Expanding the accuracy, resolution, and breadth of cell-free DNA investigation

Matthew William Snyder

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Jay Shendure, Chair

Phil Green

Maynard Olson

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Expanding the accuracy, resolution, and breadth
of cell-free DNA investigation

Matthew William Snyder

Chair of the Supervisory Committee:
Dr. Jay Shendure
Department of Genome Sciences

When cells die, they don't simply vanish without a trace. Instead, they leave behind fingerprints of their genetic and epigenetic identities in the form of cell-free DNA (cfDNA), or the scant amount of highly fragmented DNA circulating in human plasma. As the detritus of apoptotic and necrotic cell death in multiple tissues throughout the body, this class of molecule serves as a powerful biomarker for noninvasive detection and monitoring of disease processes and physiological conditions, including pregnancy, organ transplantation, and a growing number of cancers.

Despite this promise, current methods for interrogating cfDNA are challenged by limited resolution, imperfect accuracy, and constrained breadth. Taken as a whole, these factors restrict the set of conditions that might in principle be detected or monitored with this molecular evidence. In this dissertation, I directly address these limitations with the goal of expanding the scope and precision of the "liquid biopsy," or the noninvasive monitoring of health status through cfDNA analysis.

I first address the limited resolution of cfDNA testing in the context of pregnancy by developing statistical methods for inference of the entire fetal genome at the single nucleotide level, including both inherited and *de novo* variation. I show that the use of parental haplotypes and maternal cfDNA in a hidden Markov model can yield highly accurate prediction

of inherited fetal genotypes. I next determine that the length of parental haplotype blocks is a key parameter driving the prediction accuracy, and demonstrate a method for increasing block length and downstream inference. I explain how these approaches, coupled with improved methods for detection of *de novo* variation, open the door to a single, noninvasive test with the possibility of prenatal detection of more than 3,000 highly penetrant single-gene disorders.

I next demonstrate a method for improving the accuracy and positive predictive value (PPV) of noninvasive screening for fetal aneuploidy. In the most popular screening methodologies, PPVs of cfDNA-based tests are limited by a combination of the low incidence of trisomic pregnancies and the small number of false-positive tests in which truly euploid pregnancies are incorrectly classified as aneuploid. I investigate the causes of false-positive test results in a small cohort, and determine that maternal copy-number variants (CNVs) substantially contributes to the burden of spurious findings. I further develop a statistical framework for quantifying the likely impact of maternal CNVs by size and tested chromosome. I then propose a straightforward method for addressing this analytical limitation and improving test accuracy.

Finally, I develop a new approach to disentangle the various tissues or cell types contributing to cfDNA in a biological sample, potentially expanding the breadth of physiological conditions that can be monitored in this way. Here, I show that the locations of cfDNA fragment endpoints evidence the positions of proteins on the DNA *in vivo* in the contributing cells, and use these endpoints to infer the spacing of nucleosomes and transcription factors genome-wide. I demonstrate that these positions correlate with gene expression profiles, and use these data to model cell type contributions in healthy individuals, where the expected myeloid and lymphoid cell lineages are recovered. I then apply this analytical framework to a cohort of individuals with advanced cancers, and recover the tissue-of-origin of the primary tumor for a subset of the cancers.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Chapter 1: Introduction 1
1.1 Detritus of cell death: the origins of cell-free DNA 1
1.2 Prenatal testing with cell-free DNA 3
1.3 The "liquid biopsy" for cancer 7
1.4 Noninvasive monitoring of allograft health 9
1.5 Tissues of origin: expanding the breadth of cfDNA-based testing 10
1.6 Organization of this dissertation 13

Chapter 2: Noninvasive whole genome sequencing of a human fetus 15
2.1 Introduction 17
2.2 Results 19
2.3 Discussion 32
2.4 Methods 37

Chapter 3: Copy-number variation and false positives in prenatal aneuploidy screening 48
3.1 Introduction 49
3.2 Results 51
3.3 Discussion 58
3.4 Methods 62

Chapter 4: Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin 77
4.1 Introduction 78

4.2 Results 80
4.3 Discussion 98
4.4 Methods 105

Chapter 5: Conclusions and future directions 125
5.1 Noninvasive prenatal screening 125
5.2 Noninvasive monitoring of cancer and other diseases 134
5.3 Closing thoughts 142

LIST OF FIGURES

Figure Number Page

2.1 Schematic of dense haplotyping 18
2.2 Schematic of experimental approach 20
2.3 Inference of the fetal genome on a site-by-site basis 22
2.4 Accuracy of fetal genotype inference from maternal plasma DNA sequencing 23
2.5 Prediction of maternally transmitted alleles 24
2.6 Hidden Markov model-based detection of recombination events and haplotype assembly switch errors 25
2.7 Accuracy of maternal transmission inference, by haplotype block size 26
2.8 Improving density and contiguity of haplotype blocks with reference panels 28
2.9 Impact of coverage, haplotype length, and fetal fraction on statistical inference 29
2.10 Inference accuracy for paternal transmission at paternal-only heterozygous sites, by depth 30
2.11 Detecting sites of *de novo* mutation among maternal fetal plasma sequences 31
2.12 SVM-based classification of candidate *de novo* mutations 32
2.13 Inference of the fetal genome from haplotype blocks 34

3.1 Schematic representation of cfDNA analysis 51
3.2 Maternal CNVs and sampling distributions 52
3.3 Copy number profile in pregnancy cohort 53
3.4 Validation of CNVs with multiplex PCR 54
3.5 Structure and validation of maternal duplication in Patient 1 55
3.6 Structure and validation of maternal duplication in Patient 3 56
3.7 Population frequency and estimated impact on false-positive test rates of maternal CNVs 57
3.8 Population frequency and estimated impact on false-positive test rates of maternal CNVs with less stringent filtering 60

4.1 Schematic overview of cfDNA fragmentation 78
4.2 Characteristics of conventional cell-free DNA sequencing libraries 80

4.3 Schematic of single-stranded library preparation protocol 82
4.4 Characteristics of single-stranded cell-free DNA sequencing libraries 83
4.5 Schematic of inference of nucleosome positioning 84
4.6 Strongly positioned nucleosomes at a well-studied alpha-satellite array . . . 85
4.7 Inferred nucleosome positioning around an example DNase-I Hypersensitive Site 86
4.8 Schematic of peak calling and scoring 87
4.9 Distances between adjacent peaks, by sample 88
4.10 Comparison of peak calls between samples 89
4.11 Distances between adjacent peaks, samples CH01 90
4.12 Nucleosome positioning and spacing correlates with genomic features 91
4.13 Nucleosome spacing in A/B compartments 92
4.14 Short cfDNA fragments footprint CTCF and other TF binding sites 93
4.15 Nucleosome spacing around DHSes in 116 callsets 94
4.16 WPS profiles stratified by gene expression 95
4.17 Correlation of gene expression and inferred nucleosome spacing 96
4.18 Inference of mixtures of cell types contributing to cell-free DNA in healthy states and cancer 97
4.19 Stability of inference of cfDNA tissues-of-origin to downsampling 99
4.20 Stability of inference of cfDNA tissues-of-origin to dilution 101
4.21 Comparison of uniformity and completeness of nucleosome callsets 102
4.22 Comparison of peak locations between samples and call sets 103

5.1 Capture of cfDNA targets with smMIPs, by input 138

LIST OF TABLES

Table Number Page

2.1 Summary of sequencing 43
2.2 Accuracy of fetal genome inference 44
2.3 *De novo* point mutations identified by whole-genome shotgun sequencing . 45

3.1 Calculation of standard deviations in number of reads derived from chromosomes 13, 18, and 21 66
3.2 Control CNV cohorts and call sources 67
3.3 Demographic and pregnancy characteristics of study subjects 71
3.4 Breakpoints and PCR primers for detection of maternal CNVs 73
3.5 The role of CNV size and fetal fraction in false-positive NIPT results 74

4.1 Sequencing statistics for samples BH01, IH01, IH02, and CH01 113
4.2 Synthetic oligos used in preparation of single stranded sequencing libraries . 113
4.3 Correlation of WPS FFT intensities with gene expression datasets 114
4.4 Clinical diagnoses and cfDNA yield for cancer panel samples 120
4.5 Sequencing statistics for additional samples included in CA01 122

ACKNOWLEDGMENTS

I am deeply indebted to the many individuals who have offered academic, financial, and emotional support during the preparation of this dissertation.

I first thank my Ph.D. supervisor, Jay Shendure, for the freedom to develop this research project, the funding to execute it, the guidance to improve it, the motivation to see it through, and the excitement to publicize it widely.

The remaining members of my committee each played different but important roles in my training. I am grateful to Bob Waterston for warning me early on that I was taking a risk with my dissertation project, inspiring me to work harder to prove him wrong. Maynard Olson steered me away from a questionable decision and indulged me in my deep dives into the history of our young science. Phil Green's course proved immensely valuable for the work described in Chapter 2. Lee Nelson's genuine, honest-to-goodness interest in my work motivated me and suggested several new avenues for future exploration. Brian Browning bailed me out of a tight spot, and generously shared his expertise regarding the work described in Chapter 2.

While I have benefited enormously from interactions and collaborations with every member of the Shendure lab, I would like to single out several individuals for contributions that merit particular thanks. Jacob Kitzman took me on as an inexperienced rotation student and nearly single-handedly turned me into a competent scientist. His attention to detail, his focus on experimental design, and his willingness to share his exceptional skills in the lab are inspirational. Voni Simmons taught me the best collaborators are the ones who ask the hard questions, who challenge your assumptions, and who can get you the really choice samples. Martin Kircher told me very early on that he had nothing he could teach me, and

then proceeded to spend the next four years teaching me things. Andrew Adey, Akash Kumar, and Greg Findlay both generously made the lab a more creative, collaborative space and asked little in return.

I have been very lucky to complete my doctoral research in the Department of Genome Sciences at this moment in time, surrounded by an amazing group of faculty, administrators, and trainees – the very reason I was attracted to the department in the first place. Thanks to David Young, Jeff Vierstra, Ben Vernot, Matt Rich, Max Press, Alan Rubin, Nancy Cameron, Dawn Counts. Thanks also to my frequent collaborator Hilary Gammill, whose efforts were critical to the success of much of the work presented here.

I thank my family for never questioning – at least not openly – the circuitous route I took to get here, and for showing even more excitement than I did about my progress.

Finally, I thank my wife, Cecilia Roussel, for her willingness to follow me to Seattle, her encouragement during my many setbacks, her cheerleading during my fewer successes, and her patience and love through many late nights.

DEDICATION

This work is dedicated to my mother, perhaps the only person not surprised that I am actually completing it.

Chapter 1

INTRODUCTION

A sea change is underway in noninvasive clinical screening and testing. Tests based on quantifying and characterizing a blood analyte, cell-free DNA, have quickly emerged in prenatal medicine and oncology as promising, low-risk alternatives to "gold standard" invasive tests such as amniocentesis or tumor biopsy. In recent years, the resolution, accuracy, and breadth of this testing methodology have increased, expanding the range of conditions that can be monitored noninvasively and improving the precision with which these conditions can be tracked. However, the vast quantities of genetic data these tests provide may sometimes complicate, rather than clarify, clinical decision-making.

In this chapter, I briefly review the history of cell-free DNA investigation, describe major applications of this biomarker to clinical questions, and provide background on the technologies and analytical methodologies used in these applications. I further detail the types of information that this class of noninvasive testing can provide to clinicians and patients, and the potential directions for and implications of its continued growth.

1.1 Detritus of cell death: the origins of cell-free DNA

Cell-free DNA (cfDNA) is the detritus of cell death, and consists of the many short and short-lived fragments that circulate in the blood and other bodily fluids of all individuals, regardless of disease state. Despite its discovery in blood in 1948 (Mandel and Métais, 1948), circulating DNA remained essentially a curiosity until the detection, nearly 40 years later, of tumor-shed fragments of cfDNA in the plasma of cancer patients (Stroun et al., 1987). Another decade later came the first report of fetal-derived cell-free DNA in the cir-

culation of pregnant women, along with the suggestion that the "use of maternal plasma or serum for the detection of fetal DNA for non-invasive *(sic)* prenatal diagnosis may therefore be possible" (Lo et al., 1997).

The twenty-year period since that prescient remark has witnessed explosive growth in the use of cfDNA for noninvasive testing – not just in pregnancy, but also for the detection and monitoring of a growing number of cancers, early warning of allograft rejection, and tracking of viral infection over time. Indeed, to date, cfDNA's "greatest value has been to measure the proportion of foreign genomes within an individual" (Quake, 2012), particularly when these measurements are used to guide clinical decision-making. While noninvasive, indirect cfDNA-based assays have generally not yet reached diagnostic status, their clinical uptake has grown rapidly both in the US and abroad.

The rapid development of cfDNA-based methods with direct application to clinical questions, discussed below, has in some ways meant putting the cart before the horse. Several fundamental questions about the origins and properties of cfDNA remain unanswered, some of which are directly addressed in Chapter 4 of this dissertation. Nevertheless, many of the basic parameters have been sketched out. The major biological source of plasma-borne cfDNA in healthy individuals is apoptotic nucleated blood cells (Koh et al., 2014; Lui et al., 2002; Sun et al., 2015), with contributions from other tissues in the context of altered physiology. Typically, cfDNA is present at approximately 1-10 nanograms (ng) per milliliter (mL) of plasma (Fleischhacker and Schmidt, 2007), so that a 10 mL blood draw will contain roughly 1500-15,000 genomic equivalents in the plasma fraction. In individuals with physiological conditions such as obesity, pregnancy, or cancer, the concentration of cfDNA can increase markedly. Most fragments of cfDNA are double-stranded and short, overwhelmingly less than 200 base-pairs (bp). Although some debate remains about the length distribution of cfDNA fragments and the extent to which it may vary based on disease processes (Thierry et al., 2010), typical fragments observed in high-throughput sequencing libraries tend to be 147-167 bp (Fan et al., 2008; Lo et al., 2010), leading to the hypothesis that cfDNA fragments are predominantly nucleosomal in origin. The half-life of these

fragments is quite short, on the order of 15 minutes (Lo et al., 1999), suggesting a model of ongoing degradation and filtration of these fragments from the circulation, although the exact mechanisms of this clearance are not yet clear. At least some cfDNA is filtered from the circulation by the kidneys, leading to detectable cfDNA in urine (Botezatu et al., 2000). Fragments of cfDNA purified from urine or from cerebrospinal fluid may also have limited clinical utility in specific contexts (Angert et al., 2004; Rhodes et al., 1994), although to date, most studies and applications have focused on plasma-borne cfDNA.

Below, I discuss some of the success stories of these applications of cfDNA analysis to clinical questions. I also highlight the limitations of existing analytical methodologies, and discuss ways in which these challenges might be overcome.

1.2 Prenatal testing with cell-free DNA

The impact of cfDNA-based testing on clinical practice has nowhere been stronger than in prenatal and reproductive medicine. The application of high-throughput, massively parallel sequencing to cfDNA, fittingly termed "next-generation sequencing and the next generation" (Chitty and Bianchi, 2015), has dramatically increased the number of potential diseases and disorders that can be profiled *in utero*. While noninvasive screening for fetal abnormalities is not new – well-established techniques including maternal serum screening, triple and quad screening, and fetal ultrasounds present virtually zero risk to the mother or the fetus – the accuracy and resolution of the new class of cfDNA-based genetic tests for aneuploidy, collectively termed "noninvasive prenatal testing" or "NIPT," has led to an unparalleled rate of adoption in clinical practice. Even where high resolution is less important – for example, in screening for fetal trisomies – cfDNA-based testing now rivals diagnostic "gold standard" assays for accuracy and eliminates risks involved with invasive tests (Bianchi et al., 2014).

Despite the lack of FDA approval for NIPT, recent decisions by major health insurance payors Anthem and Blue Cross Blue Shield to reimburse NIPT costs for average-risk pregnancies now open the door to increasing uptake in the United States. Indeed, a recent

study of maternal-fetal medicine specialists in the US and internationally found that 94.3% of respondents were offering NIPT to at least some patients (Haymon et al., 2014). Clinical uptake of this new testing paradigm has been so strong that noninvasive prenatal testing for aneuploidy using cfDNA may be the fastest growing molecular test in the history of medicine (Chitty and Bianchi, 2015), and represents the biggest success story for genomic medicine to date.

How can profiling of maternal cfDNA yield insight into fetal health? During gestation, a portion of cfDNA in the maternal circulation is derived from apoptotic trophoblast cells in the placenta (Alberry et al., 2007; Masuzaki et al., 2004). These molecules constitute the so-called "fetal fraction," which varies from approximately 5% to 25% of total cfDNA and generally increases with gestational age (Nygren et al., 2010). Proceeding on the assumption that the placental genome mirrors the genome of the developing fetus, cfDNA-based analysis can make inference about the fetus while only "indirectly" accessing its genetic material. The rapid clearance of this placenta-derived cfDNA after delivery suggests that the half-life of this placental material is roughly 15 minutes (Lo et al., 1999) and argues that a biological sample from a pregnant mother contains, in essence, a snapshot of fetal health development over a narrow time window, and importantly a snapshot that contains the entirety of the fetal genome (Lo et al., 2010).

NIPT methodologies to identify common fetal aneuploidies typically involve one of three approaches. In the first method, cell-free DNA is sequenced, and fragments unambiguously derived from each chromosome are tallied. In the rare event of a fetal trisomy or monosomy, the number of fragments from a given chromosome will deviate significantly from the expected result – in other words, there will be a detectable over- or underrepresentation of the relevant chromosome compared to a reference cohort of euploid pregnancies (Chiu et al., 2008; Fan et al., 2008). A second method modifies this approach to sequence only targeted regions of selected chromosomes, potentially reducing the likelihood of incidental findings (Sparks et al., 2012). In the third method, a genome-wide panel of common SNPs is investigated by multiplex PCR of cfDNA followed by sequencing of the resulting amplicons.

Jointly considered, the ratios of the alleles observed at each site should be consistent across chromosomes, and deviations from this expectation signal possible aneuploidy (Zimmermann et al., 2012).

Test performance from each of these methodologies has been consistently high in both high- and average-risk cohorts (Bianchi et al., 2014; Dar et al., 2014; Norton et al., 2015), with most test metrics comparing favorably with standard screening. However, as with any rare disease, even a small number of false-positive tests can severely limit the positive predictive value (PPV) of screening. This number – the probability that a positive test result truly indicates fetal aneuploidy – is particularly relevant in average-risk cohorts, where the prior probability of disease is by definition lower than in cohorts defined by advanced maternal age, medical history, or other risk factor. Concerns about the PPVs of NIPT, among other factors, have led to the continued and joint recommendation by two professional organizations, the American College of Obstetricians and Gynecologists (ACOG) and the Society for Maternal-Fetal Medicine (SMFM), that NIPT be considered non-diagnostic, with invasive followup procedures suggested for all positive tests (Committee on Genetics, Society for Maternal-Fetal Medicine, 2015).

As the number of administered tests in average-risk patient groups increases, the number false-positive results can be expected to grow concomitantly. If ACOG and SMFM advice is heeded, the uptake of this noninvasive testing framework may paradoxically lead to growth in the number of invasive, diagnostic procedures – specifically, procedures that likely would not have happened had noninvasive alternatives *not* been available. In order to avoid this unfortunate eventuality, attention has turned to the enumeration of the various causes of false-positive tests and the ways in which such results might be minimized or avoided. Plausible causes of spurious findings include inevitable statistical errors, poor sample handling, and other technical issues with the tests themselves; or true biological phenomena including maternal copy-number variation or confined placental mosaicism, the latter of which in effect violates the assumption that the placental genome provides a mirror into the genetic makeup of the fetus. A more complete treatment of this topic is

provided in Chapter 3.

