of CD95/Fas account for increased susceptibility to spontaneous and activation-induced lymphocyte cell death. *Bone Marrow Transplant* **23:** 451-457.

Calin, G.A., C.G. Liu, C. Sevignani, M. Ferracin, N. Felli, C.D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono et al. 2004. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc Natl Acad Sci U S A* **101:** 11755-11760.

Campbell, P.J., P.J. Stephens, E.D. Pleasance, S. O'Meara, H. Li, T. Santarius, L.A. Stebbings, C. Leroy, S. Edkins, C. Hardy et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40:** 722-729.

Caron, F., C. Jacq, and J. Rouviere-Yaniv. 1979. Characterization of a histone-like protein extracted from yeast mitochondria. *Proc Natl Acad Sci U S A* **76:** 4265-4269.

Cavelier, L., A. Johannisson, and U. Gyllensten. 2000. Analysis of mtDNA copy number and composition of single mitochondrial particles using flow cytometry and PCR. *Exp Cell Res* **259:** 79-85.

Chan, K.C., C. Ding, A. Gerovassili, S.W. Yeung, R.W. Chiu, T.N. Leung, T.K. Lau, S.S. Chim, G.T. Chung, K.H. Nicolaides et al. 2006. Hypermethylated RASSF1A in maternal plasma: A universal fetal DNA marker that improves the reliability of noninvasive prenatal diagnosis. *Clin Chem* **52:** 2211-2218.

Chan, K.C., S.F. Leung, S.W. Yeung, A.T. Chan, and Y.M. Lo. 2008. Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin Cancer Res* **14:** 4141-4145.

Chan, K.C., J. Zhang, A.T. Chan, K.I. Lei, S.F. Leung, L.Y. Chan, K.C. Chow, and Y.M. Lo. 2003a. Molecular characterization of circulating EBV DNA in the plasma of nasopharyngeal carcinoma and lymphoma patients. *Cancer Res* **63:** 2028-2032.

Chan, K.C., J. Zhang, A.B. Hui, N. Wong, T.K. Lau, T.N. Leung, K.W. Lo, D.W. Huang, and Y.M. Lo. 2004. Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem* **50:** 88-92.

Chan, L.Y., T.N. Leung, K.C. Chan, H.L. Tai, T.K. Lau, E.M. Wong, and Y.M. Lo. 2003b. Serial analysis of fetal DNA concentrations in maternal plasma in late pregnancy. *Clin Chem* **49:** 678-680.

Chan, R.W., C.A. Graham, T.H. Rainer, N.Y. Lam, R.W. Chiu, K.W. Chik, V. Lee, and Y.M. Lo. 2007. Use of a bone marrow transplantation model system to demonstrate the hematopoietic origin of plasma S100B mRNA. *Clin Chem* **53:** 1874-1876.

Chen, W., R. Ullmann, C. Langnick, C. Menzel, Z. Wotschofsky, H. Hu, A. Doring, Y. Hu, H. Kang, A. Tzschach et al. 2009. Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *Eur J Hum Genet.*

Chim, S.S., S. Jin, T.Y. Lee, F.M. Lun, W.S. Lee, L.Y. Chan, Y. Jin, N. Yang, Y.K. Tong, T.Y. Leung et al. 2008a. Systematic search for placental DNA-methylation markers on chromosome 21: toward a maternal plasma-based epigenetic test for fetal trisomy 21. *Clin Chem* **54:** 500-511.

Chim, S.S., T.K. Shing, E.C. Hung, T.Y. Leung, T.K. Lau, R.W. Chiu, and Y.M. Lo. 2008b. Detection and characterization of placental microRNAs in maternal plasma. *Clin Chem* **54:** 482-490.

Chim, S.S., Y.K. Tong, R.W. Chiu, T.K. Lau, T.N. Leung, L.Y. Chan, C.B. Oudejans, C. Ding, and Y.M. Lo. 2005. Detection of the placental epigenetic signature of the maspin gene in maternal plasma. *Proc Natl Acad Sci U S A* **102:** 14753-14758.

Chin, L. and J.W. Gray. 2008. Translating insights from the cancer genome into clinical practice. *Nature* **452:** 553-563.

Chinnapapagari, S.K., W. Holzgreve, O. Lapaire, B. Zimmermann, and S. Hahn. 2005. Treatment of maternal blood samples with formaldehyde does not alter the proportion of circulatory fetal nucleic acids (DNA and mRNA) in maternal plasma. *Clin Chem* **51:** 652-655.

Chiu, R.W., C.R. Cantor, and Y.M. Lo. 2009a. Non-invasive prenatal diagnosis by single molecule counting technologies. *Trends Genet* **25:** 324-331.

Chiu, R.W., K.C. Chan, Y. Gao, V.Y. Lau, W. Zheng, T.Y. Leung, C.H. Foo, B. Xie, N.B. Tsui, F.M. Lun et al. 2008. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105:** 20458-20463.

Chiu, R.W., L.Y. Chan, N.Y. Lam, N.B. Tsui, E.K. Ng, T.H. Rainer, and Y.M. Lo. 2003. Quantitative analysis of circulating mitochondrial DNA in plasma. *Clin Chem* **49:** 719-726.

Chiu, R.W., S.S. Chim, I.H. Wong, C.S. Wong, W.S. Lee, K.F. To, J.H. Tong, R.K. Yuen, A.S. Shum, J.K. Chan et al. 2007. Hypermethylation of RASSF1A in human and rhesus placentas. *Am J Pathol* **170:** 941-950.

Chiu, R.W., T.K. Lau, P.T. Cheung, Z.Q. Gong, T.N. Leung, and Y.M. Lo. 2002a. Noninvasive prenatal exclusion of congenital adrenal hyperplasia by maternal plasma analysis: a feasibility study. *Clin Chem* **48:** 778-780.

Chiu, R.W., T.K. Lau, T.N. Leung, K.C. Chow, D.H. Chui, and Y.M. Lo. 2002b. Prenatal exclusion of beta thalassaemia major by examination of maternal plasma. *Lancet* **360:** 998-1000.

Chiu, R.W., W.B. Lui, M.C. Cheung, N. Kumta, A. Farina, I. Banzola, S. Grotti, N. Rizzo, C.J. Haines, and Y.M. Lo. 2006. Time profile of appearance and disappearance of circulating placenta-derived mRNA in maternal plasma. *Clin Chem* **52:** 313-316.

Chiu, R.W., L.L. Poon, T.K. Lau, T.N. Leung, E.M. Wong, and Y.M. Lo. 2001. Effects of blood-processing protocols on fetal and total DNA quantification in maternal plasma. *Clin Chem* **47:** 1607-1613.

Chiu, R.W., H. Sun, R. Akolekar, C. Clouser, C. Lee, K. McKernan, D. Zhou, K.H. Nicolaides, and Y.M. Lo. 2009b. Maternal Plasma DNA Analysis with Massively Parallel Sequencing by Ligation for Noninvasive Prenatal Diagnosis of Trisomy 21. *Clin Chem.*

Choi, M., U.I. Scholl, W. Ji, T. Liu, I.R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106:** 19096-19101.

Chu, T., K. Bunce, W.A. Hogge, and D.G. Peters. 2009. Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. *Bioinformatics* **25:** 1244-1250.

Chung, G.T., R.W. Chiu, K.C. Chan, T.K. Lau, T.N. Leung, and Y.M. Lo. 2005. Lack of dramatic enrichment of fetal DNA in maternal plasma by formaldehyde treatment. *Clin Chem* **51:** 655-658.

Church, G.M. and S. Kieffer-Higgins. 1988. Multiplex DNA sequencing. *Science* **240:** 185-188.

Clarke, J., H.C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4:** 265-270.

Cohen, A.S., D.R. Najarian, A. Paulus, A. Guttman, J.A. Smith, and B.L. Karger. 1988. Rapid separation and purification of oligonucleotides by high-performance capillary gel electrophoresis. *Proc Natl Acad Sci U S A* **85:** 9660-9663.

Consortium, I.H.G.S. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Costa, J.M., A. Benachi, and E. Gautier. 2002. New strategy for prenatal diagnosis of X-linked disorders. *N Engl J Med* **346:** 1502.

Cremonesi, L., S. Galbiati, B. Foglieni, M. Smid, D. Gambini, A. Ferrari, E. Viora, M. Campogrande, M. Pagliano, M. Travi et al. 2004. Feasibility study for a microchip-based approach for noninvasive prenatal diagnosis of genetic diseases. *Ann N Y Acad Sci* **1022:** 105-112.

Cronn, R., A. Liston, M. Parks, D.S. Gernandt, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res* **36:** e122.

DeFrancesco, L. and G. Attardi. 1981. In situ photochemical crosslinking of HeLa cell mitochondrial DNA by a psoralen derivative reveals a protected region near the origin of replication. *Nucleic Acids Res* **9:** 6017-6030.

Deligezer, U., F. Yaman, N. Erten, and N. Dalay. 2003. Frequent copresence of methylated DNA and fragmented nucleosomal DNA in plasma of lymphoma patients. *Clin Chim Acta* **335:** 89-94.

Dhallan, R., W.C. Au, S. Mattagajasingh, S. Emche, P. Bayliss, M. Damewood, M. Cronin, V. Chou, and M. Mohr. 2004. Methods to increase the percentage of free fetal DNA recovered from the maternal circulation. *JAMA* **291:** 1114-1119.

Diehl, F., M. Li, D. Dressman, Y. He, D. Shen, S. Szabo, L.A. Diaz, Jr., S.N. Goodman, K.A. David, H. Juhl et al. 2005. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A* **102:** 16368-16373.

Dijckmeester, W.A., B.P. Wijnhoven, D.I. Watson, M.P. Leong, M.Z. Michael, G.C. Mayne, T. Bright, D. Astill, and D.J. Hussey. 2009. MicroRNA-143 and -205 expression in neosquamous esophageal epithelium following Argon plasma ablation of Barrett's esophagus. *J Gastrointest Surg* **13:** 846-853.

Ding, C., R.W. Chiu, T.K. Lau, T.N. Leung, L.C. Chan, A.Y. Chan, P. Charoenkwan, I.S. Ng, H.Y. Law, E.S. Ma et al. 2004. MS analysis of single-nucleotide differences in circulating nucleic acids: Application to noninvasive prenatal diagnosis. *Proc Natl Acad Sci U S A* **101:** 10762-10767.

Dohm, J.C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36:** e105.

Driscoll, D.A. and S. Gross. 2009. Clinical practice. Prenatal screening for aneuploidy. *N Engl J Med* **360:** 2556-2562.

Dunham, A. L.H. Matthews J. Burton J.L. Ashurst K.L. Howe K.J. Ashcroft D.M. Beare D.C. Burford S.E. Hunt S. Griffiths-Jones et al. 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428:** 522-528.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman et al. 2008. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133-138.

El-Maarri, O., T. Becker, J. Junen, S.S. Manzoor, A. Diaz-Lacava, R. Schwaab, T. Wienker, and J. Oldenburg. 2007. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Hum Genet* **122:** 505-514.

Emlen, W. and M. Mannik. 1984. Effect of DNA size and strandedness on the in vivo clearance and organ localization of DNA. *Clin Exp Immunol* **56:** 185-192.

Enomoto, R., H. Tatsuoka, Y. Yoshida, T. Komai, K. Node, R. Nogami, A. Yamauchi, and E. Lee. 2002. Thymocyte apoptosis induced by phosphorylation of histones is associated

with the change in chromatin structure to allow easy accessibility of DNase. *IUBMB Life* **54:** 123-127.

Enomoto, R., Y. Yoshida, T. Komai, C. Sugahara, Y. Yasuoka, and E. Lee. 2003. Involvement of the change in chromatin structure in thymocyte apoptosis induced by phosphorylation of histones. *Ann N Y Acad Sci* **1010:** 218-220.

Fan, H.C., Y.J. Blumenfeld, U. Chitkara, L. Hudgins, and S.R. Quake. 2008. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* **105:** 16266-16271.

Fan, H.C. and S.R. Quake. 2007. Detection of aneuploidy with digital polymerase chain reaction. *Anal Chem* **79:** 7576-7579.

Fan, H.C. and S.R. Quake. 2010. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS One* **5:** e10439.

Farrer, R.A., E. Kemen, J.D. Jones, and D.J. Studholme. 2009. De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol Lett* **291:** 103-111.

Fedurco, M., A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34:** e22.

Finning, K., P. Martin, and G. Daniels. 2004. A clinical service in the UK to predict fetal Rh (Rhesus) D blood group using free fetal DNA in maternal plasma. *Ann N Y Acad Sci* **1022:** 119-123.

Fleischhacker, M. and B. Schmidt. 2007. Circulating nucleic acids (CNAs) and cancer--a survey. *Biochim Biophys Acta* **1775:** 181-232.

Flusberg, B.A., D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, J. Korlach, and S.W. Turner. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*.

Fournie, G.J., J.P. Courtin, F. Laval, J.J. Chale, J.P. Pourrat, M.C. Pujazon, D. Lauque, and P. Carles. 1995. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. *Cancer Lett* **91:** 221-227.

Galbiati, S., M. Smid, D. Gambini, A. Ferrari, G. Restagno, E. Viora, M. Campogrande, S. Bastonero, M. Pagliano, S. Calza et al. 2005. Fetal DNA detection in maternal plasma throughout gestation. *Hum Genet* **117:** 243-248.

Gautier, E., A. Benachi, Y. Giovangrandi, P. Ernault, M. Olivi, T. Gaillon, and J.M. Costa. 2005. Fetal RhD genotyping by maternal serum analysis: a two-year experience. *Am J Obstet Gynecol* **192:** 666-669.

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

Giacona, M.B., G.C. Ruben, K.A. Iczkowski, T.B. Roos, D.M. Porter, and G.D. Sorenson. 1998. Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls. *Pancreas* **17:** 89-97.