The resolution of NIPT continues to improve, enabling clinicians to evaluate the risk of additional classes of genetic insults beyond whole-chromosome aneuploidy. Recently, cfDNA-based tests able to detect sub-chromosomal abnormalities, such as microdeletions implicated in Prader-Willi or Angelman syndromes, have been described (Srinivasan et al., 2013; Wapner et al., 2015) and added to some existing NIPT offerings, although it should be noted that ACOG and SMFM guidelines currently discourage such screening (Committee on Genetics, Society for Maternal-Fetal Medicine, 2015). Assays for risk assessment in the context of monogenic disorders, in which detection of specific risk alleles is required[1], are either in development or already in limited use for a variety of conditions, including Huntington's disease (van den Oever et al., 2015) and cystic fibrosis (Hill et al., 2015).

Rather than designing hundreds or thousands of cfDNA-based tests targeting single-gene disorders, microdeletion syndromes, and other diseases, each of which must be individually validated, it may be preferable to develop a single, comprehensive test that can simultaneously interrogate a pregnancy for risk of aneuploidy, pathogenic structural variation, and the more than 3,000 Mendelian disorders with known molecular bases (Amberger et al., 2009): in other words, to sequence the fetal genome noninvasively. The comprehensive determination of the inherited and *de novo* variation present in a given fetus, described in Chapter 2, is now technically within reach (Fan et al., 2012; Kitzman et al., 2012). Despite the promise of such a test for impacting reproductive outcomes, multiple challenges to widespread adoption remain. First and foremost is the challenge of interpretation of exhaustive test results: how much information is too much for a clinician, a genetic counselor, or a prospective parent? Second, substantial technical and logistical obstacles – including experimental complexity, scalability, necessary expertise, and cost – remain significant impediments to clinical adoption. Early reports of "first draft" versions of such tests col-

[1]Although outside the scope of this dissertation, carrier screening of parents for risk alleles – for example, on the basis of genetic ancestry or family history of genetic disease – followed by preimplantation genetic diagnosis of fertilized embryos represents an alternative course of action for prospective parents (Kumar et al., 2015).

lectively have examined only a handful of pregnancies, with no larger studies planned or underway. Finally, the ethical considerations surrounding prenatal testing, which are not unique to cfDNA-based NIPT, are magnified in light of increasing resolution, with potential for eventual prenatal prediction of traits such as color blindness or male pattern baldness.

1.3 The "liquid biopsy" for cancer

Fragments of cfDNA can be shed not only by myeloid and lymphoid cells and placental trophoblasts, but by other tissues in the body, including neoplastic cells during oncogenesis (Stroun et al., 1987). These tumor-derived fragments are often termed "[c]irculating [t]umor DNA" or "ctDNA," and represent a fraction of total cfDNA that correlates imperfectly with tumor stage, treatment status, and cancer type. The analysis of ctDNA to provide insights into disease progression, or to guide or measure response to treatment, is now called the "liquid biopsy" for its potential to supplement or eventually replace direct biopsies of cancerous tissue.

How do liquid biopsies provide insight into disease status? Most such tests rely on the fact that tumor genotypes will differ from the germline genetic makeup of an individual patient. The number and classes of mutations in a given individual's tumor may vary, but all tumors should be marked by at least a small handful of somatic mutations, the collection of which constitutes a genetic fingerprint of a tumor. This list of variant loci can be obtained directly by conventional tumor biopsy and sequencing of the purified tumor DNA, or less accurately by assuming the presence of one or more stereotyped, typically activating mutations in a given tumor. By scanning the plasma-borne cfDNA for this panel of mutations, the burden of mutant fragments can be assessed. This burden correlates with tumor stage and volume (Newman et al., 2014), and the whorls and loops of this genetic fingerprint may in some cases suggest therapeutic strategies (discussed below).

The identification of mutant fragments to quantify disease burden is challenged on several fronts. First, the human genome is large, requiring substantial sequencing costs to cover each base repeatedly. Second, the number of mutant fragments in a sample at any

given locus may be quite small, often less than a single copy per mL of plasma (Diehl et al., 2007), suggesting that highly sensitive methods may be needed to cover the full spectrum of ctDNA load. Third, while some cancer-associated mutations, such as the *BRAF* V600E mutation in melanoma or the *IDH1* R132H mutation in glioma, are repeatedly observed across a large number of patients, many mutations in a given patient may be rare and specific to a single tumor, while simultaneously being relevant for directing the selection of a therapeutic intervention. Thus, while technologies that target specific, stereotyped mutations in the cfDNA using quantitative or digital PCR (Li et al., 2006; Thierry et al., 2014) may be useful in limited contexts, more promising are frameworks that capture and interrogate panels of relevant genes with high sensitivity (Forshew et al., 2012; Newman et al., 2014). Whole-genome strategies, even those using molecular barcoding to enable precise counting of unique mutant molecules (Kinde et al., 2011), suffer from greatly expanded search space and inefficiencies introduced during library preparation, which typically yields conversion rates under 5% when using cfDNA as input material.

The extent to which tumors contribute ctDNA to the plasma compartment varies both within and between cancer subtypes. While noninvasive detection of many brain tumors is complicated by the blood-brain barrier's role in restricting ctDNA from entering the circulation, the picture is somewhat less grim for a wide variety of other subtypes. In a recent study (Bettegowda et al., 2014), ctDNA was detected in 82% of individuals with solid tumors outside the brain, with particularly high rates of detection in bladder, colorectal, gastroesophageal, and ovarian cancers, albeit with a limited number of cases examined. Within a cancer subtype, the number of mutant fragments detected can vary widely, with one study of colorectal cancer finding anywhere from just over one to several thousand copies of a mutant allele at a single gene within a sample (Diehl et al., 2007), in broad agreement with another study using a different methodology (Thierry et al., 2014). Although one hypothesis for this high variance between samples is the copy-number amplification in some tumors, the fact that consistent estimates obtained across genes within an individual sample (Diehl et al., 2007) argues that other factors may play a larger role in this variability. Finally, it is

important to note that these studies identify mutant fragments by profiling a small number of recurrently mutated genes, and thus may underestimate the true number of cancers with detectable ctDNA.

At least two other uses for the liquid biopsy have been proposed. First is the assessment of tumor heterogeneity both within and between tumors in a single individual, a goal that otherwise would require multiple biopsies within and across primary and metastatic tumors in various parts of the body along with concomitant risks (Diaz and Bardelli, 2014). Second, ctDNA-based monitoring allows detection of markers of resistance to treatment, either prior to or in response to treatment (Diaz et al., 2013; Murtaza et al., 2013). Interestingly, a single test may provide insight into both of these clinical considerations: by applying a target enrichment strategy to a ctDNA sample from an individual with non-small cell lung cancer, two tumor subclones were identified, one with an activating *EGFR* mutation, suggesting treatment by erlotinib; and another with a different *EGFR* mutation conferring resistance to erlotinib (Newman et al., 2014).

Despite continued improvements to both data acquisition and analysis, ctDNA-based screening is still in its infancy. Uptake of this class of noninvasive monitoring has yet to rival the growth in the use of prenatal screening, but the promise of tailoring cancer therapeutics to the specific genetic makeup of the constellation of tumors within an individual with late-stage metastatic disease suggests continued refinement to and availability of these methods in clinical settings. Whether or not patient outcomes are directly impacted, these methods may open the door to a greater understanding of the heterogeneity and evolution of cancers, particularly those typically grouped on the basis of histological or anatomical similarity.

1.4 Noninvasive monitoring of allograft health

A third success story for noninvasive monitoring with cfDNA is in the surveillance of the health of transplanted organs. Anecdotal correlations between the burden of cfDNA in a transplant recipient and the probability of acute rejection of the allograft have led to more complete investigations into the potential link between these two phenomena. Donor geno-

types – present in the transplanted organ – typically differ sufficiently from recipient genotypes to allow assignment of the origin of specific cfDNA fragments to one or the other genomes, and thus to the transplanted organ or to uninvolved tissue. By monitoring the levels of donor-derived cfDNA over time, increasing evidence now suggests the health of the allograft can be noninvasively followed, and in some cases, acute rejection events may even be predictable.

The presence of donor-derived cfDNA fragments was described as early as 1998, when fragments of the Y chromosome were detected by PCR in the plasma of female liver and kidney transplant recipients receiving organs from male donors (Lo et al., 1998b). Despite similar findings in a cohort of sex-mismatched bone marrow transplant recipients (Lui et al., 2002), contradictory reports from modest numbers of heart, liver, and kidney transplant recipients argued that such fragments might make up a small fraction of total cfDNA (Lui et al., 2003). The advent of high-throughput sequencing in the ensuing years allowed more precise quantification of rare fragments, and confirmed contributions from donated heart (De Vlaminck et al., 2014; Snyder et al., 2011) and lung (De Vlaminck et al., 2015), often detectable within a single postoperative day.

Although a full treatment of this topic is outside the scope of this dissertation, the increasing numbers of transplant recipients surveilled in this way may lead to more accurate modeling for prediction of acute rejection. Already, classification models based on quantifying the burden of donor DNA can group patients into those likely to have the most severe rejection events and those likely to have no rejection with areas under the curve (AUC) of 0.9. In particular, by combining this testing framework with additional evidence from standard clinical tests, classification performance rivals “gold standard” assays in at least some clinical contexts (De Vlaminck et al., 2015).

1.5 Tissues of origin: expanding the breadth of cfDNA-based testing

In principle, since cfDNA represents a snapshot of cell death in many cell types and tissues within a narrow time window, the application of high-throughput genomic technologies

to cfDNA investigation should allow the noninvasive and temporally relevant monitoring of the health of a wide variety of tissues with a routine blood draw. This monitoring could include not only the detection and classification of disease processes, but also the evaluation of the response to treatment and potential disease recurrence. Given this promise, why have such screens lagged?

In practice, the deconvolution of this mixture of cfDNA – the assignment of fragments to the tissues or cell types from which they originated – has proved a limiting factor in the uptake of this approach in the clinic. This problem is exacerbated by the possibility that many of the contributions may be both modest in magnitude and clinically relevant, such that this deconvolution process should ideally identify not just the dominant contributors, but minority ones as well. Circulating cfDNA typically consists of a small mass of short, degraded fragments with brief half-lives. What features of these molecules might be interrogated to learn about their origins?

Existing methods for disentangling the complex mixture of cell types and tissues whose apoptosis and necrosis give rise to cfDNA most commonly involve identification and quantification of genotypic differences between tissues in an individual. These genotypic differences can take several forms. Perhaps most intuitive is the application of cfDNA to the detection and monitoring of cancer: the so-called “liquid biopsy” currently in limited clinical use and discussed briefly above. Here, tumor-specific somatic variation, which may include stereotyped driver mutations, copy-number changes, or aneuploidy, can be discovered through conventional tumor biopsy followed by sequencing or microarray analysis. Once identified, these variants can be searched for in the circulating cfDNA, a portion of which will be derived from the tumor, and used to build a model of disease burden or tumor progression. A similar approach can be applied to questions in pregnancy, where fetal or placental cfDNA fragments should evidence paternally inherited genotypes distinguishable from the maternal background, and in organ transplantation, where the donated organ sheds fragments of the donor’s genome, again separable on the basis of DNA sequence from the patient’s own genetic material. Downstream analysis proceeds on the basis of observ-

ing the concentration or kinetics of such foreign fragments and comparing these findings to expectations derived from reference cohorts.

This straightforward and attractive approach is hamstrung, though, by its reliance on the presence of such genotypic differences between contributing cells. In most cases, such differences may not exist, or if they do exist, the location and composition of the differences may not be known *a priori*. As a concrete example, severe inflammation of the kidneys may result in the shedding of renal cfDNA fragments into the circulation, but the genetic makeup of those dead cells is not likely to differ from that of the white blood cells that contribute the majority of cfDNA fragments. How would a clinical assay for nephritis try to identify kidney-derived cfDNA fragments?

Two strategies have been proposed to address this question. The first strategy relies on epigenetic differences between tissues in the body: specifically, the DNA methylation profiles of these tissues. The collected cfDNA is subjected to bisulfite conversion and sequencing, and the methylation status of the observed sample is compared to previously ascertained methylation maps of various tissues using a quadratic programming approach to model the number and proportions of contributing tissues (Sun et al., 2015). This strategy is challenged by the difficulty of constructing complex cfDNA sequencing libraries following bisulfite conversion, which has the effect of limiting the number of informative molecules in a typical experiment, and by the modest number of tissue-specific DNA methylation loci – no more than a few dozen sites for a given tissues (Sun et al., 2015). In addition, the DNA methylation pattern within a tissue can vary with age or environmental exposures, potentially confounding inference. A second strategy relies on the detection and quantification of a different class of circulating nucleic acid, in this case cell-free RNA (cfRNA). As cell types differ in their gene expression profiles, identifying tissue-specific transcripts or transcriptional programs and modeling cfRNA contributions on the basis of transcript counting can provide insight into the composition of the cfDNA (Koh et al., 2014). As of this writing, the interrogation of cfRNA has been applied in very limited contexts, and performance characteristics of the methodology are as yet unknown. Typical cfRNA yield is

low in biological samples, and because of the risk of co-purification of cfDNA and cfRNA, quantitation of cell type-specific fragments may be confounded by the presence of cfDNA in a sample. Additionally, because cfRNA is highly degraded, detection of tissue-specific isoforms is challenging. Nevertheless, this strategy may prove particularly useful for specific applications, such as the identification of RNA virus infection.

In Chapter 4, I revisit the question of deconvolution of cfDNA fragments into their tissues of origin. I investigate whether the very enzymatic processes that give rise to cfDNA during cell death might inform the tissues-of-origin of this mixed population. In particular, I ask whether plasma-borne cfDNA fragments are protected from nuclease degradation by association with nucleosomes, transcription factors (TFs), and other DNA-binding proteins, and whether the footprints of such factors might be used to model the various contributors.

1.6 Organization of this dissertation

In this dissertation, I aim to improve the resolution, accuracy, and breadth of cfDNA-based testing using high-throughput sequencing. I explain how a more complete understanding of cfDNA ontology suggests new strategies for cfDNA-based screening and demonstrate proof-of-concept embodiment of one such strategy.

In the second chapter of this dissertation, I describe a strategy for increasing the *resolution* of cfDNA-based testing in prenatal medicine by determining the whole genome sequence, including both inherited and *de novo* variation at the single nucleotide level, of a human fetus. I explain how this increased resolution in principle represents a step towards the prenatal diagnosis of the more than 3,000 genetic diseases with known causes.

In the third chapter, I explain how the *accuracy* of prenatal, cfDNA-based screening for aneuploidy is negatively impacted by maternal copy-number variation, and develop a statistical framework for estimating the likely impact of this finding based on specific testing frameworks. I further describe straightforward modifications to current analytical approaches that, if applied, mitigate this potential confounder.

In the fourth chapter, I first describe key biological features of cfDNA, shedding light on the mechanisms through which it is created in healthy and diseased individuals. I show that contributions from cancerous tissues can be identified in cfDNA solely on the basis of the locations of the fragment endpoints and comparison to reference gene expression datasets. I use these findings to explain how the *breadth* of cfDNA-based screening, currently limited to specific cancers, pregnancy, and organ transplantation, might be expanded to additional physiological conditions not currently amenable to this type of investigation.

In the final chapter, I discuss opportunities and challenges of these distinct areas of investigation moving forward, particularly with regard to clinical adoption. I close with thoughts about the promise and pitfalls of precision medicine informed by genomic investigation.

Chapter 2

NONINVASIVE WHOLE GENOME SEQUENCING OF A HUMAN FETUS

This chapter has been adapted with changes from: Kitzman, JO; Snyder, MW; Ventura, M; Lewis, AP; Qiu, R; Simmons, LE; Gammill, HS; Rubens, CE; Santillan, DA; Murray, JC; Tabor, HK; Bamshad, MJ; Eichler, EE; and Shendure, J. Non-invasive whole genome sequencing of a human fetus. *Science Translational Medicine* 4(137):137ra76 (2012)

and from: Snyder, MW; Simmons, LE; Kitzman, JO; Santillan, DA; Santillan, MK; Gammill, HS; and Shendure, J. Noninvasive fetal genome sequencing: a primer. *Prenatal Diagnosis* 33(6):547-554 (2013)

and from: Snyder, MW; Adey, A; Kitzman, JO; and Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics* 16(6):344-358 (2015).

My specific contributions to this project were the creation of the computational and statistical methods used for fetal genome and *de novo* mutation inference, preparation of all figures excluding 2.2 and 2.4, performing the analysis of downsampling, and drafting portions of the manuscript.

Abstract

Analysis of cell-free fetal DNA in maternal plasma holds promise for the development of noninvasive prenatal genetic diagnostics. Previous studies have been restricted to detection of fetal trisomies (Chiu et al., 2008; Fan et al., 2008), specific paternally inherited mutations (Lo and Chiu, 2007), or to genotyping common polymorphisms using material obtained invasively e.g. through chorionic villus sampling (Lo et al., 2010). Here, we combine genome sequencing of two parents, genome-wide haplotyping (Kitzman et al., 2011) of a mother and deep sequencing of maternal plasma DNA to noninvasively determine the genome sequence of a human fetus at 18.5 weeks gestation. Inheritance was predicted at $2.8x10^6$ parental heterozygous sites with 98.1% accuracy. Furthermore, 39 of 44 *de novo* point mutations in the fetal genome were detected, albeit with limited specificity. Subsampling these data and analyzing a second family trio by the same approach indicate that ~300

kilobase parental haplotype blocks combined with shallow sequencing of maternal plasma DNA is sufficient to substantially determine the inherited complement of a fetal genome. However, ultra-deep sequencing of maternal plasma DNA is necessary for the practical detection of fetal *de novo* mutations genome-wide. Although technical and analytical challenges remain, we anticipate that noninvasive analysis of inherited variation and *de novo* mutations in fetal genomes will facilitate prenatal diagnosis of both recessive and dominant Mendelian disorders.

2.1 Introduction

On average, ~13% of cell-free DNA isolated from maternal plasma during pregnancy is fetal in origin (Nygren et al., 2010). The concentration of cell-free fetal DNA in the maternal circulation varies between individuals, increases during gestation, and is rapidly cleared postpartum (Lo et al., 1998a, 1999). Despite this variability, cell-free fetal DNA has been successfully targeted for noninvasive prenatal diagnosis including for development of targeted assays for single gene disorders (Lo and Chiu, 2007). More recently, several groups have demonstrated that shotgun, massively parallel sequencing of cell-free DNA from maternal plasma is a robust approach for noninvasively diagnosing fetal aneuploidies such as trisomy 21 (Chiu et al., 2008; Fan et al., 2008).

Ideally, it should be possible to noninvasively predict the whole genome sequence of a fetus to high accuracy and completeness, potentially enabling the comprehensive prenatal diagnosis of Mendelian disorders and obviating the need for invasive prenatal diagnostic procedures such as chorionic villus sampling with their attendant risks. However, several key technical obstacles must be overcome for this goal to be achieved using cell-free DNA from maternal plasma. First, the sparse representation of fetal-derived sequences poses the challenge of detecting low frequency alleles inherited from the paternal genome as well as those arising from *de novo* mutations in the fetal genome. Second, maternal DNA predominates in the mother's plasma making it difficult to assess maternally inherited variation at individual sites in the fetal genome.