Gilad, S., E. Meiri, Y. Yogev, S. Benjamin, D. Lebanony, N. Yerushalmi, H. Benjamin, M. Kushnir, H. Cholakh, N. Melamed et al. 2008. Serum microRNAs are promising novel biomarkers. *PLoS One* **3:** e3148.

Ginsburg, G.S. and H.F. Willard. 2009. Genomic and personalized medicine: foundations and applications. *Transl Res* **154:** 277-287.

Gonzalez-Gonzalez, M.C., M. Garcia-Hoyos, M.J. Trujillo, M. Rodriguez de Alba, I. Lorda-Sanchez, J. Diaz-Recasens, E. Gallardo, C. Ayuso, and C. Ramos. 2002. Prenatal detection of a cystic fibrosis mutation in fetal DNA from maternal plasma. *Prenat Diagn* **22:** 946-948.

Gonzalez-Gonzalez, M.C., M.J. Trujillo, M. Rodriguez de Alba, M. Garcia-Hoyos, I. Lorda-Sanchez, J. Diaz-Recasens, C. Ayuso, and C. Ramos. 2003. Huntington disease-unaffected fetus diagnosed from maternal plasma using QF-PCR. *Prenat Diagn* **23:** 232-234.

Guibert, J., A. Benachi, A.G. Grebille, P. Ernault, J.R. Zorn, and J.M. Costa. 2003. Kinetics of SRY gene appearance in maternal serum: detection by real time PCR in early pregnancy after assisted reproductive technique. *Hum Reprod* **18:** 1733-1736.

Guliaeva, N.A., E.A. Kuznetsova, and A.I. Gaziev. 2006. [Proteins associated with mitochondrial DNA protect it against the action of X-rays and hydrogen peroxide]. *Biofizika* **51:** 692-697.

Gupta, P.K. 2008. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26:** 602-611.

Hahn, S. and W. Holzgreve. 2002. Prenatal diagnosis using fetal cells and cell-free fetal DNA in maternal blood: what is currently feasible? *Clin Obstet Gynecol* **45:** 649-656; discussion 730-642.

Hahn, T., K.S. Drese, and C.K. O'Sullivan. 2009. Microsystem for isolation of fetal DNA from maternal plasma by preparative size separation. *Clin Chem* **55:** 2144-2152.

Harris, T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J.W. Efcavitch et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320:** 106-109.

Hattori, M., A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311-319.

Heffner, L.J. 2004. Advanced maternal age--how old is too old? *N Engl J Med* **351:** 1927-1929.

Hibi, K., H. Nakayama, T. Yamazaki, T. Takase, M. Taguchi, Y. Kasai, K. Ito, S. Akiyama, and A. Nakao. 2001. Detection of mitochondrial DNA alterations in primary tumors and corresponding serum of colorectal cancer patients. *Int J Cancer* **94:** 429-431.

Hillier, L.W., G.T. Marth, A.R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J.I. Glasscock, M. Hickenbotham, W. Huang et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* **5:** 183-188.

Holdenrieder, S., P. Stieber, H. Bodenmuller, M. Busch, G. Fertig, H. Furst, A. Schalhorn, N. Schmeller, M. Untch, and D. Seidel. 2001. Nucleosomes in serum of patients with benign and malignant diseases. *Int J Cancer* **95:** 114-120.

Holdenrieder, S., P. Stieber, L.Y. Chan, S. Geiger, A. Kremer, D. Nagel, and Y.M. Lo. 2005. Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. *Clin Chem* **51:** 1544-1546.

Honda, H., N. Miharu, Y. Ohashi, O. Samura, M. Kinutani, T. Hara, and K. Ohama. 2002. Fetal gender determination in early pregnancy through qualitative and quantitative analysis of fetal DNA in maternal serum. *Hum Genet* **110:** 75-79.

Huang, X.C., M.A. Quesada, and R.A. Mathies. 1992. DNA sequencing using capillary array electrophoresis. *Anal Chem* **64:** 2149-2154.

Hung, E.C., T.K. Shing, S.S. Chim, P.C. Yeung, R.W. Chan, K.W. Chik, V. Lee, N.B. Tsui, C.K. Li, C.S. Wong et al. 2009. Presence of donor-derived DNA and cells in the urine of sex-mismatched hematopoietic stem cell transplant recipients: implication for the transrenal hypothesis. *Clin Chem* **55:** 715-722.

Huppertz, B., D.S. Tews, and P. Kaufmann. 2001. Apoptosis and syncytial fusion in human placental trophoblast and skeletal muscle. *Int Rev Cytol* **205:** 215-253.

Invernizzi, P., M.L. Biondi, P.M. Battezzati, F. Perego, C. Selmi, F. Cecchini, M. Podda, and G. Simoni. 2002. Presence of fetal DNA in maternal plasma decades after pregnancy. *Hum Genet* **110:** 587-591.

Ishihara, N., H. Matsuo, H. Murakoshi, J.B. Laoag-Fernandez, T. Samoto, and T. Maruo. 2002. Increased apoptosis in the syncytiotrophoblast in human term placentas complicated by either preeclampsia or intrauterine growth retardation. *Am J Obstet Gynecol* **186:** 158-166.

Jahr, S., H. Hentze, S. Englisch, D. Hardt, F.O. Fackelmayer, R.D. Hesch, and R. Knippers. 2001. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res* **61:** 1659-1665.

Jiang, C. and B.F. Pugh. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10:** 161-172.

Jiang, W.W., M. Zahurak, D. Goldenberg, Y. Milman, H.L. Park, W.H. Westra, W. Koch, D. Sidransky, and J. Califano. 2006. Increased plasma DNA integrity index in head and neck cancer patients. *Int J Cancer* **119:** 2673-2676.

Jimenez, D.F. and A.F. Tarantal. 2003. Quantitative analysis of male fetal DNA in maternal serum of gravid rhesus monkeys (Macaca mulatta). *Pediatr Res* **53:** 18-23.

Johnson-Hopson, C.N. and C.M. Artlett. 2002. Evidence against the long-term persistence of fetal DNA in maternal plasma after pregnancy. *Hum Genet* **111:** 575.

Jones, C.J. and H. Fox. 1980. An ultrastructural and ultrahistochemical study of the human placenta in maternal pre-eclampsia. *Placenta* **1:** 61-76.

Keijser, B.J., E. Zaura, S.M. Huse, J.M. van der Vossen, F.H. Schuren, R.C. Montijn, J.M. ten Cate, and W. Crielaard. 2008. Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res* **87:** 1016-1020.

Kel-Margoulis, O.V., D. Tchekmenev, A.E. Kel, E. Goessling, K. Hornischer, B. Lewicki-Potapov, and E. Wingender. 2003. Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol* **3:** 145-171.

Khodarev, N.N., T. Bennett, N. Shearing, I. Sokolova, J. Koudelik, S. Walter, M. Villalobos, and A.T. Vaughan. 2000. LINE L1 retrotransposable element is targeted during the initial stages of apoptotic DNA fragmentation. *J Cell Biochem* **79:** 486-495.

Kim, J.I., Y.S. Ju, H. Park, S. Kim, S. Lee, J.H. Yi, J. Mudge, N.A. Miller, D. Hong, C.J. Bell et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460:** 1011-1015.

Kirkegaard, I., O.B. Petersen, N. Uldbjerg, and N. Torring. 2008. Improved performance of first-trimester combined screening for trisomy 21 with the double test taken before a gestational age of 10 weeks. *Prenat Diagn* **28:** 839-844.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lau, T.W., T.N. Leung, L.Y. Chan, T.K. Lau, K.C. Chan, W.H. Tam, and Y.M. Lo. 2002. Fetal DNA clearance from maternal plasma is impaired in preeclampsia. *Clin Chem* **48:** 2141-2146.

Lazar, L., B. Nagy, Z. Ban, G.R. Nagy, and Z. Papp. 2006. Presence of cell-free fetal DNA in plasma of women with ectopic pregnancies. *Clin Chem* **52:** 1599-1601.

Leary, J., I.K., Frank Diehl, Kerstin Schmidt, Chris Clouser, Cisilya Duncan, Alena Antipova, Clarence Lee, Kevin McKernan, Francisco M. De La Vega, Kenneth W. Kinzler, Bert Vogelstein, Luis A. Diaz Jr. and Victor E. Velculescu 2010. Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing. *Sci Transl Med 24 February 2010* **Vol. 2, Issue 20, p. 20ra14**

Leon, S.A., G.E. Ehrlich, B. Shapiro, and V.A. Labbate. 1977a. Free DNA in the serum of rheumatoid arthritis patients. *J Rheumatol* **4:** 139-143.

Leon, S.A., B. Shapiro, D.M. Sklaroff, and M.J. Yaros. 1977b. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* **37:** 646-650.

Leung, T.N., J. Zhang, T.K. Lau, L.Y. Chan, and Y.M. Lo. 2001. Increased maternal plasma fetal DNA concentrations in women who eventually develop preeclampsia. *Clin Chem* **47:** 137-139.

Leung, T.N., J. Zhang, T.K. Lau, N.M. Hjelm, and Y.M. Lo. 1998. Maternal plasma fetal DNA as a marker for preterm labour. *Lancet* **352:** 1904-1905.

Levine, R.J., C. Qian, E.S. Leshane, K.F. Yu, L.J. England, E.F. Schisterman, T. Wataganara, R. Romero, and D.W. Bianchi. 2004. Two-stage elevation of cell-free fetal DNA in maternal sera before onset of preeclampsia. *Am J Obstet Gynecol* **190:** 707-713.

Levy, S., G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254.

Ley, T.J., E.R. Mardis, L. Ding, B. Fulton, M.D. McLellan, K. Chen, D. Dooling, B.H. Dunford-Shore, S. McGrath, M. Hickenbotham et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66-72.

Li, E. 2002. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3:** 662-673.

Li, J.Z. and C.R. Steinman. 1989. Plasma DNA in systemic lupus erythematosus. Characterization of cloned base sequences. *Arthritis Rheum* **32:** 726-733.

Li, Y., E. Di Naro, A. Vitucci, B. Zimmermann, W. Holzgreve, and S. Hahn. 2005. Detection of paternally inherited fetal point mutations for beta-thalassemia using size-fractionated cell-free DNA in maternal plasma. *JAMA* **293:** 843-849.

Li, Y., W. Holzgreve, G.C. Page-Christiaens, J.J. Gille, and S. Hahn. 2004a. Improved prenatal detection of a fetal point mutation for achondroplasia by the use of size-fractionated circulatory DNA in maternal plasma--case report. *Prenat Diagn* **24:** 896-898.

Li, Y., B. Zimmermann, C. Rusterholz, A. Kang, W. Holzgreve, and S. Hahn. 2004b. Size separation of circulatory DNA in maternal plasma permits ready detection of fetal DNA polymorphisms. *Clin Chem* **50:** 1002-1011.

Liu, C.J., S.Y. Kao, H.F. Tu, M.M. Tsai, K.W. Chang, and S.C. Lin. 2010. Increase of microRNA miR-31 level in plasma could be a potential marker of oral cancer. *Oral Dis* **16:** 360-364.

Liu, F., L.M. Sholienberger, C.C. Conwell, X. Yuan, and L. Huang. 2007. Mechanism of naked DNA clearance after intravenous injection. *J Gene Med* **9:** 613-619.

Lo, Y.M. and R.W. Chiu. 2007. Prenatal diagnosis: progress through plasma nucleic acids. *Nat Rev Genet* **8:** 71-77.

Lo, Y.M. 2001. Circulating nucleic acids in plasma and serum: an overview. *Ann N Y Acad Sci* **945:** 1-7.

Lo, Y.M. and R.W. Chiu. 2009. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. *Clin Chem* **55:** 607-608.

Lo, Y.M., N. Corbetta, P.F. Chamberlain, V. Rai, I.L. Sargent, C.W. Redman, and J.S. Wainscoat. 1997. Presence of fetal DNA in maternal plasma and serum. *Lancet* **350:** 485-487.

Lo, Y.M., N.M. Hjelm, C. Fidler, I.L. Sargent, M.F. Murphy, P.F. Chamberlain, P.M. Poon, C.W. Redman, and J.S. Wainscoat. 1998a. Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *N Engl J Med* **339:** 1734-1738.

Lo, Y.M., T.K. Lau, J. Zhang, T.N. Leung, A.M. Chang, N.M. Hjelm, R.S. Elmes, and D.W. Bianchi. 1999a. Increased fetal DNA concentrations in the plasma of pregnant women carrying fetuses with trisomy 21. *Clin Chem* **45:** 1747-1751.

Lo, Y.M., T.N. Leung, M.S. Tein, I.L. Sargent, J. Zhang, T.K. Lau, C.J. Haines, and C.W. Redman. 1999b. Quantitative abnormalities of fetal DNA in maternal serum in preeclampsia. *Clin Chem* **45:** 184-188.

Lo, Y.M., F.M. Lun, K.C. Chan, N.B. Tsui, K.C. Chong, T.K. Lau, T.Y. Leung, B.C. Zee, C.R. Cantor, and R.W. Chiu. 2007a. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc Natl Acad Sci U S A* **104:** 13116-13121.

Lo, Y.M., L. Noakes, E. Roux, M. Jeannet, B. Chapuis, K.A. Fleming, and J.S. Wainscoat. 1995. Application of a polymorphic Y microsatellite to the detection of post bone marrow transplantation chimaerism. *Br J Haematol* **89:** 645-649.

Lo, Y.M., E. Roux, M. Jeannet, B. Chapuis, K.A. Fleming, and J.S. Wainscoat. 1993. Detection of chimaerism after bone marrow transplantation using the double amplification refractory mutation system. *Br J Haematol* **85:** 223-226.