Recently, Lo et al. showed that fetal-derived DNA is distributed sufficiently evenly in maternal plasma to support the inference of fetal genotypes, and furthermore they demonstrated how knowledge of parental haplotypes could be leveraged to this end (Lo et al., 2010). However, their study was limited in several ways. First, the proposed method depended on the availability of parental haplotypes, but at the time of their work, no technologies existed to measure these experimentally on a genome-wide scale. Therefore, an invasive procedure, chorionic villus sampling, was used to obtain placental material for fe-

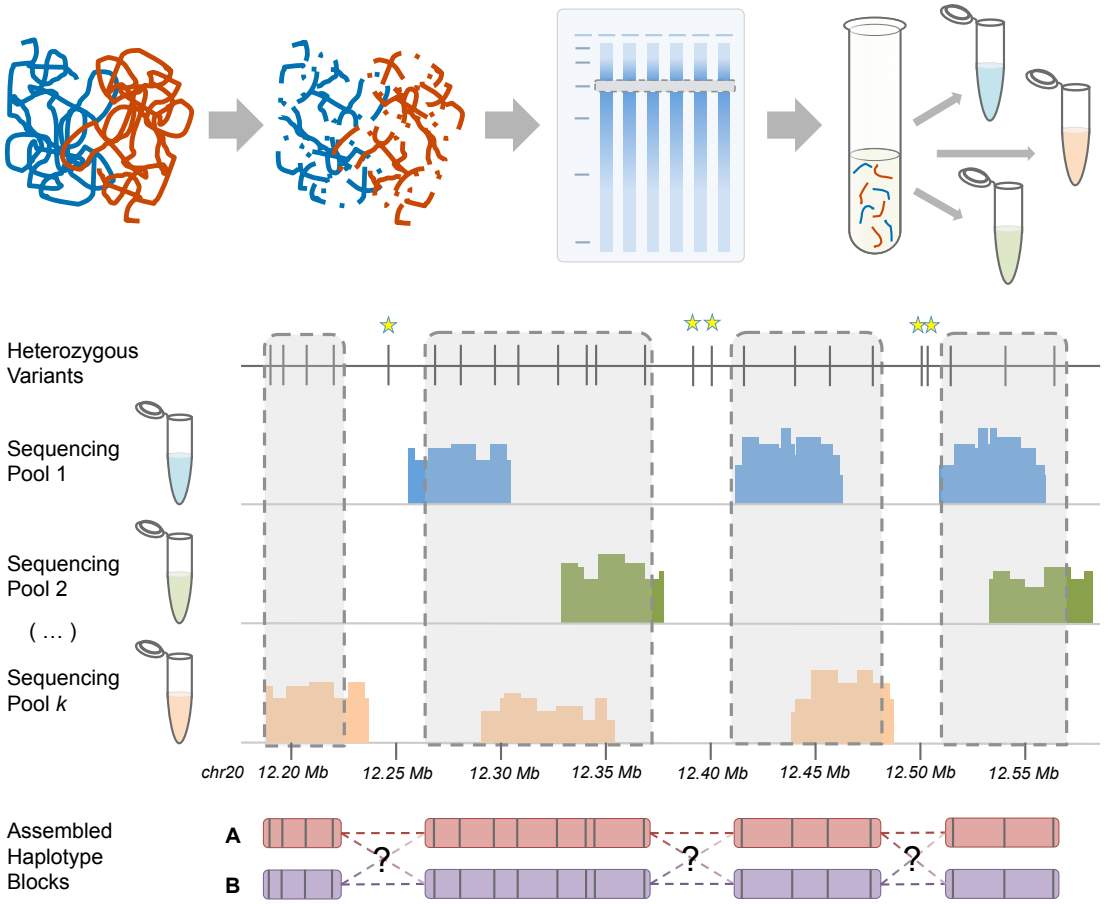


Figure 2.1: **Schematic of dense haplotyping**. Genomic DNA is extracted from a large number of cells. The resulting sample contains a mixture of both haplotypes *(blue and red fragments)* from every genomic region. After gentle fragmentation, high molecular weight DNA is size-selected, in this example by gel electrophoresis, to enrich for fragments 35-40 kilobases in length. The resulting pool of fragments is used to produce a fosmid library *(not shown)*, which is then diluted to a large number of reaction chambers, such that each chamber has zero or one copies of any given genomic region – and thus, no more than a single haplotype at any locus. An indexed sequencing library is prepared separately from each diluted pool. After sequencing, reads are aligned to the genome, and compared to a list of heterozygous sites produced by conventional shotgun sequencing followed by variant calling. Islands of coverage overlapping between two or more pools *(grey boxes)* are stitched together on the basis of allele sharing at heterozygous sites. The resulting haplotype blocks, arbitrarily labelled A and B, densely cover the variation within the windows defined by overlapping islands of coverage, with few variants left unphased *(yellow stars)*. However, contiguity between adjacent blocks is unknown, and haplotypes longer than one to two megabases are typically not obtainable.

tal genotyping. Second, parental genotypes and fetal genotypes obtained invasively were used to infer parental haplotypes. These haplotypes were then used in combination with the sequencing of DNA from maternal plasma to predict the fetal genotypes. Although necessitated by the lack of genome-wide haplotyping methods, the circularity of these inferences makes it difficult to assess how well the method would perform in practice. Third, their analysis was restricted to several hundred thousand parentally heterozygous sites of common single nucleotide polymorphisms (SNPs) represented on a commercial genotyping array. These common SNPs are only a small fraction of the several million heterozygous sites present in each parental genome, and include few of the rare variants that predominantly underlie Mendelian disorders (MacArthur et al., 2012). Fourth, Lo et al. did not ascertain *de novo* mutations in the fetal genome. As *de novo* mutations underlie a substantial fraction of dominant genetic disorders, their detection is critical for comprehensive prenatal genetic diagnostics. Therefore, although the Lo et al. study demonstrated the first successful construction of a genetic map of a fetus, it required an invasive procedure and did not attempt to determine the whole genome sequence of the fetus. We and others recently demonstrated methods for experimentally determining haplotypes for both rare and common variation on a genome-wide scale (Fan et al., 2010; Kitzman et al., 2011; Ma et al., 2010; Yang et al., 2011). In the current study, we set out to integrate the haplotype-resolved genome sequence of a mother, the shotgun genome sequence of a father, and the deep sequencing of cell-free DNA in maternal plasma to noninvasively predict the whole genome sequence of a fetus.

2.2 Results

We set out to predict the whole genome sequence of a fetus in each of two mother-father-child trios (I1, a first trio at 18.5 weeks gestation; G1, a second trio at 8.2 weeks gestation). We focus here primarily on the trio for which considerably more sequence data was generated (I1) (Table 2.1).

In brief, the haplotype-resolved genome sequence of the mother (I1-M) was determined

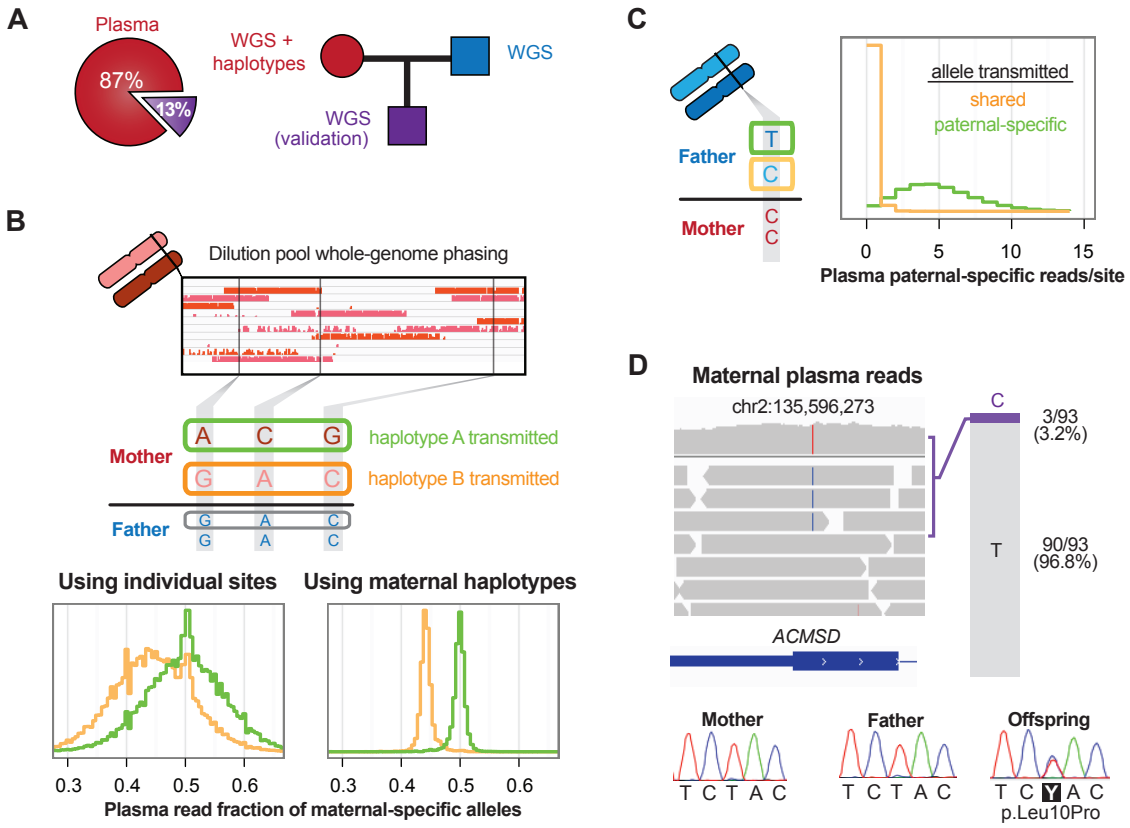


Figure 2.2: **Schematic of experimental approach**. *(A)* Sequenced individuals in a family trio. Maternal plasma DNA sequences were ~13% fetal-derived based on read depth at chrY and alleles specific to each parent. *(B)* Inheritance of maternally heterozygous alleles inferred using long haplotype blocks. Among plasma DNA sequences, maternal-specific alleles are more abundant when transmitted (expected 50% versus 43.5%), but there is substantial overlap between the distributions of allele frequencies when considering sites in isolation (left histogram, yellow=shared allele transmitted, green=maternal-specific allele transmitted). Taking average allele balances across haplotype blocks (right histogram) provides much greater separation. *(C)* Histogram of fractional read depth among plasma data at paternal-specific heterozygous sites. In the overwhelming majority of cases when the allele specific to the father was not detected, the opposite allele had been transmitted (96.8%, n=561,552). *(D)* Validated *de novo* missense mutation in the gene ACMSD detected in 3 of 93 plasma reads, creating a leucine-to-proline substitution at a conserved site in a gene implicated in Parkinson's disease by GWAS (International Parkinson Disease Genomics Consortium et al., 2011).

by first performing shotgun sequencing of maternal genomic DNA from blood to 32-fold coverage (coverage = median-fold coverage of mapping reads to the reference genome after discarding duplicates). Next, by sequencing complex haploid subsets of maternal genomic DNA while preserving long-range contiguity (Kitzman et al., 2011), we directly phased 91.4% of 1.9 x 10^6 heterozygous SNPs into long haplotype blocks (N50 of 326 kilobases (kbp) (Figure 2.1). The shotgun genome sequence of the father (I1-P) was determined by sequencing of paternal genomic DNA to 39-fold coverage, yielding 1.8 x 10^6 heterozygous SNPs. However, paternal haplotypes could not be assessed because only relatively low molecular weight DNA obtained from saliva was available. Shotgun DNA sequencing libraries were also constructed from 5 mL of maternal plasma (obtained at 18.5 weeks gestation), and this composite of maternal and fetal genomes was sequenced to 78-fold nonduplicate coverage. The fetus was male, and fetal content in these libraries was estimated at 13% (Figure 2.2A). To properly assess the accuracy of our methods for determining the fetal genome solely from samples obtained noninvasively at 18.5 weeks gestation, we also performed shotgun genome sequencing of the child (I1-C) to 40-fold coverage via cord blood DNA obtained after birth.

Our analysis comprised four parts: (1) predicting the subset of 'maternal-only' heterozygous variants (homozygous in the father) transmitted to the fetus; (2) predicting the subset of 'paternal-only' heterozygous variants (homozygous in the mother) transmitted to the fetus; (3) predicting transmission at sites heterozygous in both parents; (4) predicting sites of *de novo* mutation – that is, variants occurring only in the genome of the fetus. Allelic imbalance in maternal plasma, manifesting across experimentally determined maternal haplotype blocks, was used to predict their maternal transmission (Figure 2.2B). The observation (or lack thereof) of paternal alleles in shotgun libraries derived from maternal plasma was used to predict paternal transmission (Figure 2.2C). Finally, a strict analysis of alleles rarely observed in maternal plasma, but never in maternal or paternal genomic DNA, enabled the genome-wide identification of candidate *de novo* mutations (Figure 2.2D). Fetal genotypes are trivially predicted at sites where the parents are both homozygous (for the same or dif-

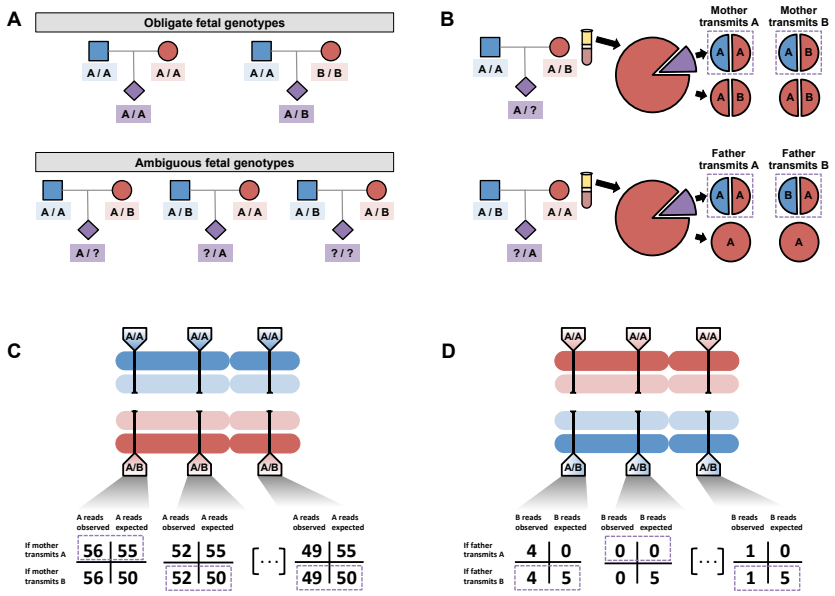


Figure 2.3: **Inference of the fetal genome on a site-by-site basis**. *(A)* Observed parental genotypes at a given site constrain the possible fetal genotypes. At the vast majority of sites, both parents are homozygotes and the fetal genotype is unambiguous. *(B)* Expected cfDNA makeup in the maternal plasma at maternal (top) and paternal (bottom) heterozygous sites. At these sites, either allele could be transmitted, yielding different expected allelic fractions. *(C)* Schematic of inference of fetal inheritance at maternal heterozygous sites. Numbers shown assume a constant sequencing depth of 100X at each site and do not represent real data. After sequencing the cfDNA, observed allele counts are compared to expected allele counts at each site, and in each case, the more likely scenario is chosen (purple boxes). *(D)* Schematic of inference of fetal inheritance at paternal heterozygous sites. Numbers are presented as in *(C)*. At each site, presence of the B allele is unexpected in cases where the father transmits the "A" allele. At the rightmost site, the single observed "B" allele could be evidence of true "B" transmission or an error introduced during the sequencing process.

ferent allele) (Figure 2.3A).

We first sought to predict transmission at 'maternal-only' heterozygous sites (Figure 2.3B-C). Given the fetal-derived proportion of ~13% in cell-free DNA, the maternal-specific allele is expected in 50% of reads aligned to such a site if it is transmitted, versus 43.5% if the allele shared with the father is transmitted. However, even with 78-fold coverage of the maternal plasma "genome", the variability of sampling is such that site-by-site prediction results in only 64.4% accuracy (Figure 2.4). We therefore examined allelic imbalance

across blocks of maternally heterozygous sites defined by haplotype-resolved genome sequencing of the mother (Figures 2.1 and 2.2B). As anticipated given the haplotype assembly N50 of 326 Kb, the vast majority of experimentally defined maternal haplotype blocks were wholly transmitted, with partial inheritance in a small minority of blocks (0.6%, n=72) corresponding to switch errors from haplotype assembly and to sites of recombination.

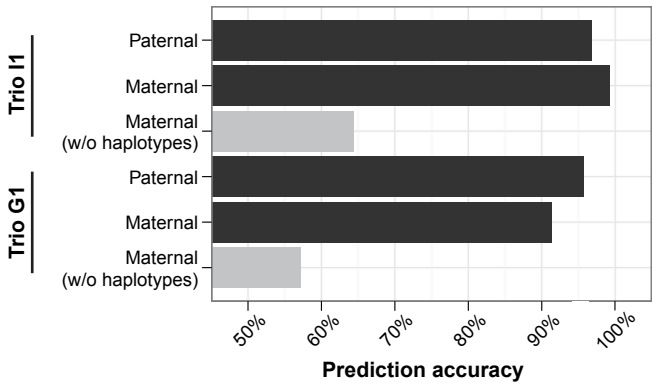


Figure 2.4: **Accuracy of fetal genotype inference from maternal plasma DNA sequencing**. Accuracy is shown for paternal-only heterozygous sites, and for phased maternal-only heterozygous sites, either using maternal phase information *(black)* or instead predicting inheritance on a site-by-site basis *(gray)*.

We developed a Hidden Markov model (HMM) to identify likely switch sites and thus more accurately infer the inherited alleles at maternally heterozygous sites (Figures 2.5 and 2.6). Using this model, accuracy of the inferred inherited alleles at 1.1 x 10^6 phased, 'maternal-only' heterozygous sites increased from 98.6% to 99.3% (Table 2.2). Remaining errors were concentrated among the shortest maternal haplotype blocks (Figure 2.7), which provide less power to detect allelic imbalance in plasma DNA data compared with long blocks. Among the top 95% of sites ranked by haplotype block length, prediction accuracy rose to 99.7%, suggesting that remaining inaccuracies can be mitigated by improvements in haplotyping.

As a result of this observation, we sought to improve both the length and the density of the haplotype blocks used for the fetal genome inference. We reasoned that, like the fetal

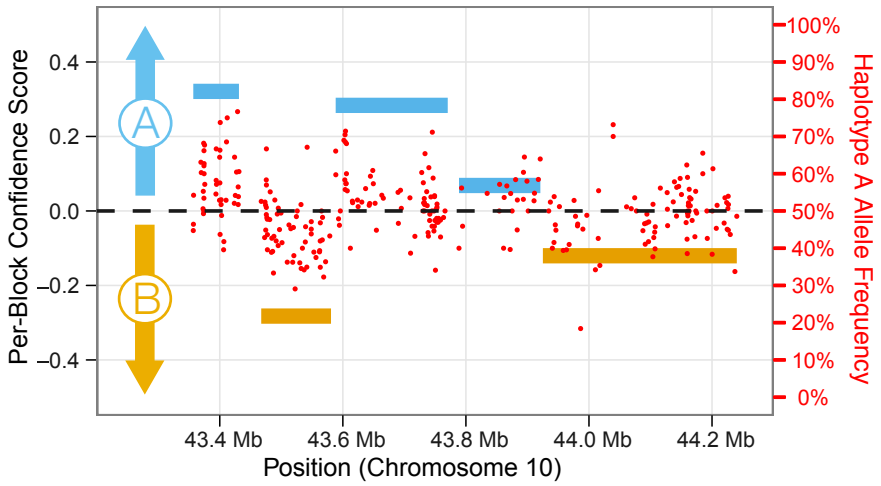


Figure 2.5: **Prediction of maternally transmitted alleles**. Hidden Markov model-based predictions (*blue and gold bars*) correctly predict maternally transmitted alleles across ~1 Mbp on chromosome 10, despite site-to-site variability of allelic representation among maternal plasma DNA sequences *(red)*.

genome is a mosaic of the parental haplotypes, the parental haplotypes are mosaics of population haplotypes. We used a reference panel of haplotypes inferred from population-level sequencing by the 1000 Genomes Project towards two goals: first, to phase a portion of the 8.6% of maternal heterozygous variants left out of the maternal haplotype blocks described above; and second, to stitch together maternal haplotype blocks into longer aggregates (Figure 2.8).