Lo, Y.M., M.S. Tein, T.K. Lau, C.J. Haines, T.N. Leung, P.M. Poon, J.S. Wainscoat, P.J. Johnson, A.M. Chang, and N.M. Hjelm. 1998b. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am J Hum Genet* **62:** 768-775.

Lo, Y.M., M.S. Tein, C.C. Pang, C.K. Yeung, K.L. Tong, and N.M. Hjelm. 1998c. Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. *Lancet* **351:** 1329-1330.

Lo, Y.M., N.B. Tsui, R.W. Chiu, T.K. Lau, T.N. Leung, M.M. Heung, A. Gerovassili, Y. Jin, K.H. Nicolaides, C.R. Cantor et al. 2007b. Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. *Nat Med* **13:** 218-223.

Lo, Y.M., J. Zhang, T.N. Leung, T.K. Lau, A.M. Chang, and N.M. Hjelm. 1999c. Rapid clearance of fetal DNA from maternal plasma. *Am J Hum Genet* **64:** 218-224.

Luger, K. 2003. Structure and dynamic behavior of nucleosomes. *Curr Opin Genet Dev* **13:** 127-135.

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

Lui, Y.Y., K.S. Woo, A.Y. Wang, C.K. Yeung, P.K. Li, E. Chau, P. Ruygrok, and Y.M. Lo. 2003. Origin of plasma cell-free DNA after solid organ transplantation. *Clin Chem* **49:** 495-496.

Lun, F.M., R.W. Chiu, K.C. Allen Chan, T. Yeung Leung, T. Kin Lau, and Y.M. Dennis Lo. 2008a. Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. *Clin Chem* **54:** 1664-1672.

Lun, F.M., N.B. Tsui, K.C. Chan, T.Y. Leung, T.K. Lau, P. Charoenkwan, K.C. Chow, W.Y. Lo, C. Wanapirak, T. Sanguansermsri et al. 2008b. Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. *Proc Natl Acad Sci U S A* **105:** 19920-19925.

Lund, J. and B.A. Parviz. 2009. Scanning probe and nanopore DNA sequencing: core techniques and possibilities. *Methods Mol Biol* **578:** 113-122.

Luo, S.S., O. Ishibashi, G. Ishikawa, T. Ishikawa, A. Katayama, T. Mishima, T. Takizawa, T. Shigihara, T. Goto, A. Izumi et al. 2009. Human villous trophoblasts express and secrete placenta-specific microRNAs into maternal circulation via exosomes. *Biol Reprod* **81:** 717-729.

Lupski, J.R., J.G. Reid, C. Gonzaga-Jauregui, D. Rio Deiros, D.C. Chen, L. Nazareth, M. Bainbridge, H. Dinh, C. Jing, D.A. Wheeler et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362:** 1181-1191.

Lyon, M.F. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190:** 372-373.

Madabhushi, R.S. 1998. Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* **19:** 224-230.

Maekawa, M., T. Taniguchi, H. Higashi, H. Sugimura, K. Sugano, and T. Kanno. 2004. Methylation of mitochondrial DNA is not a useful marker for cancer detection. *Clin Chem* **50:** 1480-1481.

Maher, C.A., N. Palanisamy, J.C. Brenner, X. Cao, S. Kalyana-Sundaram, S. Luo, I. Khrebtukova, T.R. Barrette, C. Grasso, J. Yu et al. 2009. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* **106:** 12353-12358.

Majer, S., M. Bauer, E. Magnet, A. Strele, E. Giegerl, M. Eder, U. Lang, and B. Pertl. 2007. Maternal urine for prenatal diagnosis--an analysis of cell-free fetal DNA in maternal urine and plasma in the third trimester. *Prenat Diagn* **27:** 1219-1223.

Mamanova, L., A.J. Coffey, C.E. Scott, I. Kozarewa, E.H. Turner, A. Kumar, E. Howard, J. Shendure, and D.J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7:** 111-118.

Mandavilli, B.S., J.H. Santos, and B. Van Houten. 2002. Mitochondrial DNA repair and aging. *Mutat Res* **509:** 127-151.

Mandel, P. and P. Metais. 1948. Les acides nucleiques du plasma sanguin chez l'homme. *C R Seances Soc Biol Fil* **142:** 241-243.

Mardis, E.R. 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med* **1:** 40.

Mardis, E.R., L. Ding, D.J. Dooling, D.E. Larson, M.D. McLellan, K. Chen, D.C. Koboldt, R.S. Fulton, K.D. Delehaunty, S.D. McGrath et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361:** 1058-1066.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Masuzaki, H., K. Miura, K. Yamasaki, S. Miura, K. Yoshiura, S. Yoshimura, D. Nakayama, C.K. Mapendano, N. Niikawa, and T. Ishimaru. 2005. Clinical applications of plasma circulating mRNA analysis in cases of gestational trophoblastic disease. *Clin Chem* **51:** 1261-1263.

Masuzaki, H., K. Miura, K.I. Yoshiura, S. Yoshimura, N. Niikawa, and T. Ishimaru. 2004. Detection of cell free placental mosaicismplacental DNA in maternal plasma: direct evidence from three cases of confined *J Med Genet* **41:** 289-292.

May, A. and V.A. Bohr. 2000. Gene-specific repair of gamma-ray-induced DNA strand breaks in colon cancer cells: no coupling to transcription and no removal from the mitochondrial genome. *Biochem Biophys Res Commun* **269:** 433-437.

McKernan, K.J., H.E. Peckham, G.L. Costa, S.F. McLaughlin, Y. Fu, E.F. Tsung, C.R. Clouser, C. Duncan, J.K. Ichikawa, C.C. Lee et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19:** 1527-1541.

Mennuti, M.T. and D.A. Driscoll. 2003. Screening for Down's syndrome--too many choices? *N Engl J Med* **349:** 1471-1473.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Res* **15:** 1767-1776.