Using this approach in conjunction with the haplotype inference software BEAGLE, (Browning and Browning, 2007), we phased 73,257 additional maternally heterozygous sites initially left out of the maternal haplotype assembly, representing 66% of maternal unphased sites, and predicted the inheritance of those sites, using the HMM approach described above, with 96.0% accuracy. The slight reduction in accuracy at this set of sites relative to the genome-wide accuracy at directly phased, maternally heterozygous sites likely reflects a combination of errors in the population reference panel, incomplete IBD sharing, and the indirect, heuristic nature of our approach. To address this limitation, we developed

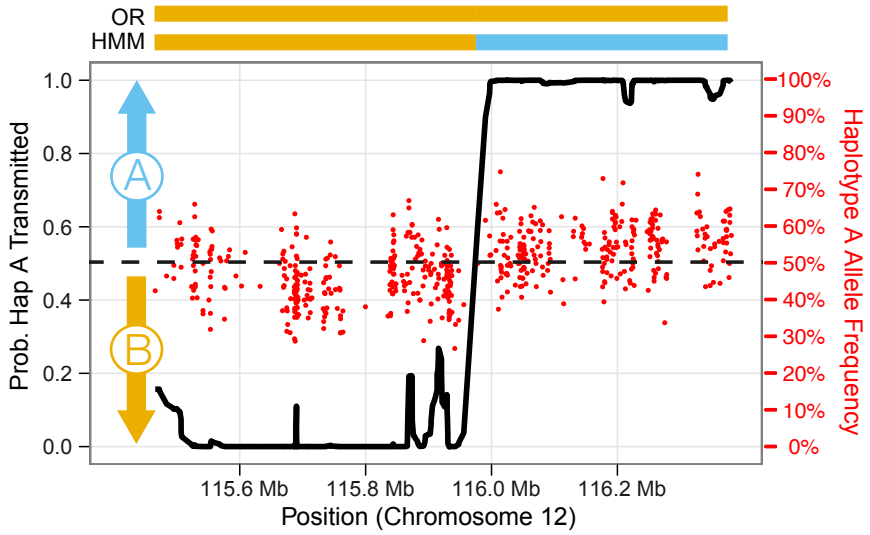


Figure 2.6: **Hidden Markov model-based detection of recombination events and haplotype assembly switch errors**. A maternal haplotype block of 917 Kbp on chromosome 12q is shown, with red points representing the frequency of haplotype A alleles among plasma reads, and the black line indicating the posterior probability of transmission for haplotype A computed by the HMM at each site. A block-wide odds ratio test (OR) predicts transmission of the entire haplotype B, resulting in incorrect prediction at 272 of 587 sites (46.3%). The HMM predicts a switch between chromosomal coordinates 115,955,900 and 115,978,082, and predicts transmission of haplotype B alleles from the centromeric end of the block to the switch point, and haplotype A alleles thereafter, resulting in correct predictions at all 587 sites. All three overlapping informative clones support the given maternal phasing of the SNPs adjacent to the switch site, suggesting that the switch predicted by the HMM results from a maternal recombination event rather than an error of haplotype assembly.

a confidence score for the phase assignment at each heterozygous site. When selecting only the top 90% of sites ranked by confidence score, accuracy improved to 97.7%. We next identified haplotype blocks containing the fewest sites, for which prediction accuracy was lowest (Figure 2.7). Using evidence from reference haplotypes, we merged 2,989 sites into neighboring, longer haplotype blocks, reducing the overall error rate of inheritance predictions at maternal heterozygous sites by 16%.

We performed simulations to characterize how the accuracy of haplotype-based fetal genotype inference depended upon haplotype block length, maternal plasma sequencing depth, and the fraction of fetal-derived DNA. To mimic the effect of less successful phasing,

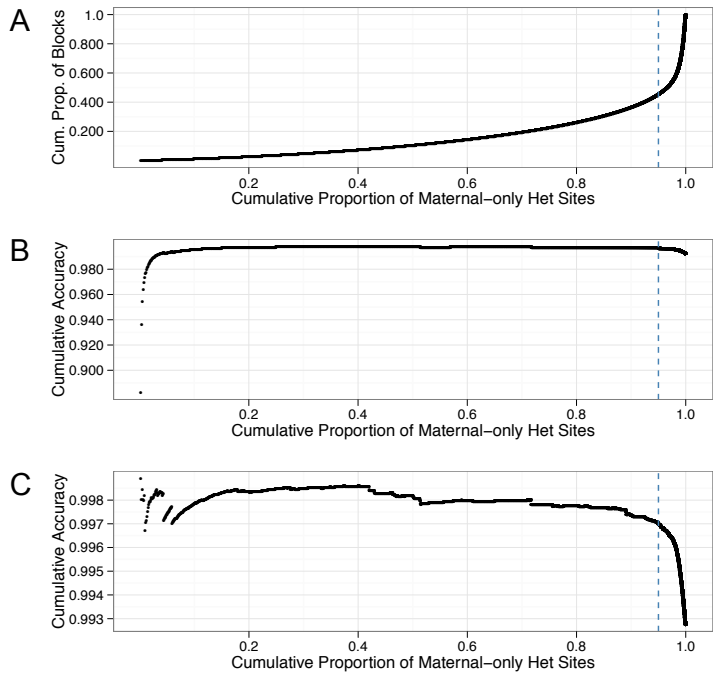


Figure 2.7: **Accuracy of maternal transmission inference, by haplotype block size**. Maternal-only heterozygous sites were ranked in decreasing order by haplotype block size (the number of other maternal-heterozygous sites phased in the same block). Blue dotted lines denote cutoffs retaining 95% of sites. *(A)* Cumulative distribution of maternal-only heterozygous sites by block; 95% of such sites are contained in the top 45% of haplotype blocks. *(B)* and *(C)*. Cumulative accuracy among maternal-only heterozygous sites ranked by block size. Cumulative accuracy is 99.7% among the 95% of sites in the largest haplotype blocks, and falls to 99.3% when the remaining 5% of sites are included. *(B)* Cumulative accuracy including all blocks. *(C)* Cumulative accuracy when removing the largest block, which resides among a duplication-rich region at 43.7 Mb - 44.3 Mb on chromosome 17q.

we split the maternal haplotype blocks into smaller fragments to create a series of assemblies with decreasing contiguity. We then subsampled a range of sequencing depths from the pool of observed alleles in maternal plasma, and predicted the maternally contributed allele at each site as above (Figure 2.9A). The results suggest that inference of the inherited allele is robust to either decreasing sequencing depth of maternal plasma, or to shorter haplotype blocks, but not both. For example, using only 10% of the plasma sequence data (median depth = 8X) in conjunction with full-length haplotype blocks, we successfully pre-

dicted inheritance at 94.9% of 'maternal-only' heterozygous sites. We achieved nearly identical accuracy (94.8%) at these sites when highly fragmented haplotype blocks (N50 = 50 Kb) were used with the full set of plasma sequences. We next simulated decreased proportions of fetal DNA in the maternal plasma by spiking in additional depth of both maternal alleles at each site and subsampling from these pools, effectively diluting away the signal of allelic imbalance used as a signature of inheritance (Figure 2.9B). Again, we found the accuracy of the model to be robust to either lower fetal DNA concentrations or shorter haplotype blocks, but not both.

We next sought to predict transmission at 'paternal-only' heterozygous sites (2.3D). At these sites, when the father transmits the shared allele, the paternal-specific allele should be entirely absent among the fetal-derived sequences. If instead the paternal-specific allele is transmitted, it will on average constitute half the fetal-derived reads within the maternal plasma "genome" (~5 reads given 78-fold coverage, assuming 13% fetal content). To assess these, we performed a site-by-site log-odds test; this amounted to taking the observation of one or more reads matching the paternal-specific allele at a given site as evidence of its transmission, and conversely the lack of such observations as evidence of nontransmission (Figure 2.2C). In contrast to maternal-only heterozygous sites, this simple site-by-site model was sufficient to correctly predict inheritance at 1.1×10^6 paternal-only heterozygous sites with 96.8% accuracy (Table 2.2). We anticipate that accuracy could likely be improved by deeper sequence coverage of the maternal plasma DNA (Figure 2.10), or alternatively by taking a haplotype-based approach if high molecular weight genomic DNA from the father is available.

We next considered transmission at sites heterozygous in both parents. We predicted maternal transmission at such shared sites phased using neighboring 'maternal-only' heterozygous sites in the same haplotype block. This yielded predictions at 576,242/631,721 (91.2%) of shared heterozygous sites with an estimated accuracy of 98.7% (Table 2.2). Although we did not predict paternal transmission at these sites, we anticipate that analogous to the case of maternal transmission, this could be done with high accuracy given paternal

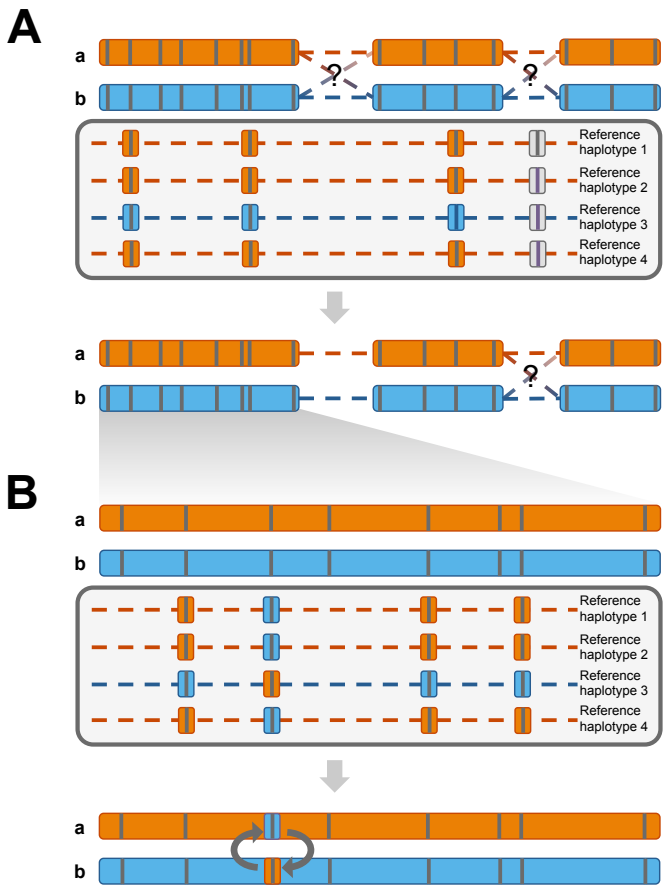


Figure 2.8: **Improving density and contiguity of haplotype blocks with reference panels**. *(A)* Dense methods for obtaining haplotypes produce phased haplotype blocks, here arbitrarily labeled a and b, that comprehensively encompass variation contained within the blocks. Contiguity between adjacent blocks, however, is unknown. Large reference panels of previously ascertained haplotypes, for example those produced by the 1000 Genomes Project, lack many of the rare or private variants phased by direct methods, but contain information about population-level linkage disequilibrium patterns between common variants. By modeling the directly phased sample as a mosaic of haplotypes segregating in the population, contiguity between pairs of nearby dense blocks can inferred. *(B)* Dilution pool-based methods may produce haplotype blocks containing a small number of common variants whose experimentally determined phase is incompatible with previously ascertained haplotypes, and may leave a portion of common variants unphased (Figure 2.1). By again using patterns of linkage disequilibrium observed in large or population-specific reference panels, the phase can be correctly assigned.

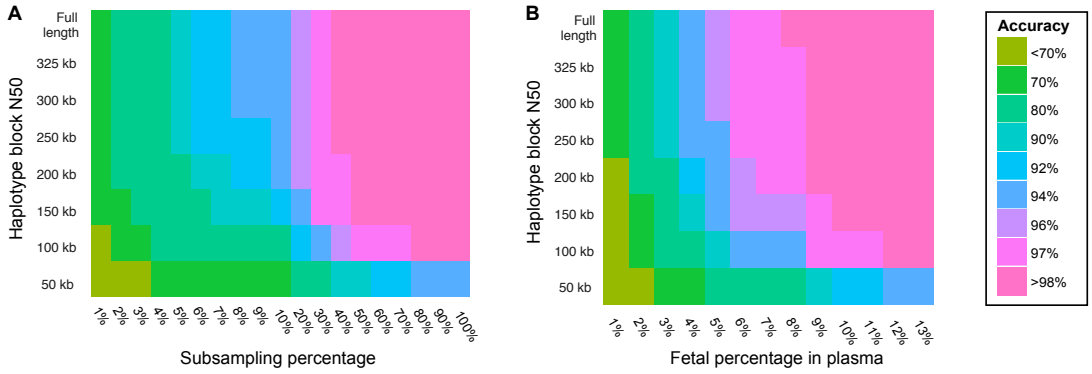


Figure 2.9: **Impact of coverage, haplotype length, and fetal fraction on statistical inference**. Simulation of effects of reduced coverage, haplotype length, and fetal DNA concentration on fetal genotype inference accuracy, defined as the percentage of sites at which the inherited allele was correctly identified out of all sites where prediction was attempted. Heatmaps of accuracy after in silico fragmentation of haplotype blocks and *(A)* shallower sequencing of maternal plasma, or *(B)* reduced fetal concentration among plasma sequences.

haplotypes. We note that shared heterozygous sites primarily correspond to common alleles (*data not shown*), which are less likely to contribute to Mendelian disorders in non-consanguineous populations.

True *de novo* mutations in the fetal genome are expected to appear within the maternal plasma DNA sequences as 'rare alleles' (Figure 2.2D), similar to transmitted paternal-specific alleles. However, the detection of *de novo* mutations poses a much greater challenge: unlike the 1.8×10^6 paternally heterozygous sites defined by sequencing the father (of which ~50% are transmitted), the search space for *de novo* sites is effectively the full genome, throughout which there may be only ~60 sites given a prior mutation rate estimate of $\sim 1 \times 10^{-8}$ (Conrad et al., 2011). Indeed, whole genome sequencing of the offspring (I1-C) revealed only 44 high-confidence point mutations ('true *de novo* sites'; Table 2.3). Taking all positions in the genome at which at least one plasma-derived read had a high-quality mismatch to the reference sequence, and excluding variants present in the parental whole genome sequencing data, we found 2.5×10^7 candidate *de novo* sites, including 39

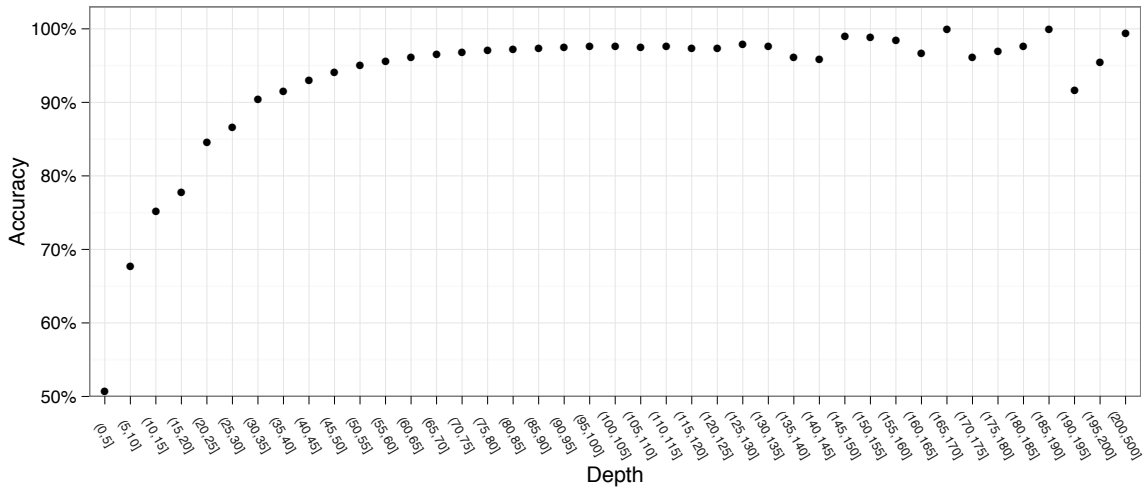


Figure 2.10: **Inference accuracy for paternal transmission at paternal-only heterozygous sites, by depth**. Inference accuracy is generally higher at more deeply sequenced sites. At sites with low overall coverage, too few fetal-derived reads are sampled, and the paternally transmitted allele is more likely to go unobserved. On the other end, sites with extremely high coverage may reside in regions of high copy number or dense repeat content that are recalcitrant to accurate mapping and variation calling with short reads.

of the 44 true *de novo* sites. At baseline, this corresponds to sensitivity of 88.6% with a signal-to-noise ratio of 1-to-6.4x10^5.

We applied a series of increasingly stringent filters (Figure 2.11) intended to remove sites prone to sequencing or mapping artifacts. We first removed alleles found in at least one read among any other individual sequenced in this study, known polymorphisms from dbSNP (release 135), and sites adjacent to 1-3mer repeats, reducing the number of candidate *de novo* sites to 1.8 x 10^7. We next filtered out sites with insufficient evidence (fewer than two independent reads supporting the variant allele, or variant base qualities summing to less than 105) as well as those with excessively many reads supporting the variant allele (uncorrected $P < 0.05$, per-site one-sided binomial test using fetal-derived fraction of 13%), cutting the total number of candidate sites to 3,884, including 17 true *de novo* sites. This candidate set is substantially depleted for sites of systematic error, and is instead likely dominated by errors originating during PCR, as even a single round of amplifi-

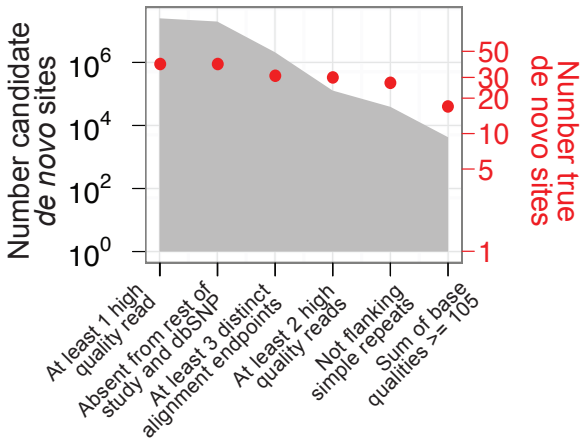


Figure 2.11: **Detecting sites of *de novo* mutation among maternal fetal plasma sequences**. Shown on a $\log_{10}$ scale are counts of candidate and true single-base *de novo* substitution variants remaining after application of successive quality filters *(filled gray area and red points, respectively)*.

cation with a proofreading DNA polymerase with an error rate of 1 x 10^{-7} would introduce hundreds of false positive candidate sites. Of note, this ~2,800-fold improvement in signal-to-noise ratio reduced the candidate set to a size that is an order of magnitude fewer than the number of candidate *de novo* sites requiring validation in a previous study involving pure genomic DNA from parent-child trios within a nuclear family (Roach et al., 2010). In a clinical setting, validation efforts would be targeted to those sites considered most likely to be pathogenic. For example, only 33 of the 3,884 candidate sites (0.84%) are predicted to alter protein sequence, and only a subset of these in genes associated with Mendelian disorders.

The improvements to the signal-to-noise ratio due to filtering were in no small part the result of careful fitting of a set of hard cutoffs to the observed sequencing data. These thresholds – for example, requiring a minimum sum of base quality scores supporting a particular variant – are unlikely to be generalizable to future data with different sequencing depths, fetal fractions, and error profiles. To attempt to address this limitation, we built a Support Vector Machine (SVM) classifier to distinguish between true *de novo* sites and errors. We

trained this classifier on high-confidence, paternally-inherited alleles at paternal-only heterozygous sites, which should mimic true *de novo* sites in the data, as well as a randomly selected set of unfiltered candidate *de novo* sites, likely to contain exclusively errors. We normalized all relevant features before training to plausibly enable classifier reuse across experiments. The performance of the resulting classifier was comparable to the results from the model built with manual filtering, achieving slightly higher sensitivity (43% vs. 38%) with slightly worse specificity (~6,000 false positives vs ~4,000) (Figure 2.12).

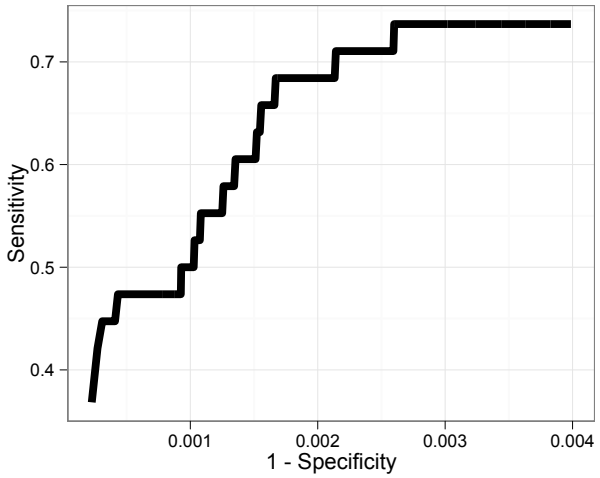


Figure 2.12: **SVM-based classification of candidate *de novo* mutations**. A Support Vector Machine-based classifier trained on inherited paternal alleles after feature normalization was used to classify candidate *de novo* mutations. Probability estimates for each site were generated using a 10-fold cross-validation approach. The sensitivity and specificity of the classifier for a range of classification thresholds is shown.

2.3 Discussion

We have demonstrated noninvasive prediction of the whole genome sequence of a human fetus through the combination of haplotype-resolved genome sequencing (Kitzman et al., 2011) of a mother, shotgun genome sequencing of a father, and deep sequencing of maternal plasma DNA. Notably, the types and quantities of materials used were consistent with those routinely collected in a clinical setting (Table 2.1). To replicate these results, we

repeated the full experiment for a second trio (G1) from which maternal plasma was collected earlier in the pregnancy, at 8.2 weeks after conception. Both the overall sequencing depth and the fetal-derived proportion were each lower relative to the first trio (by 28% and 51%, respectively), resulting in an average of fewer than four fetal-derived reads per site. Nevertheless, we achieved 95.7% accuracy for prediction of inheritance at maternal-only sites, consistent with accuracy obtained under simulation with data from the first trio (Figure 2.9). These results underscore the importance of specific technical parameters in determining performance, namely the length and completeness of haplotype-resolved sequencing of parental DNA, and the depth and complexity of sequencing libraries derived from low starting masses of plasma-derived DNA (less than 5 ng for both I1 and G1 in our study).

There remain several key avenues for improvement. First, although we predicted inheritance at 2.8×10^6 heterozygous sites with high accuracy (98.2% overall), there were 7.5×10^5 sites for which we did not attempt prediction (Table 2.2). These include 6.3×10^5 shared sites heterozygous in both parents for which we could not assess paternal transmission, and 1.2×10^5 "maternal-only" heterozygous sites which were not included in our haplotype assembly. The shared sites are in principle accessible but require haplotype-resolved (rather than solely shotgun) sequencing of paternal DNA, which was not possible here with either the I1 or G1 trio due to unavailability of high molecular weight DNA from each father. Figure 2.13 schematically illustrates a method for fetal genome inference, including at shared heterozygous sites, when paternal haplotypes are available. The unphased maternal sites are also in principle accessible but require improvements to haplotyping technology to enable phasing of SNPs residing within blocks of relatively low heterozygosity as well as within segmental duplications. More generally, despite recent innovations from our group and others (Fan et al., 2010; Kitzman et al., 2011; Ma et al., 2010; Yang et al., 2011), there remains a critical need for genome-wide haplotyping protocols that are at once robust, scalable, and comprehensive. Significant reductions in cost, along with standardization and automation, will be necessary for compatibility with large-scale clinical application.

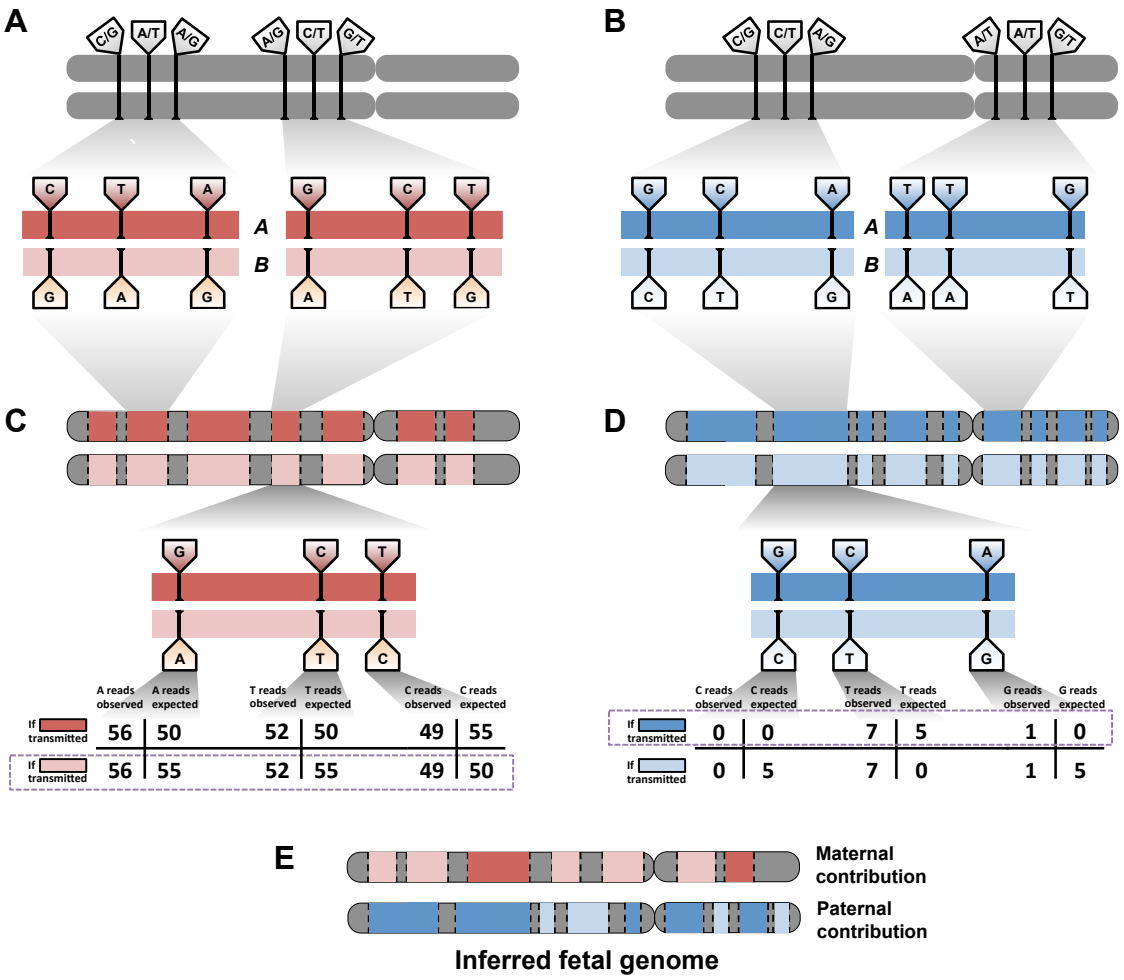


Figure 2.13: **Inference of the fetal genome from haplotype blocks**. *(a)*. Phasing of maternal heterozygous sites into haplotype blocks *(red bars)*. Haplotype blocks contain dozens or hundreds of such sites and cover over 300 kilobases on average. A single chromosome may have over 100 haplotype blocks; contiguity between blocks is not defined. Approximately 90% of all heterozygous sites are incorporated into haplotype blocks. Sites shown do not represent real data. *(b)*. Phasing of paternal heterozygous sites into haplotype blocks *(blue bars)*. Paternal and maternal blocks may overlap but are independently defined. *(c)*. Schematic of inference of fetal inheritance of maternal haplotype blocks. Numbers shown assume a constant sequencing depth of 100X at each. After sequencing the cfDNA, evidence of deviations from expected allele counts is aggregated over each site in haplotype blocks "A" and "B", and the more likely block is predicted. Block-level predictions in turn determine predictions at each contained site. The site in the center of the block would be incorrectly predicted if sites were considered independently; its inclusion in a haplotype block mitigates sampling noise and corrects the prediction. *(d)*. Schematic of inference of fetal inheritance of paternal haplotype blocks. Numbers are presented as in *(c)*. The observed "G" allele at the rightmost site, likely to cause an incorrect prediction if sites were considered independently, is now correctly identified as an error introduced during the sequencing process rather than evidence of transmission of the "G" allele. *(e)*. The inferred fetal genome is a composite of the parental haplotype blocks.

Second, although we were successful in detecting nearly 90% of *de novo* single nucleotide mutations by deep sequencing of maternal plasma DNA, this was with very low specificity. The application of a series of filters resulted in a ~2,800-fold gain in specificity at a ~2-fold cost in terms of sensitivity. However, there is clearly room for improvement if we are to enable the sensitive and specific prenatal detection of potentially pathogenic *de novo* point mutations at a genome-wide scale, a goal that will likely require deeper than 78-fold coverage of the maternal plasma "genome" (Ajay et al., 2011) in combination with targeted validation of potentially pathogenic candidate *de novo*mutations.

Third, our analyses focused exclusively on single nucleotide variants, which are by far the most common form of both non-pathogenic and pathogenic genetic variation in human genomes (Consortium, 2010; Stenson et al., 2009). Clinical application of noninvasive fetal genome sequencing will require more robust methods for detecting other forms of variation, e.g. insertion-deletions, copy number changes, repeat expansions, and structural rearrangements. Ideally, techniques for the detection of other forms of variation could derive from short sequencing reads in a manner that is directly integrated with experimental methods and algorithms for haplotype-resolved genome sequencing.

The ability to noninvasively sequence a fetal genome to high accuracy and completeness will undoubtedly have profound implications for the future of prenatal genetic diagnostics. Although individually rare, when considered collectively, the ~3,500 Mendelian disorders with a known molecular basis (Amberger et al., 2009) contribute substantially to morbidity and mortality (Bell et al., 2011). Currently, routine obstetric practice includes offering a spectrum of screening and diagnostic options to all women. Prenatal screening options have imperfect sensitivity and focus mainly on a small number of specific disorders, including trisomies, major congenital anomalies, and specific Mendelian disorders. Diagnostic tests, generally performed through invasive procedures, such as chorionic villus sampling and amniocentesis, also focus on specific disorders and confer risk of pregnancy loss that may inversely correlate with access to high-quality care. Noninvasive, comprehensive diagnosis of Mendelian disorders early in pregnancy would provide much more information

to expectant parents, with the greater accessibility inherent to a noninvasive test and without tangible risk of pregnancy loss. The less tangible implication of incorporating this level of information into prenatal decision-making raises many ethical questions that must be considered carefully within the scientific community and on a societal level. A final point is that as in other areas of clinical genetics, our capacity to generate data is outstripping our ability to interpret it in ways that are useful to physicians and patients. In other words, although the noninvasive prediction of a fetal genome may be technically feasible, its interpretation – even for known Mendelian disorders – will remain a major challenge (Cooper and Shendure, 2011).

2.4 Methods

2.4.1 Whole-genome shotgun library preparation and sequencing

Genomic DNA was extracted from whole blood, as available, or alternatively from saliva, with the Gentra Puregene Kit (Qiagen), or OrageneDx (DNA genotek), respectively. Purified DNA was fragmented by sonication with a Covaris S2 instrument (Covaris). Indexed shotgun sequencing libraries were prepared using the Kapa Library Preparation Kit (Kapa Biosystems), following the manufacturer's instructions. All libraries were sequenced on HiSeq 2000 instruments (Illumina) using paired-end 101 bp reads with an index read of 9 bp.

2.4.2 Maternal plasma library preparation and sequencing

Maternal plasma was collected by standard methods and split into 1 ml aliquots which were individually purified with the Qiaamp Circulating Nucleic Acids kit (Qiagen). DNA yield was measured with a Qubit fluorometer (Invitrogen). Sequencing libraries were prepared with the ThruPlex-FD kit (Rubicon Genomics), comprising a proprietary series of end-repair, ligation, and amplification reactions. Index read sequencing primers compatible with the whole-genome sequencing and fosmid libraries from this study were included during sequencing of maternal plasma libraries to permit detection of any contamination from other libraries. The percentage of fetal-derived sequences was estimated from plasma sequences by counting alleles specific to each parent as well as sequences mapping specifically to the Y chromosome.

2.4.3 Maternal haplotype resolution via clone pool dilution sequencing

Haplotype-resolved genome sequencing was performed essentially as previously described (Kitzman et al., 2011), with minor updates to facilitate processing in a 96-well format. Briefly, high molecular weight DNA was mechanically sheared to mean size ~38 kbp using a Hydroshear instrument (DigiLab), with the following settings: volume=120 ul, cycles=20,

speed code=16. Sheared DNA was electrophoresed through 1% Low Melting Point Ultra-Pure agarose (Invitrogen) with the buffer (0.5X TBE) chilled to 16˚C, using the following settings on a BioRad ChefDR-II pulsed field instrument: 170V, initial A=1, final A=6. After running for 17 h, lanes containing size markers (1 kbp extension ladder, Invitrogen) were excised, stained with SYBR Gold dye (Invitrogen), and placed alongside the unstained portion of the gel on a blue light transilluminator. The band between 38-40 kbp was then excised, melted for 10 min in a 70˚C water bath, spun at 15,000 rpm to pellet debris, and incubated at 47˚C for 1 h with 0.5 units beta-agarase (Promega) per 200 mg gel to digest the agarose. Sheared, size-selected DNA was precipitated onto Ampure XP beads (Beckman Coulter) as follows: 100 ul of beads in the supplied buffer were supplemented with additional binding buffer (2.5M NaCl + 20% PEG 8000) to match the volume of the digested gel and DNA. The beads and buffer were then gently mixed with the DNA/agarase reaction mixture, pelleted and rinsed following the manufacturer's directions, and finally eluted into 60 ul H2O. DNA was next end-repaired with the End-IT kit (Epicentre), cleaned up by precipitation onto 30ul Ampure XP beads supplemented with 70ul additional binding buffer, and eluted into 12ul H2O. Ligation to the fosmid vector backbone pCC1Fos and clone packaging were conducted as previously described using the CopyControl Fosmid Construction Kit (Epicentre). A single bulk infection per maternal sample was performed using each phage library and each was then split by dilution into 1.5 ml cultures (LB + 12.5ug/ml chloramphenicol) across a deep-well 96-well plate. The resulting master culture was grown overnight at 37˚C shaking at 225 rpm. The following day, subcultures were made by in 96-well plates by adding 200ul inoculum from each master culture well into fresh outgrowth media (LB + 12.5ug/ml chloramphenicol + 1X final autoinduction solution) to a final volume of 1.5ml per well. After overnight outgrowth (37˚C, 225 rpm shaking), clone pool DNA was extracted by alkaline lysis mini-preparation in 96 well plates, following standard procedures. Indexed Illumina sequencing libraries were prepared in sets of 96 using the Nextera library preparation kit as previously described (Adey et al., 2010), followed by library pooling and size selection to 350-650 bp.

2.4.4 Variant calling

Reads were split by index, allowing up to edit distance of 3 to the known barcode sequences, and then mapped to the human reference genome sequence (hg19) using bwa v0.6.1. After removing PCR duplicate read pairs using the Picard toolkit (http://picard.sourceforge.net/), local realignment around indels, variant discovery, quality score recalibration and filtering to 99% estimated sensitivity among known polymorphisms was performed using the Genome Analysis Toolkit (DePristo et al., 2011) using "best practices" parameters provided by the software package's authors (http://www.broadinstitute.org/gsa/wiki/index.php).

Of note, there was no evidence of uniparental disomy or a numerical chromosomal abnormality; the former would have manifested as a large volume of chromosome-specific Mendelian errors, whereas the latter would have been detected by chromosome-specific read-depth imbalance.

2.4.5 Haplotype assembly

Reads were split per dilution pool by barcode, and a sliding-window read depth measure was used to infer clone positions (Kitzman et al., 2011). Using custom scripts, clone pool reads were re-genotyped against heterozygous SNPs ascertained by shotgun sequencing, and overlapping clones from different pools were assembled into haplotype blocks with a custom implementation of the HapCUT algorithm (Bansal and Bafna, 2008).

2.4.6 Inference of the fetal genome sequences

A Hidden Markov model (HMM) was constructed to infer the inherited maternal allele at each maternal-specific heterozygous site. The model's latent state defines which of the two phased maternal haplotype blocks is inherited at each site, with a third state representing a between-block region at which phase is unknown. The HMM emits allele counts at each phased site, with probabilities given by binomial distribution parameterized as follows: if the maternally inherited allele is identical to the paternal (homozygous) allele at a given

"maternal-only" heterozygous site, the probability of observing k such alleles among N total reads with fetal percentage F is

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

where the first term in the second binomial parameter represents the expected allele balance in the maternally-derived DNA in the maternal plasma, the second term represents the expected contribution of the paternal allele via the fetus, and the third term represents the expected contribution of the inherited maternal allele via the fetus.

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

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

Inferred transitions within phased blocks represent either true recombination events or switch errors in maternal phasing. Transition probabilities within phased blocks were held constant at 10^{-5}; changing this parameter did not substantially affect either the number of inferred transitions within blocks nor the final accuracy. Finally, the most probable path through the observed data was determined using the Viterbi algorithm for inference of the latent state at each site, corresponding to a prediction of the inherited maternal allele. Prediction accuracy was determined by comparing the predicted to actual inheritance determined from the offspring's genotype.

Inheritance at "paternal-only" heterozygous sites was predicted using a binomial model. At each such site, either the paternal-specific allele or the allele shared with the mother can be transmitted. Let F represent the fetal DNA concentration in the maternal plasma and N represent the depth at a given site. If the paternal-specific allele is transmitted, we expect to observe it in $N \times \frac{F}{2}$ times in the maternal plasma. Similarly, if the paternal-specific allele is not transmitted, we expect to observe it 0 times. The likelihoods of observing K such alleles from N total under each inheritance models were compared, and prediction was determined by choosing the model that yielded a higher likelihood.