Metzker, M.L. 2009. Sequencing technologies - the next generation. *Nat Rev Genet* **11:** 31-46.
Meyer, M., A.W. Briggs, T. Maricic, B. Hober, B. Hoffner, J. Krause, A. Weihmann, S. Paabo, and M. Hofreiter. 2008. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res* **36:** e5.
Meyer, M., U. Stenzel, S. Myles, K. Prufer, and M. Hofreiter. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* **35:** e97.
Mitchell, P.S., R.K. Parkin, E.M. Kroh, B.R. Fritz, S.K. Wyman, E.L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K.C. O'Briant, A. Allen et al. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* **105:** 10513-10518.
Morozova, O. and M.A. Marra. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92:** 255-264.
Nagata, S., H. Nagase, K. Kawane, N. Mukae, and H. Fukuyama. 2003. Degradation of chromosomal DNA during apoptosis. *Cell Death Differ* **10:** 108-116.
Nelson, M., C. Eagle, M. Langshaw, H. Popp, and H. Kronenberg. 2001. Genotyping fetal DNA by non-invasive means: extraction from maternal plasma. *Vox Sang* **80:** 112-116.
Ng, E.K., A. El-Sheikhah, R.W. Chiu, K.C. Chan, M. Hogg, R. Bindra, T.N. Leung, T.K. Lau, K.H. Nicolaides, and Y.M. Lo. 2004. Evaluation of human chorionic gonadotropin beta-subunit mRNA concentrations in maternal serum in aneuploid pregnancies: a feasibility study. *Clin Chem* **50:** 1055-1057.
Ng, E.K., T.N. Leung, N.B. Tsui, T.K. Lau, N.S. Panesar, R.W. Chiu, and Y.M. Lo. 2003a. The concentration of circulating corticotropin-releasing hormone mRNA in maternal plasma is increased in preeclampsia. *Clin Chem* **49:** 727-731.
Ng, E.K., N.B. Tsui, N.Y. Lam, R.W. Chiu, S.C. Yu, S.C. Wong, E.S. Lo, T.H. Rainer, P.J. Johnson, and Y.M. Lo. 2002. Presence of filterable and nonfilterable mRNA in the plasma of cancer patients and healthy individuals. *Clin Chem* **48:** 1212-1217.
Ng, E.K., N.B. Tsui, T.K. Lau, T.N. Leung, R.W. Chiu, N.S. Panesar, L.C. Lit, K.W. Chan, and Y.M. Lo. 2003b. mRNA of placental origin is readily detectable in maternal plasma. *Proc Natl Acad Sci U S A* **100:** 4748-4753.
Ng, S.B., E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E.E. Eichler et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272-276.
Nicolaides, K.H., G. Azar, D. Byrne, C. Mansur, and K. Marks. 1992. Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy. *BMJ* **304:** 867-869.
Nusbaum, C., M.C. Zody, M.L. Borowsky, M. Kamal, C.D. Kodira, T.D. Taylor, C.A. Whittaker, J.L. Chang, C.A. Cuomo, K. Dewar et al. 2005. DNA sequence and analysis of human chromosome 18. *Nature* **437:** 551-555.
Nyren, P., B. Pettersson, and M. Uhlen. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem* **208:** 171-175.
Ohashi, Y., N. Miharu, H. Honda, O. Samura, and K. Ohama. 2002. Correlation of fetal DNA and human chorionic gonadotropin concentrations in second-trimester maternal serum. *Clin Chem* **48:** 386-388.
Okochi, O., K. Hibi, T. Uemura, S. Inoue, S. Takeda, T. Kaneko, and A. Nakao. 2002. Detection of mitochondrial DNA alterations in the serum of hepatocellular carcinoma patients. *Clin Cancer Res* **8:** 2875-2878.
Old, R.W., F. Crea, W. Puszyk, and M.A. Hulten. 2007. Candidate epigenetic biomarkers for non-invasive prenatal diagnosis of Down syndrome. *Reprod Biomed Online* **15:** 227-235.
Orozco, A.F., F.Z. Bischoff, C. Horne, E. Popek, J.L. Simpson, and D.E. Lewis. 2006. Hypoxia-induced membrane-bound apoptotic DNA particles: potential mechanism of fetal DNA in maternal plasma. *Ann N Y Acad Sci* **1075:** 57-62.
Palacios, G., J. Druce, L. Du, T. Tran, C. Birch, T. Briese, S. Conlan, P.L. Quan, J. Hui, J. Marshall et al. 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med* **358:** 991-998.
Papageorgiou, E.A., H. Fiegler, V. Rakyan, S. Beck, M. Hulten, K. Lamnissou, N.P. Carter, and P.C. Patsalis. 2009. Sites of differential DNA methylation between placenta and peripheral blood: molecular markers for noninvasive prenatal diagnosis of aneuploidies. *Am J Pathol* **174:** 1609-1618.

Pettersson, E., J. Lundeberg, and A. Ahmadian. 2009. Generations of sequencing technologies. *Genomics* **93:** 105-111.

Plath, K., S. Mlynarczyk-Evans, D.A. Nusinow, and B. Panning. 2002. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet* **36:** 233-278.

Pohl, G. and M. Shih Ie. 2004. Principle and applications of digital PCR. *Expert Rev Mol Diagn* **4:** 41-47.

Poon, L.L., T.N. Leung, T.K. Lau, K.C. Chow, and Y.M. Lo. 2002. Differential DNA methylation between fetus and mother as a strategy for detecting fetal DNA in maternal plasma. *Clin Chem* **48:** 35-41.

Poon, L.L., T.N. Leung, T.K. Lau, and Y.M. Lo. 2000. Presence of fetal RNA in maternal plasma. *Clin Chem* **46:** 1832-1834.

Prober, J.M., G.L. Trainor, R.J. Dam, F.W. Hobbs, C.W. Robertson, R.J. Zagursky, A.J. Cocuzza, M.A. Jensen, and K. Baumeister. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238:** 336-341.

Pushkarev, D., N.F. Neff, and S.R. Quake. 2009. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27:** 847-852.

Puszyk, W.M., F. Crea, and R.W. Old. 2009. Unequal representation of different unique genomic DNA sequences in the cell-free plasma DNA of individual donors. *Clin Biochem* **42:** 736-738.

Qin, J., R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59-65.

Quail, M.A., I. Kozarewa, F. Smith, A. Scally, P.J. Stephens, R. Durbin, H. Swerdlow, and D.J. Turner. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5:** 1005-1010.

Rijnders, R.J., C.E. van der Schoot, B. Bossers, M.A. de Vroede, and G.C. Christiaens. 2001. Fetal sex determination from maternal plasma in pregnancies at risk for congenital adrenal hyperplasia. *Obstet Gynecol* **98:** 374-378.

Roach, J.C., G. Glusman, A.F. Smit, C.D. Huff, R. Hubley, P.T. Shannon, L. Rowen, K.P. Pant, N. Goodman, M. Bamshad et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328:** 636-639.

Roberts, J.M. and C.W. Redman. 1993. Pre-eclampsia: more than pregnancy-induced hypertension. *Lancet* **341:** 1447-1451.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242:** 84-89.

Ronaghi, M., M. Uhlen, and P. Nyren. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281:** 363, 365.

Rosl, F. 1992. A simple and rapid method for detection of apoptosis in human cells. *Nucleic Acids Res* **20:** 5243.

Saito, H., A. Sekizawa, T. Morimoto, M. Suzuki, and T. Yanaihara. 2000. Prenatal DNA diagnosis of a single-gene disorder from maternal plasma. *Lancet* **356:** 1170.

Saller, D.N., Jr. and J.A. Canick. 2008. Current methods of prenatal screening for Down syndrome and other fetal abnormalities. *Clin Obstet Gynecol* **51:** 24-36.

Sanger, F. and A.R. Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94:** 441-448.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Santacroce, R., G. Vecchione, M. Tomaiyolo, F. Sessa, M. Sarno, D. Colaizzo, E. Grandone, and M. Margaglione. 2006. Identification of fetal gender in maternal blood is a helpful tool in the prenatal diagnosis of haemophilia. *Haemophilia* **12:** 417-422.