At each shared heterozygous site (i.e., heterozygous in both parents), the maternally contributed allele was predicted based on the inferred inheritance of the block in which the site is situated, as determined by "maternal-only" heterozygous sites within the same block. In the rare event that a block was identified to be partially inherited, either due to a real recombination event or a switch error in phasing, the inferred inheritance of the nearest "maternal-only" heterozygous site within the block was used to assign a prediction.

True *de novo* mutations in each offspring were identified from the trio shotgun whole genome sequences as follows: starting with all sites called as heterozygous in the offspring and homozygous reference in both parents, known variants were removed (dbSNP v135 and 1000 Genomes Pilot 1 sites), as were sites with low coverage in either parent (< 15 reads for the I1 trio, < 10 reads for the G1 trio). Candidate *de novo* alleles present at high quality positions in at least one read in any other individual (Phred-scaled base quality >= 10 and mapping quality >= 20) were removed. Finally, a minimum variant quality score threshold of 230 was applied. *De novo* mutations were validated by PCR and direct capillary sequencing (Table 2.3).

2.4.7 Downsampling

The effect of reduced fetal contribution to the maternal plasma sequences was investigated by diluting the fetal-specific sequences *in silico* and reanalyzing the modified data. Simulated dilution of fetal content was carried out as follows. At each maternal-specific heterozygous site, alleles A and B were observed with counts N_A and N_B among the full dataset, with $N_{TOTAL} = N_A + N_B$. For a given dilution coefficient $\frac{D}{F}$ where $0 < D < F$, the total pool of observed counts was diluted by first increasing N_{TOTAL} by a factor of $\frac{F}{D}$, with additional counts allocated by assigning each new allele randomly to N_A or N_B with equal probability, and then sampling counts from the temporarily expanded pool by discarding each allele from N_A and N_B with probability $1-\frac{D}{F}$. Updated counts and fetal content estimates were used as input into the Hidden Markov model described above. Reduced coverage within plasma data was separately simulated by subsampling a portion of the observed counts at

each site. For a given proportion S, each observed base was discarded with probability $1-S$. Updated counts were then used as input into the Hidden Markov model as described.

Acknowledgements

We thank C. Lee, B. Munson, D. Nickerson and the Northwest Genomics Center (U.W.) for assistance with sequencing; J. Langmore and E. Kamberov (Rubicon Genomics) for early access to reagents; M. McMillin (U.W.) for sample coordination; B. Browning (U.W.) for helpful discussion and early access to software; and members of the Shendure Lab for helpful discussions.

Table 2.1: **Summary of sequencing**. Individuals sequenced, type of starting material, and final fold-coverage of the reference genome after discarding PCR or optical duplicate reads. GA, gestational age.

Individual	**Biological sample**	**Depth of coverage**
Mother (I1-M)	Plasma (5 ml, GA 18.5 wk)	78
	Whole blood (<1 mL)	32
Father (I1-P)	Saliva	39
Offspring (I1-C)	Cord blood at delivery	39

Table 2.2: **Accuracy of fetal genome inference**. Number of sites and accuracy of fetal genotype inference from maternal plasma sequencing (percentage of transmitted alleles correct out of all predicted) by parental genotype and phasing status. Sites later determined by trio sequencing (including the offspring) to have poor genotype quality scores or genotypes that violated Mendelian inheritance were discarded the purpose of evaluating accuracy (14,000 maternal-only, 32,233 paternal-only, and 480 shared heterozygous sites, or 1.5% of all sites). †Among biparentally heterozygous sites, accuracy was assessed only where the offspring was homozygous (48.8%, n=631,721), allowing the "true" transmitted alleles to be unambiguously inferred from trio genotypes.

Individual	**Site**	**Other parental genotype**	**Number of sites**	**Accuracy**
Mother (I1-M)	Heterozygous, phased	Homozygous	1,064,255	99.3%
		Heterozygous	576,242	98.7% †
	Heterozygous, not phased	All	121,425	N.D.
Father (I1-P)	Heterozygous	Homozygous	1,134,192	96.8%
		Heterozygous	631,721	N.D.

Table 2.3: ***De novo* point mutations identified by whole-genome shotgun sequencing**. Each event in trio "I1" was targeted for validation by PCR and direct capillary sequencing. Amplification and sequencing succeed at 35 of 44 sites; of those, all 35 validated as true *de novo* point mutations (i.e., offspring heterozygous and parents homozygous for reference allele).

Chrom.	Position	Ref. Allele	Alt. Allele	Validated?	Genes overlapped
chr1	14827232	A	C	Yes	Intergenic
chr1	21959596	G	A	Yes	Intron of RAP1GAP
chr1	62642578	C	T	Yes	Intergenic
chr1	158061739	G	A	Yes	Intron of KIRREL
chr1	176538426	C	T	Yes	Intron of PAPPA2
chr1	197602948	G	A	Yes	Intron of DENND1B
chr2	32296201	A	T	Yes	Intron of SPAST
chr2	58060266	T	C	Assay failed	Intergenic
chr2	135596281	T	C	Yes	ACMSD exon 1, p.Leu10Pro
chr2	238760708	G	T	Yes	Intergenic
chr3	17614899	C	T	Yes	Intron of TBC1D5
chr3	18023875	C	T	Yes	Intergenic
chr3	36828198	T	C	Yes	Intergenic
chr3	79639506	G	A	Assay failed	Intron of ROBO1
chr3	188400668	G	A	Assay failed	Intron of LPP
chr4	28535313	G	C	Yes	Intergenic
chr4	32286182	G	A	Yes	Intergenic
chr4	38294675	G	A	Yes	Intergenic

Continued on Next Page...

Table 2.3 – Continued

Chrom.	Position	Ref. allele	Alt. allele	Validated?	Genes overlapped
chr5	5059500	T	C	Yes	Intergenic
chr5	19601463	T	A	Yes	Intron of CDH18
chr5	133747799	G	T	Yes	Intergenic
chr6	63446051	T	G	Yes	Intergenic
chr7	77442453	A	G	Assay failed	Intron of PHTF2
chr7	85735259	T	C	Yes	Intergenic
chr9	18393440	A	G	Yes	Intergenic
chr9	31764904	G	C	Assay failed	Intergenic
chr9	36929059	A	G	Yes	Intron of PAX5
chr9	38730375	G	A	Assay failed	Intergenic
chr10	92799212	G	A	Yes	Intergenic
chr11	18977014	A	G	Yes	Intergenic
chr11	74729193	C	T	Yes	Intergenic
chr13	43751146	G	T	Yes	Intergenic
chr14	27178898	A	G	Yes	Intergenic
chr14	38414572	G	A	Yes	Intergenic
chr15	59850650	C	T	Assay failed	Intergenic
chr15	67184470	C	T	Yes	Intergenic
chr16	74871390	G	C	Yes	Intergenic
chr17	38242084	C	G	Assay failed	Intron of THRA
chr18	67786028	C	T	Yes	Intron of RTTN
chr21	32940951	A	C	Yes	Intergenic

Continued on Next Page...

Table 2.3 – Continued

Chrom.	Position	Ref. allele	Alt. allele	Validated?	Genes overlapped
chr21	41369734	C	T	Yes	Intergenic
chr22	26530011	C	T	Assay failed	Intergenic
chr22	47178447	A	T	Yes	Intron of TBC1D22A
chrX	47330402	C	T	Yes	Intron of ZNF51

Chapter 3

COPY-NUMBER VARIATION AND FALSE POSITIVES IN PRENATAL ANEUPLOIDY SCREENING

This chapter has been adapted with minor changes from: Snyder, MW; Simmons, LE; Kitzman, JO; Coe, BP; Henson, JM; Daza, RM; Eichler, EE; Shendure, J; and Gammill, HS. Copy-number variation and false positive prenatal aneuploidy screening results. *New England Journal of Medicine* 372(17):1639-45 (2014).

In the published manuscript, I share first author credit with Dr. Simmons. My specific contributions to this project were the processing of biological samples, preparation of sequencing libraries, data analysis, statistical modeling, preparation of all figures excluding 3.3, and writing the first draft of the manuscript.

Abstract

Investigations of noninvasive prenatal testing (NIPT) for aneuploidy by analysis of circulating cell-free DNA have demonstrated high sensitivity and specificity in high- and low-risk cohorts. However, the overall low incidence of aneuploidy limits the positive predictive value of these tests. Currently, the causes of false-positive results are poorly understood. We investigated four pregnancies with discordant prenatal test results, and found in two cases that maternal duplications on chromosome 18 were the likely cause of the discordant results. Modeling based on population-level copy number variation supports the possibility that some false-positive results with NIPT may be attributable to maternal copy number gains.

3.1 Introduction

Methods for noninvasive prenatal testing (NIPT) have rapidly matured in clinical practice, with aneuploidy screening based on analysis of circulating cell-free DNA (cfDNA) now routinely offered to women with high-risk pregnancies. Due to the high reported accuracy of these tests (Ashoor et al., 2012; Palomaki et al., 2012), attention has shifted to low-risk cohorts, where reduced incidence of aneuploidy may limit the positive predictive value (PPV) of NIPT (Nicolaides et al., 2012). A recent prospective analysis of cfDNA-based NIPT in 1,914 low-risk pregnancies reported false-positive rates of 0.3%, 0.2%, and 0.1% for trisomies 21, 18, and 13, respectively – outperforming standard screening (Bianchi et al., 2014). However, the PPVs were 45.5% and 40.0% for trisomies 21 and 18 (Bianchi et al., 2014), respectively, highlighting the need for follow-up diagnostic testing. Norton et al. report higher PPVs for cfDNA-based NIPT with a different method, albeit with a higher "no call" rate that may include ambiguous results (Norton et al., 2015).

Mechanisms underlying false-positive cfDNA-based NIPT results remain incompletely elucidated (Mennuti et al., 2013). Explanatory hypotheses include maternal mosaicism (Lau et al., 2013; Wang et al., 2013), undetected tumors (Osborne et al., 2013), vanishing twin syndrome (Lau et al., 2014) or confined placental mosaicism (CPM) (Grati et al., 2014; Mao et al., 2014), as well as technical errors. While case reports have documented examples of underlying causes of both false positives and other aberrant results, only a small proportion have been comprehensively explained (Lau et al., 2013).

Methodologies for cfDNA-based NIPT include massively multiplex PCR (Zimmermann et al., 2012), shotgun sequencing (Chiu et al., 2008; Fan et al., 2008), or targeted sequencing (Sparks et al., 2012). Illumina Verifi and Sequenom MaterniT21 PLUS are based on counting statistics naturally arising from shotgun sequencing of total cfDNA in maternal plasma. After isolation, sequencing, and alignment of cfDNA fragments, a minority of which are feto-placentally derived (mean 13%, but with considerable variance during and between pregnancies (Nygren et al., 2010)), the reads are sorted into bins. Each bin con-

tains reads unambiguously derived from a specific chromosome, and the distributions for each chromosome are converted to standard normal distributions. A newly analyzed cfDNA sample is compared to reference distributions, yielding per-chromosome z-scores that estimate the likelihood of fetal aneuploidies. In diploid pregnancies, false-positive detection of trisomy may occur due to infrequent sampling of extreme values – typically z-scores above 4.0 – from the normalized distributions of reads derived from the relevant chromosomes. In statistical terms, this probability is given by `Pr(Z > 4.0)`, or roughly 3 in 100,000.

This approach implicitly assumes that every woman carries the same proportion of genetic material on a given chromosome. In fact, chromosomes vary slightly in composition and size between individuals due to inherited or *de novo* copy number variants (CNVs), in which a genomic region is deleted or duplicated. For example, a maternal duplication effectively increases the length of the chromosome on which it resides, thereby increasing the proportion of cfDNA derived from that chromosome. In such an individual, sequencing of cfDNA would yield overrepresentation of reads deriving from the CNV-containing chromosome relative to reference individuals, potentially leading to false interpretation as fetal trisomy (Figure 3.1).

The capacity of a maternal CNV to alter interpretation of NIPT is augmented by the fact that the vast majority of cfDNA is maternally derived. In a diploid pregnancy in which the mother carries a duplication, the increased number of reads derived from the additional copy of the duplicated region shifts the sampling distribution for this pregnancy to the right relative to the underlying reference distribution (Figure 3.2). The probability of a false-positive statistical test would then exceed `Pr(Z > 4.0)`, with the extent of excess driven primarily by the size of the duplication.

We sought to investigate whether maternal CNVs could give rise to false-positive NIPT results. As a proof-of-principle, we enrolled four pregnant women who had discordant findings: positive cfDNA-based screening with normal clinical outcomes. Subsequently, we modeled the potential population-level impact of maternal CNVs on false-positive results.

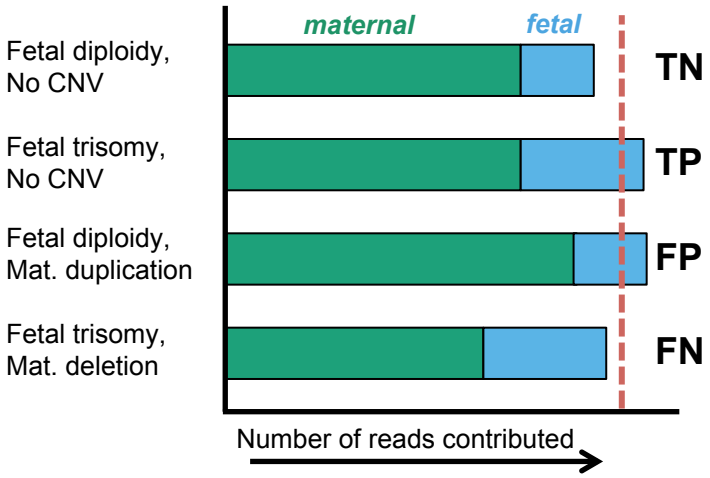


Figure 3.1: **Schematic representation of cfDNA analysis**. Schematic representation of cfDNA analysis. The cfDNA in maternal plasma contains primarily maternal cfDNA (green) and a smaller proportion of fetal cfDNA (blue). The threshold for triggering a positive cfDNA test is indicated by the vertical dashed line. From top: First, the combination of fetal diploidy and the absence of a maternal copy number variant (CNV) results in a true negative test (TN). Second, fetal trisomy results in a true positive test (TP). Third, the combination of fetal diploidy and the presence of a maternal CNV duplicating a portion of a relevant chromosome results in a false positive test (FP). Fourth, hypothetically, the combination of fetal trisomy on a specific chromosome and the presence of a maternal CNV deleting a portion of the same chromosome could result in a false negative test (FN).

3.2 Results

We enrolled four subjects, each with discordant NIPT results and clinical findings. In each case, NIPT was performed by Illumina Verifi. Three subjects received screen results positive for trisomy 18 and one for trisomy 13 (complete clinical information in Table 3.3). In two of the three cases screen-positive for trisomy 18, we identified maternal CNVs on chromosome 18 (Figure 3.3).

Case 1 was a 35-year-old primigravida. NIPT at 18 gestational weeks reported fetal trisomy 18. Ultrasound at 20 gestational weeks was consistent with normal fetal anatomy and concordant biometry. Diagnostic testing by genetic amniocentesis was consistent with a diploid male pregnancy. The remainder of the pregnancy was uncomplicated and a healthy male infant was delivered at term.

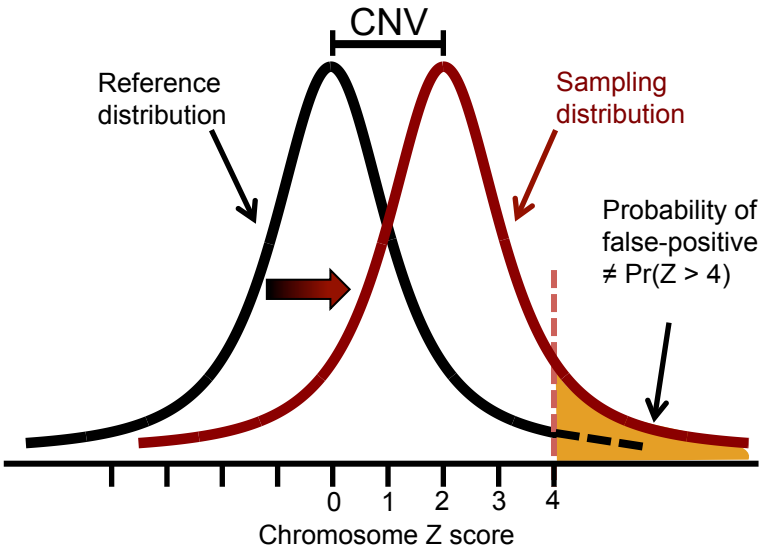


Figure 3.2: **Maternal CNVs and sampling distributions**. Schematic representation of the impact of a maternal CNV on the probability of a false positive test result. Maternal duplications shift the test's sampling distribution to the right, while the underlying reference distribution is unchanged.

Case 3 was a 34-year-old multigravida. NIPT at 12 gestational weeks was consistent with fetal trisomy 18. At 12, 16 and 20 gestational weeks, fetal ultrasound examinations demonstrated normal anatomy and concordant biometry. Genetic amniocentesis was declined. The remainder of the pregnancy was uncomplicated and a healthy female infant was delivered at term.

Analysis of Case 1 cfDNA identified a duplicated region on chromosome 18 containing portions of 18p11.31 and 18p11.23 (1.15 Mb). Analysis of Case 3 cfDNA identified a duplication on chromosome 18 covering a region of 18p11.31 (487 kb) (Figure 3.3). For both cases, maternal DNA from PBMC was used to validate the CNV by PCR and Sanger sequencing (Table 3.4, Figures 3.5 and 3.6).

To model the effect of these duplications on the risk of false-positive NIPT results, we calculated the theoretical fold-increase in the probability of false-positive results for a range of CNV sizes on chromosomes 13, 18, and 21 (Figures 3.7 and 3.8). The calculated increase

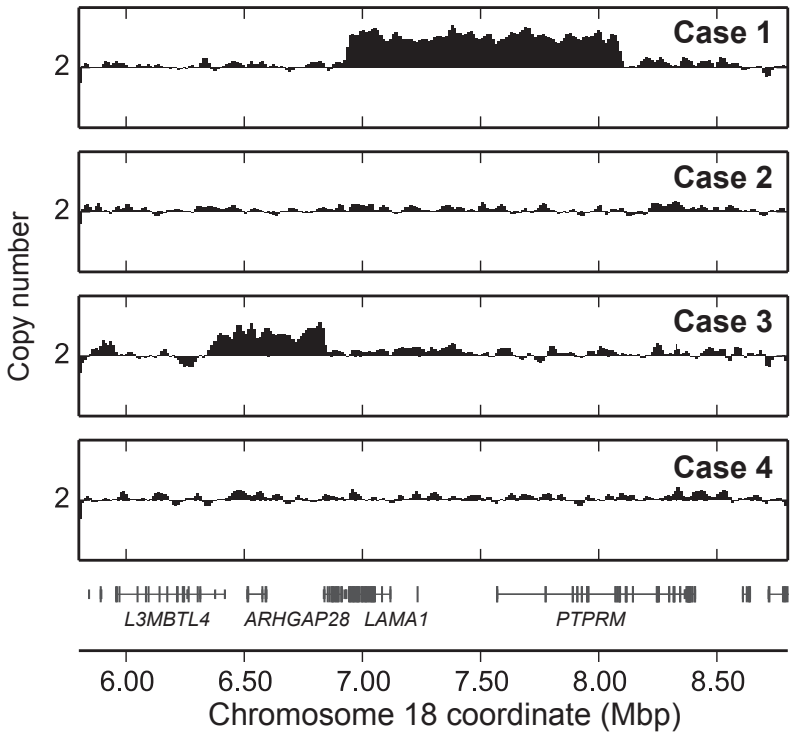


Figure 3.3: **Copy number profile in pregnancy cohort**. Copy number profile based on normalized cfDNA read-depth demonstrating duplicated sequence on chromosome 18 in two of four analyzed cases. Profiles of Cases 2 and 4 are consistent with two copies throughout the region of interest. Cases 1 and 3 each demonstrate increased copy number in contiguous regions, suggestive of duplications.

depends on several factors, including total number of reads per sample, coefficient of variation for the chromosome in question (Rava et al., 2013), fetal fraction, and fetal inheritance of the maternal CNV (Table 3.5). As the fetal fraction increases, the signal of overrepresentation is dampened if the CNV is not transmitted to the fetus, and increasingly large duplications are necessary to reach the same fold increase in probability of a false-positive. Conversely, if the CNV is transmitted to the fetus, the maternally inherited chromosome also contributes to the signal of overrepresentation, obviating dependency on the fetal fraction. We estimate that the CNV present in Case 1, duplicating 1.15 Mb and inherited by

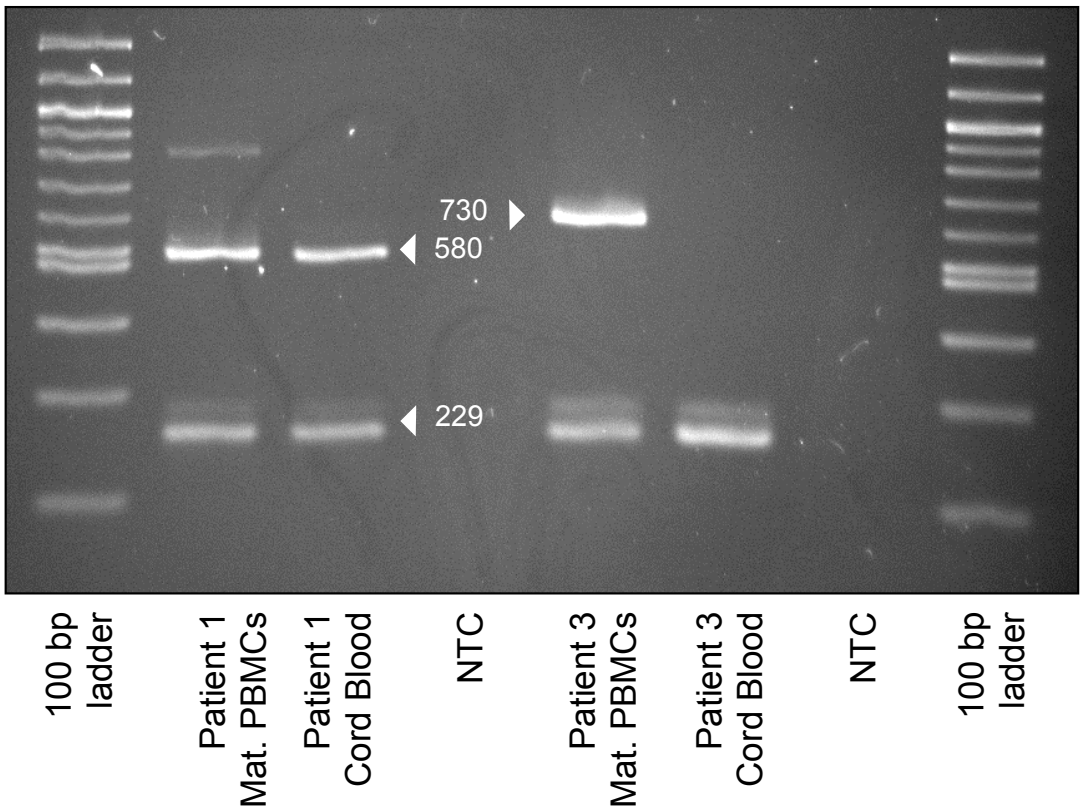


Figure 3.4: **Validation of CNVs with multiplex PCR**. PCR primers were designed to yield a product of the expected size in the event of a tandem duplication in the region of interest. Patient 1 demonstrated the expected 730 bp product in maternal PBMC DNA; the infant inherited the CNV. Patient 3 demonstrated the expected 580 bp product in maternal PBMC DNA; the infant did not inherit the CNV. Additional PCR primers for chromosome 18 were designed to yield a 229 bp product for all samples as a positive control. PCR products were purified and Sanger sequenced for breakpoint confirmation (Figures 3.5 and 3.6). NTC: no-template control.

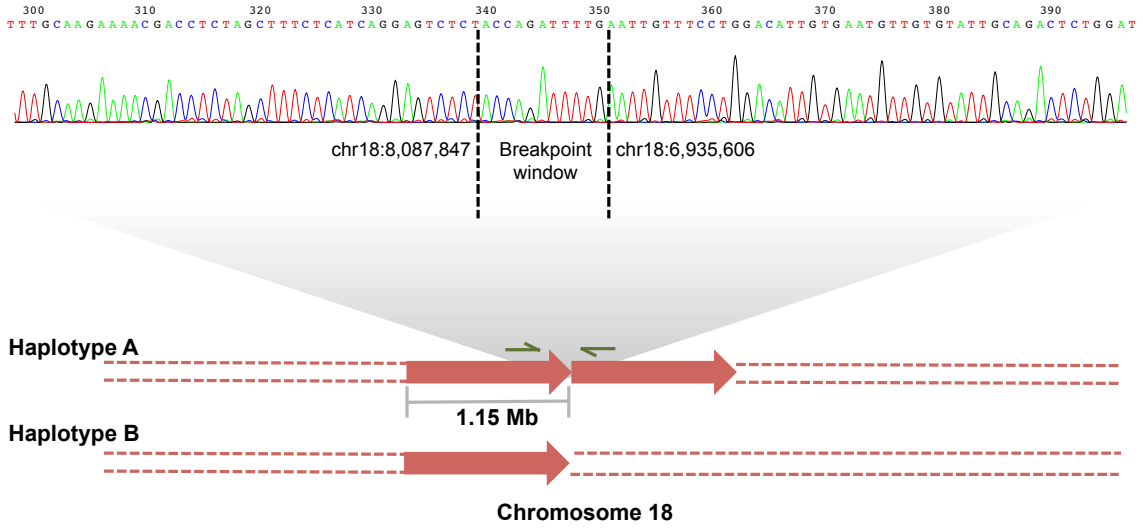


Figure 3.5: **Structure and validation of maternal duplication in Patient 1**. Structure and validation of maternal duplication in Patient 1. A schematic of 18p11.31 and 18p11.23 is shown at bottom, with red block arrows representing copies of the 1.15Mb duplicated region. Green arrows represent placement of primers for PCR and Sanger sequencing of the breakpoint window (meaning the window to which the breakpoint has been narrowed; the precise breakpoint cannot be determined with single nucleotide resolution due to microhomology). A partial Sanger sequencing trace of the PCR amplicon using the forward primer is shown.

the fetus, increased the probability of a false-positive statistical test on chromosome 18 approximately 15,650-fold, such that in the absence of fetal aneuploidy, the test was nearly equivalent to flipping a coin. The 487 kb CNV present in Case 3, but not inherited by the fetus, had a more modest estimated effect, yielding a 128- to 262-fold increase in the probability of false-positive results for plausible fetal fractions between 5% and 20%.

We identified two population factors that contribute to the impact of maternal CNVs. First, the distribution of CNV sizes varies by chromosome length, with chromosomes 13 and 18 having excesses of large duplications relative to the smaller chromosome 21 (Figure 3.7). Chromosomes with higher CNV burdens should be more susceptible to false-positive results caused by maternal CNVs, suggesting that false-positive calls, consequent to maternal CNV,

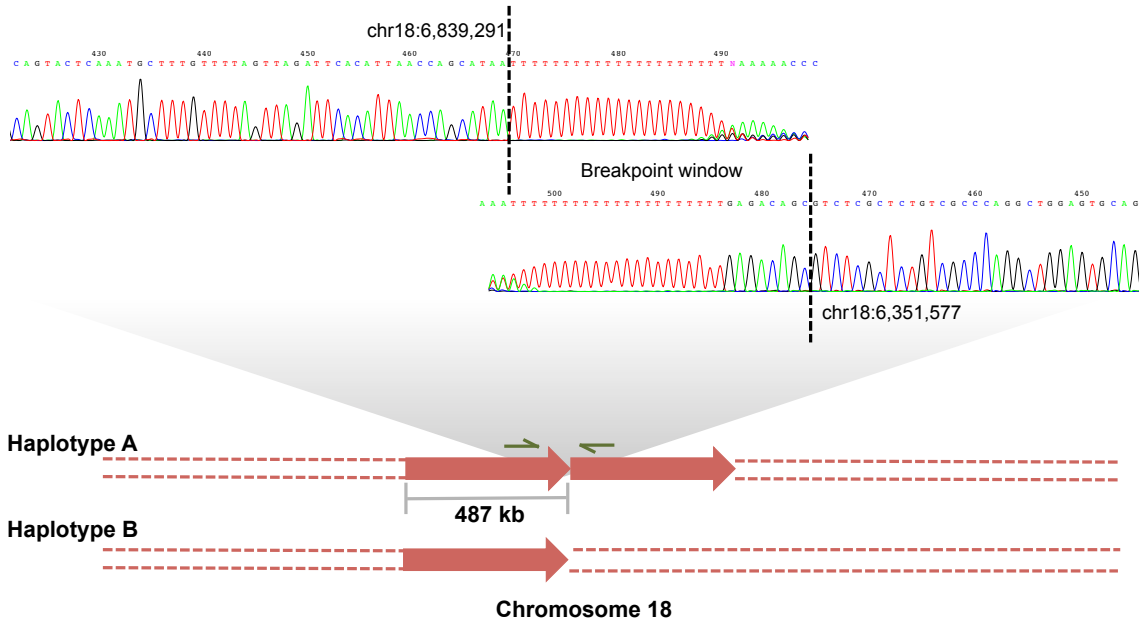


Figure 3.6: **Structure and validation of maternal duplication in Patient 3**. Structure and validation of maternal duplication in Patient 3. A schematic of 18p11.31 is shown at bottom, with red block arrows representing copies of the 487 kb duplicated region. Green arrows represent placement of primers for PCR and Sanger sequencing of the breakpoint window (meaning the window to which the breakpoint has been narrowed; the precise breakpoint cannot be determined with single nucleotide resolution due to the long polyT stretch) Partial Sanger sequencing traces of the PCR amplicon using the forward (top) and reverse (bottom) primers are shown and aligned at the breakpoint window. The base calls and trace from the reverse primer were reverse complemented for clarity. The duplication is mediated by microhomology in the breakpoint window, where a long polyT stretch challenges Sanger sequencing instruments.

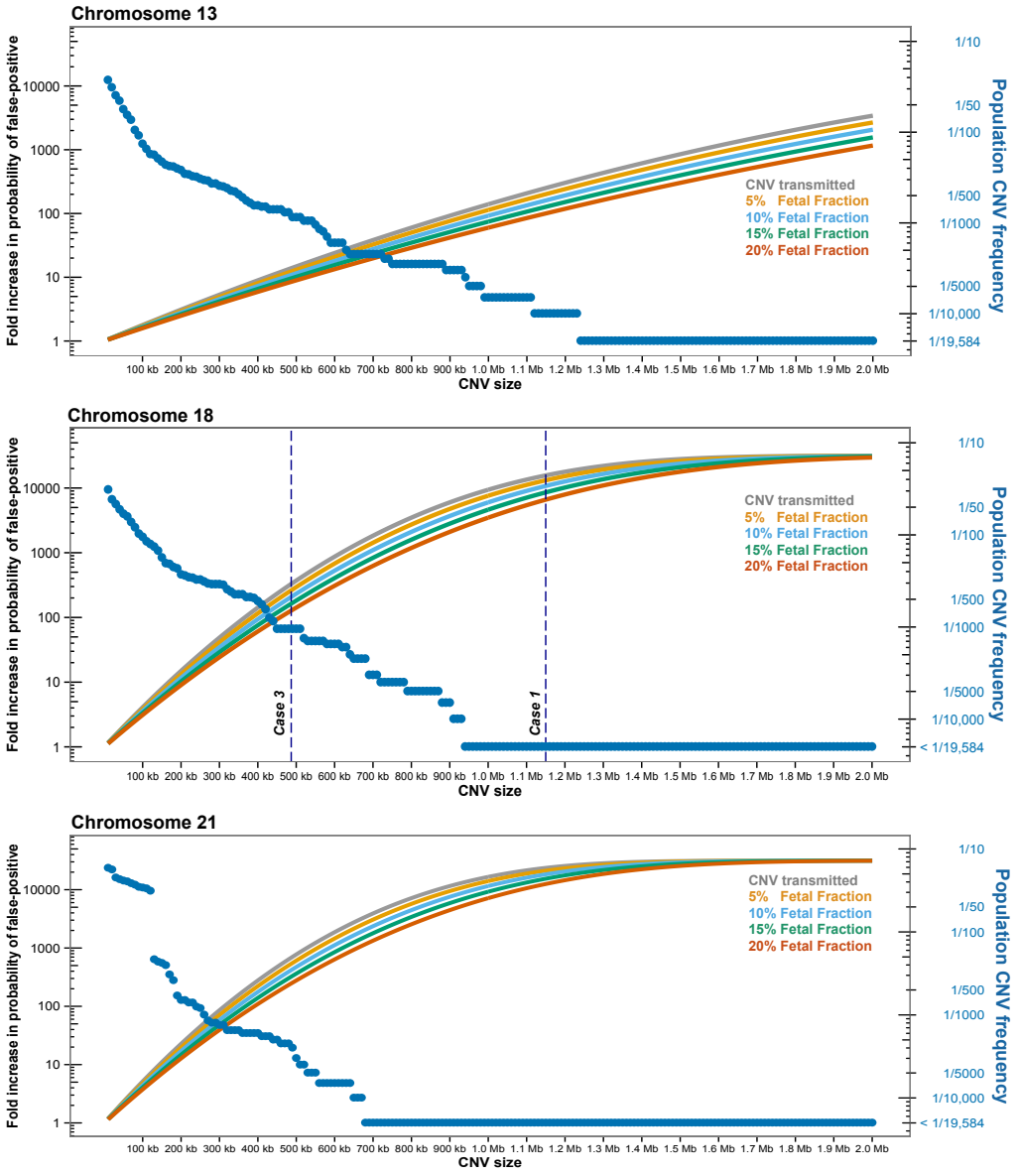


Figure 3.7: **Population frequency and estimated impact on false-positive test rates of maternal CNVs**. Population frequency and estimated impact on false-positive test rates of maternal CNVs. The burden of non-pathogenic copy-number increases on chromosomes 13, 18 and 21 in a cohort of 19,584 individuals predominantly of European ancestry is displayed for a range of CNV sizes *(blue, right vertical axis)*. CNV frequencies in each size bin refer to CNVs of the given size or larger. For each size bin, the estimated fold increase of the probability of a false-positive test resulting from the copy-number increase is shown for a range of fetal fractions (gray and colored lines, left vertical axis). The sizes of the CNVs present in Cases 1 and 3 are highlighted *(dashed vertical lines)*.

of trisomy 13 and trisomy 18 are more likely than trisomy 21. Second, the coefficient of variation of sequence reads for each chromosome modulates the effect of CNV size on the probability of false-positive results. For example, chromosome 13, which has the highest of the three examined coefficients of variation, is the most buffered from the effects of CNVs (Figure 3.7).

3.3 Discussion

Recent advances in cfDNA-based NIPT have yielded screening techniques with substantially better test performance characteristics than previous approaches (Bianchi et al., 2014; Norton et al., 2015). However, PPV remains limited in both high- and low-risk populations, and optimization of these screens, including delineation of potential mechanisms of false-positive results, will be essential as uptake of this form of screening continues to accelerate. We demonstrated and validated large maternal CNVs on chromosome 18 as plausible causes of discordant results in two of four pregnancies with false-positive NIPT results. Furthermore, using CNV frequencies from a largely European cohort, we estimated that maternal CNVs may substantially contribute to an elevated risk of false-positive results.

Our study has several limitations. First, the study samples were not obtained at the same time as the initial samples sent for commercial testing, potentially masking underlying biological changes during gestation. While the presence of maternal CNVs is invariant to sample collection timing, the impact of these CNVs in statistical inference of fetal ploidy does depend on fetal inheritance and fetal fraction, the latter of which increases with gestational age. Thus, later in gestation, marginally larger CNVs are generally required to achieve the same fold increase in false-positive probability when the CNV is not transmitted (Figure 3.7, Table 3.5). Second, we did not directly observe any large CNVs underlying NIPT false-positives on chromosomes 13 or 21. Third, our preliminary estimates of impact from modeling based on population-wide CNV frequencies are only as good as the assumptions and data that went into them, which include the methods themselves (which are not necessarily optimal or static), the coefficients of variation for each chromosome, the set of unique

genomic regions potentially harboring CNVs (Figure 3.8), and the joint distribution of CNV sizes and allele frequencies. For example, the spectra of CNV sizes and frequencies may differ between European and non-European populations, underscoring the importance of future studies with diverse patient groups.

A small cohort such as ours is insufficient to determine the precise impact of maternal CNVs on aggregate cfDNA-based NIPT false-positive rates. Other cfDNA-based NIPT methodologies, such as the targeted analysis of cfDNA from selected genomic regions, may be more or less susceptible to false-positive results owing to maternal CNVs. However, even as larger studies are warranted, implementations of NIPT based on counting statistics arising from shotgun sequencing may be immediately modifiable to reduce maternal CNVs as a source of false-positives. For example, when a maternal CNV is identified (in cfDNA or PBMC-derived DNA, with the latter unconfounded by the fetus), reads derived from the affected region could be discarded or proportionally discounted, or the effective size of the chromosome adjusted. Alternatively, analogous to methods developed by others (Srinivasan et al., 2013), z-scores could be calculated in fixed genome bin sizes, rather than for whole chromosomes, such that region-specific outliers potentially corresponding to maternal CNVs could be flagged or discarded.

Our study has several potential implications for the spectrum of causes of discordant prenatal test results. First, while the incidence of fetal aneuploidy increases with maternal age, the risk of a false-positive NIPT result caused by a maternal CNV would not depend on maternal age, with affected women likely to experience recurrent false-positive results in subsequent pregnancies. Second, while not directly addressed here, the presence of maternal copy number losses or deletions of sequence could potentially induce the opposite effect – that is, false-negative NIPT results in truly aneuploid pregnancies (Figure 3.1). Although the impact of CNV size for hypothetical false-negatives cannot be quantified without coefficients of variation based on truly aneuploid pregnancies, the co-occurrence of trisomic pregnancy and statistically relevant deletions is expected to be very infrequent.