Satoh, M. and T. Kuroiwa. 1991. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp Cell Res* **196:** 137-140.

Scheffer, P.G., C.E. van der Schoot, G.C. Page-Christiaens, B. Bossers, F. van Erp, and M. de Haas. 2009. Reliability of fetal sex determination using maternal plasma. *Obstet Gynecol* **115:** 117-126.

Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat Methods* **5:** 16-18.

Schuster, S.C., W. Miller, A. Ratan, L.P. Tomsho, B. Giardine, L.R. Kasson, R.S. Harris, D.C. Petersen, F. Zhao, J. Qi et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463:** 943-947.

Segal, E., Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, and J. Widom. 2006. A genomic code for nucleosome positioning. *Nature* **442:** 772-778.

Sekizawa, A., M. Jimbo, H. Saito, M. Iwasaki, Y. Sugito, Y. Yukimoto, J. Otsuka, and T. Okai. 2002. Increased cell-free fetal DNA in plasma of two women with invasive placenta. *Clin Chem* **48:** 353-354.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728-1732.

Shi, W.J., Y.H. Sun, L.J. Cui, Y. Zhang, and S.R. Li. 2003. [Detection of cell-free fetal DNA in the urine of pregnant women by nested polymerase chain reaction]. *Zhonghua Yi Xue Za Zhi* **83:** 482-484.

Shimada, K., K. Murakami, M. Shozu, T. Segawa, H. Sumitani, and M. Inoue. 2004. Sex-determining region Y levels in maternal plasma: Evaluation in abnormal pregnancy. *J Obstet Gynaecol Res* **30:** 148-154.

Shmookler Reis, R.J. and S. Goldstein. 1983. Mitochondrial DNA in mortal and immortal human cells. Genome number, integrity, and methylation. *J Biol Chem* **258:** 9078-9085.

Sikora, A., B.G. Zimmermann, C. Rusterholz, D. Birri, V. Kolla, O. Lapaire, I. Hoesli, V. Kiefer, L. Jackson, and S. Hahn. 2009. Detection of increased amounts of cell-free fetal DNA with short PCR amplicons. *Clin Chem* **56:** 136-138.

Smid, M., S. Galbiati, A. Vassallo, D. Gambini, A. Ferrari, E. Viora, M. Pagliano, G. Restagno, M. Ferrari, and L. Cremonesi. 2003. No evidence of fetal DNA persistence in maternal plasma after pregnancy. *Hum Genet* **112:** 617-618.

Smid, M., A. Vassallo, F. Lagona, L. Valsecchi, L. Maniscalco, L. Danti, A. Lojacono, A. Ferrari, M. Ferrari, and L. Cremonesi. 2001. Quantitative analysis of fetal DNA in maternal plasma in pathological conditions associated with placental abnormalities. *Ann N Y Acad Sci* **945:** 132-137.

Smith, L.M., J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B. Kent, and L.E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321:** 674-679.

Smith, S.C., P.N. Baker, and E.M. Symonds. 1997. Placental apoptosis in normal human pregnancy. *Am J Obstet Gynecol* **177:** 57-65.

Sorenson, G.D., D.M. Pribish, F.H. Valone, V.A. Memoli, D.J. Bzik, and S.L. Yao. 1994. Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol Biomarkers Prev* **3:** 67-71.

Staley, K., A.J. Blaschke, and J. Chun. 1997. Apoptotic DNA fragmentation is detected by a semi-quantitative ligation-mediated PCR of blunt DNA ends. *Cell Death Differ* **4:** 66-75.

Stenhouse, E.J., J.A. Crossley, D.A. Aitken, K. Brogan, A.D. Cameron, and J.M. Connor. 2004. First-trimester combined ultrasound and biochemical screening for Down syndrome in routine clinical practice. *Prenat Diagn* **24:** 774-780.

Stroun, M., P. Anker, J. Lyautey, C. Lederrey, and P.A. Maurice. 1987. Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol* **23:** 707-712.

Stroun, M., P. Anker, P. Maurice, J. Lyautey, C. Lederrey, and M. Beljanski. 1989. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46:** 318-322.

Stroun, M., J. Lyautey, C. Lederrey, H.E. Mulcahy, and P. Anker. 2001a. Alu repeat sequences are present in increased proportions compared to a unique gene in plasma/serum DNA: evidence for a preferential release from viable cells? *Ann N Y Acad Sci* **945:** 258-264.

Stroun, M., J. Lyautey, C. Lederrey, A. Olson-Sand, and P. Anker. 2001b. About the possible origin and mechanism of circulating DNA apoptosis and active DNA release. *Clin Chim Acta* **313:** 139-142.

Stroun, M., P. Maurice, V. Vasioukhin, J. Lyautey, C. Lederrey, F. Lefort, A. Rossier, X.Q. Chen, and P. Anker. 2000. The origin and mechanism of circulating DNA. *Ann N Y Acad Sci* **906:** 161-168.

Suzuki, N., A. Kamataki, J. Yamaki, and Y. Homma. 2008. Characterization of circulating DNA in healthy human plasma. *Clin Chim Acta* **387:** 55-58.

Swinkels, D.W., J.B. de Kok, J.C. Hendriks, E. Wiegerinck, P.L. Zusterzeel, and E.A. Steegers. 2002. Hemolysis, elevated liver enzymes, and low platelet count (HELLP) syndrome as a complication of preeclampsia in pregnant women increases the amount of cell-free fetal and maternal DNA in maternal plasma and serum. *Clin Chem* **48:** 650-653.

Tamkovich, S.N., A.V. Cherepanova, E.V. Kolesnikova, E.Y. Rykova, D.V. Pyshnyi, V.V. Vlassov, and P.P. Laktionov. 2006. Circulating DNA and DNase activity in human blood. *Ann N Y Acad Sci* **1075:** 191-196.

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

Tjoa, M.L., T. Cindrova-Davies, O. Spasic-Boskovic, D.W. Bianchi, and G.J. Burton. 2006. Trophoblastic oxidative stress and the release of cell-free feto-placental DNA. *Am J Pathol* **169:** 400-404.

Tong, Y.K., C. Ding, R.W. Chiu, A. Gerovassili, S.S. Chim, T.Y. Leung, T.N. Leung, T.K. Lau, K.H. Nicolaides, and Y.M. Lo. 2006. Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations. *Clin Chem* **52:** 2194-2202.

Tong, Y.K., S. Jin, R.W. Chiu, C. Ding, K.C. Chan, T.Y. Leung, L. Yu, T.K. Lau, and Y.M. Lo. 2009. Noninvasive prenatal detection of trisomy 21 by an epigenetic-genetic chromosome-dosage approach. *Clin Chem* **56:** 90-98.

Tsui, N.B., S.S. Chim, R.W. Chiu, T.K. Lau, E.K. Ng, T.N. Leung, Y.K. Tong, K.C. Chan, and Y.M. Lo. 2004. Systematic micro-array based identification of placental mRNA in maternal plasma: towards non-invasive prenatal gene expression profiling. *J Med Genet* **41:** 461-467.

Tsui, N.B., E.K. Ng, and Y.M. Lo. 2002. Stability of endogenous and added RNA in blood specimens, serum, and plasma. *Clin Chem* **48:** 1647-1653.

Tsui, N.B., B.C. Wong, T.Y. Leung, T.K. Lau, R.W. Chiu, and Y.M. Lo. 2009. Non-invasive prenatal detection of fetal trisomy 18 by RNA-SNP allelic ratio analysis using maternal plasma SERPINB2 mRNA: a feasibility study. *Prenat Diagn* **29:** 1031-1037.

Tsumita, T. and M. Iwanaga. 1963. Fate of injected deoxyribonucleic acid in mice. *Nature* **198:** 1088-1089.

Turcatti, G., A. Romieu, M. Fedurco, and A.P. Tairi. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **36:** e25.

Turnbaugh, P.J., M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit et al. 2009. A core gut microbiome in obese and lean twins. *Nature* **457:** 480-484.

Umbach, J.L., M.F. Kramer, I. Jurak, H.W. Karnowski, D.M. Coen, and B.R. Cullen. 2008. MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature* **454:** 780-783.

Umetani, N., A.E. Giuliano, S.H. Hiramatsu, F. Amersi, T. Nakagawa, S. Martino, and D.S. Hoon. 2006a. Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J Clin Oncol* **24:** 4270-4276.

Umetani, N., J. Kim, S. Hiramatsu, H.A. Reber, O.J. Hines, A.J. Bilchik, and D.S. Hoon. 2006b. Increased integrity of free circulating DNA in sera of patients with colorectal or periampullary cancer: direct quantitative PCR for ALU repeats. *Clin Chem* **52:** 1062-1069.

Van der Schoot, C.E., A.A. Soussan, J. Koelewijn, G. Bonsel, L.G. Paget-Christiaens, and M. de Haas. 2006. Non-invasive antenatal RHD typing. *Transfus Clin Biol* **13:** 53-57.

van der Vaart, M. and P.J. Pretorius. 2007. The origin of circulating free DNA. *Clin Chem* **53:** 2215.

van der Vaart, M. and P.J. Pretorius. 2008a. Circulating DNA. Its origin and fluctuation. *Ann N Y Acad Sci* **1137:** 18-26.

van der Vaart, M. and P.J. Pretorius. 2008b. A method for characterization of total circulating DNA. *Ann N Y Acad Sci* **1137:** 92-97.

Vasilescu, C., S. Rossi, M. Shimizu, S. Tudor, A. Veronese, M. Ferracin, M.S. Nicoloso, E. Barbarotto, M. Popa, O. Stanciulea et al. 2009. MicroRNA fingerprints identify miR-150 as a plasma prognostic marker in patients with sepsis. *PLoS One* **4:** e7405.

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

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Visigen. 2009. Visigen sequencing_PPT.

Voelkerding, K.V., S.A. Dames, and J.D. Durtschi. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* **55:** 641-658.

Wallace, D.C. 1999. Mitochondrial diseases in man and mouse. *Science* **283:** 1482-1488.

Wang, B.G., H.Y. Huang, Y.C. Chen, R.E. Bristow, K. Kassauei, C.C. Cheng, R. Roden, L.J. Sokoll, D.W. Chan, and M. Shih Ie. 2003. Increased plasma DNA integrity in cancer patients. *Cancer Res* **63:** 3966-3968.

Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R.W. Shafer. 2007a. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17:** 1195-1201.

Wang, G.P., A. Ciuffi, J. Leipzig, C.C. Berry, and F.D. Bushman. 2007b. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17:** 1186-1194.

Wang, J., W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, J. Zhang, J. Li, Y. Guo et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456:** 60-65.

Wataganara, T., E.S. LeShane, A. Farina, G.M. Messerlian, T. Lee, J.A. Canick, and D.W. Bianchi. 2003. Maternal serum cell-free fetal DNA levels are increased in cases of trisomy 13 but not trisomy 18. *Hum Genet* **112:** 204-208.

Wekerle, T., J. Kurtz, M. Sayegh, H. Ito, A. Wells, S. Bensinger, J. Shaffer, L. Turka, and M. Sykes. 2001. Peripheral deletion after bone marrow transplantation with costimulatory blockade has features of both activation-induced cell death and passive cell death. *J Immunol* **166:** 2311-2316.

Wheeler, D.A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.J. Chen, V. Makhijani, G.T. Roth et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452:** 872-876.

Widlak, P., P. Li, X. Wang, and W.T. Garrard. 2000. Cleavage preferences of the apoptotic endonuclease DFF40 (caspase-activated DNase or nuclease) on naked DNA and chromatin substrates. *J Biol Chem* **275:** 8226-8232.

Widom, J. 1992. A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc Natl Acad Sci U S A* **89:** 1095-1099.

Williams, R., S.G. Peisajovich, O.J. Miller, S. Magdassi, D.S. Tawfik, and A.D. Griffiths. 2006. Amplification of complex gene libraries by emulsion PCR. *Nat Methods* **3:** 545-550.

Wu, J.Q., L. Habegger, P. Noisa, A. Szekely, C. Qiu, S. Hutchison, D. Raha, M. Egholm, H. Lin, S. Weissman et al. 2010. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc Natl Acad Sci U S A* **107:** 5254-5259.

Wyllie, A.H. 1980. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation. *Nature* **284:** 555-556.

Xu, M., D. Fujita, and N. Hanagata. 2009. Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small* **5:** 2638-2649.

Zeschnigk, M., M. Martin, G. Betzl, A. Kalbe, C. Sirsch, K. Buiting, S. Gross, E. Fritzilas, B. Frey, S. Rahmann et al. 2009. Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. *Hum Mol Genet* **18:** 1439-1448.

Zhong, X.Y., M.R. Burk, C. Troeger, L.R. Jackson, W. Holzgreve, and S. Hahn. 2000. Fetal DNA in maternal plasma is elevated in pregnancies with aneuploid fetuses. *Prenat Diagn* **20:** 795-798.

Zhong, X.Y., W. Holzgreve, and S. Hahn. 2002. The levels of circulatory cell free fetal DNA in maternal plasma are elevated prior to the onset of preeclampsia. *Hypertens Pregnancy* **21:** 77-83.

Zhong, X.Y., H. Laivuori, J.C. Livingston, O. Ylikorkala, B.M. Sibai, W. Holzgreve, and S. Hahn. 2001. Elevation of both maternal and fetal extracellular circulating

deoxyribonucleic acid concentrations in the plasma of pregnant women with preeclampsia. *Am J Obstet Gynecol* **184:** 414-419.

Zimmermann, B.G., S. Grill, W. Holzgreve, X.Y. Zhong, L.G. Jackson, and S. Hahn. 2008. Digital PCR: a powerful new tool for noninvasive prenatal diagnosis? *Prenat Diagn* **28:** 1087-1093.