In conclusion, although prospective studies have demonstrated excellent performance

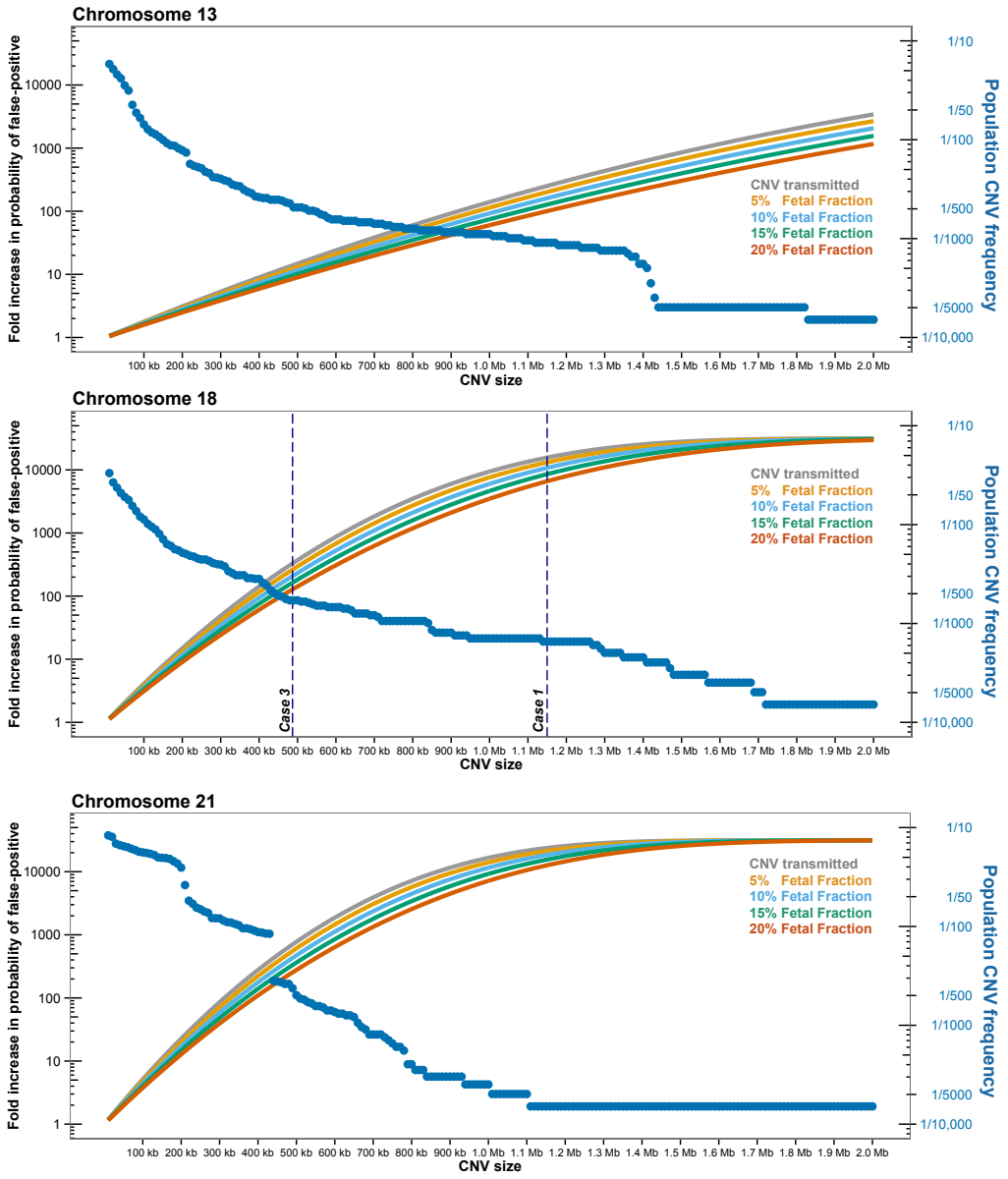


Figure 3.8: **Population frequency and estimated impact on false-positive test rates of maternal CNVs with less stringent filtering**. The burden of non-pathogenic CNVs on chromosomes 13, 18, and 21 in a cohort of 19,584 individuals predominantly of European ancestry is displayed for a range of CNV sizes *(blue, right vertical axis)*, as in Figure 3.7. Here, unlike in Figure 3.7, reference panel CNVs were not filtered for overlap with unique genomic regions. This less conservative approach accounts for the uncertainty of reference panel CNV breakpoints and unspecified coordinates of unique genomic regions used by NIPT providers. CNV frequencies in each size bin refer to CNVs of the given size or larger. For each size bin, the estimated fold increase of the probability of a false-positive test is shown for a range of fetal fractions (gray and colored lines, left vertical axis). The sizes of the CNVs carried by Patient 1 and Patient 3 are highlighted *(top, dashed vertical lines)*.

for cfDNA-based NIPT, the PPV remains limited and follow-up diagnostic testing remains essential. The impact of false-positive screening results goes beyond the clinical risks and financial costs of diagnostic testing and includes potential significant psychological stress imposed on patients. Our modeling based on population-wide CNV frequencies provides initial estimates upon which larger, more definitive studies can be based. Though currently focused clinically on high-risk populations, cfDNA-based NIPT will likely become increasingly broadly utilized as a primary screening test over time. Throughout this transition, continued investigation and refinement of methodology to improve NIPT performance will be critical.

3.4 Methods

3.4.1 Patients and sample processing

Subjects were identified from a population of consecutive false-positive cases referred for perinatal genetic counseling at UW. After delivery, normal clinical outcomes were confirmed; all subjects had fetal diploidy based on antenatal genetic amniocentesis and/or normal newborn exams.

From each pregnancy, maternal peripheral blood samples were drawn at the time of enrollment, and cord blood samples were collected at delivery. Plasma was purified, and cfDNA was isolated, sequenced, and aligned to the reference genome with standard methods. Maternal peripheral blood mononuclear cells (PBMC) were concurrently collected, and DNA was isolated from these cells for validation of CNVs.

3.4.2 Maternal plasma library preparation and sequencing

Maternal plasma was collected by standard methods and stored in 1 mL aliquots at -80°C until use. Circulating cfDNA was purified from plasma with the QIAamp Circulating Nucleic Acid kit (Qiagen). DNA yield was measured with a Qubit fluorometer (Invitrogen). Sequencing libraries were prepared with the ThruPLEX-FD kit (Rubicon Genomics). Library amplification was monitored by real-time PCR to avoid over-amplification. All libraries were sequenced on HiSeq 2000 instruments (Illumina) using paired-end 101 bp reads with an index read of 9 bp.

3.4.3 Read mapping

Reads were mapped to the 1000 Genomes human reference genome sequence including decoy sequences (hs37d5) with BWA v0.7.3a (Li and Durbin, 2009). PCR duplicate read pairs were removed using the Picard toolkit (`http://picard.sourceforge.net/`).

3.4.4 Identification of CNVs

Read depth from shotgun sequencing of maternal cfDNA was calculated in non-overlapping genomic windows of varying sizes, each containing 10,000 singly unique k-mers (SUNKs), as described previously (Sudmant et al., 2010). Read depth profiles include only those reads unambiguously derived from each chromosome. GC-correction was applied separately to each window. Read depth profiles were examined for overrepresentation of portions of chromosomes 13, 18, or 21. Candidate CNVs >250 kb in size were first identified by visual inspection of read-depth profiles on relevant chromosomes and validated by PCR and Sanger sequencing of maternal PBMC DNA and cord blood DNA.

3.4.5 Modeling

For a range of maternal CNV sizes, we calculated the fold-increase in the probability of a false-positive statistical test for each chromosome, based on the properties of the Z distribution underlying the test. The mean numbers of additional reads expected to be derived from the duplicated regions were converted to chromosome-specific standard deviation units, which were then used to calculate adjusted probabilities of false-positive results. Next, we estimated the population frequencies of non-pathogenic CNVs on chromosomes 13, 18, and 21. For this, a CNV reference panel of 19,584 individuals predominantly of European descent (Table 3.2) was filtered for duplications on relevant chromosomes and with at least 50% overlap with unique genomic regions.

The Illumina Verifi test (from which NIPT results for all four subjects were derived) reportedly uses an average of 26.2 million reads per sample (Chiu et al., 2008) and retains only reads uniquely assignable to a single genomic origin. Per-chromosome coefficients of variation (Rava et al., 2013) were used to estimate the standard deviation of the number of reads mapping to each chromosome i using the formula

$$CV_i = \frac{sd_i}{\mu_i}$$

where μ_i was estimated from the average total number of reads (26.2M) for each sample

and the total length of uniquely mappable sequence on each chromosome (Table 3.4). Given sd_i, the minimum number of additional reads required to reach a k unit shift in z-score (i.e., k standard deviations above the mean) is $k \times sd_i$. For a given copy-number gain (i.e., duplication of sequence) of size s on autosomal chromosome i, $C_{s,i}$ represents the proportion of unique sequence on i duplicated by the CNV. The number of additional reads expected to result from this duplication is

$$\mu_i \times \frac{C_{s,i}}{2} \times [1.0 - (FF \times I)]$$

where FF represents the fetal fraction of the sample and I is an indicator variable equal to 0 if the CNV is inherited by the fetus and 1 if the CNV is not inherited. The CNV size required to reach an expected z-score of k for a given autosome can be calculated by setting this expression equal to $k \times sd_i$ and solving this equation for $C_{s,i}$. Table 3.3 provides details of this calculation for representative fetal fractions. Again given sd_i, the minimum deficit of reads required to reach a $-k$ unit shift in z-score (i.e., k standard deviations below the mean) is $k \times sd_i$. For a given copy-number loss (i.e., deletion of sequence) of size s on the X chromosome, $C_{s,X}$ represents the proportion of unique sequence on the X chromosome deleted by the CNV. The deficit of reads expected to result from this deletion is

$$\mu_X \times \frac{C_{s,X}}{2} \times [1.0 - (FF \times I)]$$

if the fetal karyotype is 46,XX, where FF represents the fetal fraction of the sample and I is an indicator variable equal to 0 if the CNV is inherited by the fetus and 1 if the CNV is not inherited. The CNV size, in this case a deletion of sequence, required to reach an expected z-score of $-k$ for the X chromosome can be calculated by setting this expression equal to $k \times sd_i$ and solving this equation for $C_{s,X}$. If the fetal karyotype is 46,XY, the expectation of the number of reads derived from the X chromosome may be adjusted downward by an additional factor of $[1-(0.5 \times FF)]$ to account for the different sex chromosome compositions of the maternal and fetal compartments of the cfDNA. This calculation assumes that the presence of absence of reads derived from the Y chromosome is not relevant to the z-score calculation or inference of fetal karyotype. Table 3.4 provides details of this calculation.

The fold increase in the probability of false-positive results was determined for each CNV size by first finding the value of k associated with a CNV of that size on a particular chromosome (as detailed above), and then calculating

$$\frac{Pr\left(Z > (4.0 - k)\right)}{Pr\left(Z > 4.0\right)}$$

for $0 < k < 4$.

Acknowledgements

Supported by grants from the National Institutes of Health (K08HD067221, to Dr. Gammill; DP1HG007811, to Dr. Shendure; and 1R01MH101221, to Dr. Eichler), and from the Washington State Obstetrical Association (to Dr. Simmons). Dr. Eichler is an Investigator of the Howard Hughes Medical Institute.

Table 3.1: **Calculation of standard deviations in number of reads derived from chromosomes 13, 18, and 21**. From a fixed number of uniquely mapped reads, the mean number of reads derived from each chromosome is calculated based on the amount of unique sequence on that chromosome. In combination with this mean, the coefficient of variation (CV) for each chromosome (Rava et al., 2013) determines the standard deviation in truly diploid pregnancies.

Uniquely mapped reads per sample	Chrom.	CV	Uniquely mappable sequence (bp)	Proportion of all mappable sequence	Mean # of reads from 26.2M	1 SD in diploid samples
26,200,000	13	0.0045	80,382,000	0.03543	928,190	4176.86
	18	0.0023	62,670,000	0.02762	723,644	1664.38
	21	0.0044	27,966,000	0.01233	322,929	1420.89

Table 3.2: **Control CNV cohorts and call sources**. Sources, size, and description of each cohort used for population modeling of CNV frequencies.

Cohort	Array platform	# of samples	Description	Raw data source	CNV call source
HGDP	Human-Hap650Yv3_A	983	The HGDP consists of 1064 individuals sampled from 51 different world populations. N=983 after sample quality control.	PMID:18292342	dbVar: nstd54
NINDS (Coriell 550K)	Human-Hap550v3_A	441	Genotype data from NINDS were derived from two sets of neurological disease controls totaling 790 people and consist of individuals of European descent with no family history of or any first-degree relative with amyotrophic lateral sclerosis, ataxia, autism, brain aneurysm, dystonia, Parkinson disease, or schizophrenia.	dbGaP Accession: phs000089	dbVar: nstd54
NINDS (317K+240K)	Illumina 317K+240K	227	Genotype data from NINDS were derived from two sets of neurological disease controls totaling 790 people and consist of individuals of European descent with no family history of or any first-degree relative with amyotrophic lateral sclerosis, ataxia, autism, brain aneurysm, dystonia, Parkinson disease, or schizophrenia.	dbGaP Accession: phs000089	dbVar: nstd54

Continued on Next Page...

Table 3.2 – Continued

Cohort	Array platform	# of samples	Description	Raw data source	CNV call source
PARC (CAP and PRINCE)	Illumina 317K	936	The PARC samples are a subset of the cohorts used in two statin trials, CAP and PRINCE and consist of 960 middle-age (40-70 years) individuals of European descent living in the United States with moderately high levels of total cholesterol.	PMIDs: 11434828, 16516587	dbVar: nstd54
London (Parents)	Illumina 550K	760	The London samples represent parents of asthmatic children from Mexico City.	PMID: 19714205	dbVar: nstd54
PARC2 (CAP2)	Human610-Quadv1_B	232	The PARC samples are a subset of the cohorts used in two statin trials, CAP2 and PRINCE2, and consist of middle-age (40-70 years) individuals of European descent living in the United States with moderately high levels of total cholesterol	PMIDs: 11434828, 16516587	dbVar: nstd54
PARC2 (PRINCE2)	Illumina 610K Quad	534	The PARC samples are a subset of the cohorts used in two statin trials, CAP2 and PRINCE2, and consist of middle-age (40-70 years) individuals of European descent living in the United States with moderately high levels of total cholesterol.	PMIDs: 11434828, 16516587	dbVar: nstd54

Continued on Next Page...

Table 3.2 – Continued

Cohort	Array platform	# of samples	Description	Raw data source	CNV call source
FHCRC	Human610-Quadv1_B	1430	The FHCRC set are part of an ongoing Genome-wide Association Study to Identify Genetic Components of Hip Fracture in the Women's Health Initiative. Samples represent post-menopausal (50-79 years) female controls for pancreatic cancer, colon cancer, and cases and controls for a hip fracture study.	FHCRC	dbVar: nstd54
inChianti	Human-Hap550v3_a	695	Population-based study of older persons living in the Chianti geographic area.	http://www.inchiantistudy.net/	dbVar: nstd54
WTCCC2(NBS)	Custom Illumina 1.2M	2090	UK Blood Service Control Group (blood donors, age range 18-69 years). Custom Illumina 1.2M Data.	http://www.wtccc.org.uk/	dbVar: nstd54

Continued on Next Page...

Table 3.2 – Continued

Cohort	Array platform	# of samples	Description	Raw data source	CNV call source
ARIC	SNP6	8733	The Atherosclerosis Risk in Communities (ARIC) Cohort Component samples are from a prospective epidemiologic study conducted in four U.S. communities, designed to investigate the etiology and natural history of atherosclerosis, the etiology of clinical atherosclerotic diseases, and variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date.	dbGaP Accession: phs000090	Affymetrix GTC 4.1 + Filtering
WTCCC2(58C)	SNP6	2523	1958 British Birth Cohort	http://www.wtccc.org.uk/	Affymetrix GTC 4.1 + Filtering

Table 3.3: **Demographic and pregnancy characteristics of study subjects**. Gravidity and Parity status at time of study enrollment. Note: though classification of neonatal outcomes according to newborn examinations has been traditionally accepted Bianchi et al. (2014) and was considered a normal outcome in our study, a newborn examination does have limitations. Specifically, a normal newborn examination rules out a diagnosis of Trisomy 18 but does not provide an evaluation for the presence of a small supernumerary marker chromosome, which could be present in the setting of a normal phenotype (Eckmann-Scholz et al., 2012; Marle et al., 2014)

Patient	Maternal age	Ethnicity	Gravidity and parity	BMI	Indication for cfDNA testing	Gestational age at cfDNA testing and result	Diagnostic testing and clinical outcome
1	36	Cau-casian	G1P0	28	Maternal age 35 years or older at delivery	18 weeks: Trisomy 18	20 week ultrasound: normal fetal anatomy. Amniocentesis at 20 weeks: 46,XY. Term delivery, normal newborn exam.
2	25	Cau-casian	G2P1	65	Fetal ultrasonographic findings indicating an increased risk of aneuploidy	25 weeks: Trisomy 18	20 week ultrasound: Echogenic intracardiac focus, otherwise normal fetal anatomy. Declined amniocentesis. 31 and 36 week ultrasounds with normal interval growth. Term delivery, normal newborn exam. No infant karyotype performed.
3	34	Cau-casian	G2P0	25	Maternal age 35 years or older at delivery	12 weeks: Trisomy 18	12 week, 16 week and 20 week ultrasounds with normal fetal anatomy. Declined amniocentesis. Term delivery, normal newborn exam. No infant karyotype performed.

Continued on Next Page...

Table 3.3 – Continued

Patient	Maternal age	Ethnicity	Gravidity and parity	BMI	Indication for cfDNA testing	Gestational age at cfDNA testing and result	Diagnostic testing and clinical outcome
4	38	Cau-casian	G4P0	24	Maternal age 35 years or older at delivery	12 weeks: Trisomy 13	13 week, 16 week and 20 week ultrasounds with normal fetal anatomy. Amniocentesis at 16 weeks: 46,XX. 29 week ultrasound with normal growth. Term delivery, normal newborn exam.

Table 3.4: **Breakpoints and PCR primers for detection of maternal CNVs**. Breakpoints are in hg19 coordinates. The chromosome 18 amplification positive control, used in Figure 3.4, is expected to produce a 229 bp amplicon.

Patient	Breakpoints	Band	CNV size	PCR Primers
1	chr18: 6,935,598 - 8,087,852	18p11.31 – 18p11.23	1.15 Mb	5'-TGACCACTTTCAGCATGCCA-3' 5'-GCTTGGAAGAAGACTCAGTGGA-3'
3	chr18: 6,351,540 - 6,839,310	18p11.31	487 kb	5'-AGGGACTTTCTACTTGAGAAGCA-3' 5'-CCTTCTTGGCAGGGGGAAAT-3'
Amplification positive control				5'-TCGAAGTGTGCTTTCCCTGA-3' 5'-ACATTTTCCAGAGGCCGACA-3'

Table 3.5: **The role of CNV size and fetal fraction in false-positive NIPT results**. Z score increases refer to the number of normalized units the sampling distribution is shifted relative to the underlying reference distribution by a maternal CNV. The CNV size depends on the transmission of the CNV and, for CNVs that are not transmitted, on the fetal fraction of cfDNA in the maternal plasma. The population burden of such CNVs, as estimated from a cohort of 19,584 controls of predominantly European ancestry (Table 3.2), is given for each fetal fraction. All calculations assume that the pregnancy is diploid, and are based on 26.2 million reads per sample.
Estimated false-positive rates are reported by Bianchi and colleagues (Bianchi et al., 2014). The rate for chromosome 13 includes the 899 patients for whom standard screening was available as well as the 1,015 patients for whom only cfDNA-based results were available.
FP: false-positive NIPT result for the given chromosome.

Chrom	False-positive rate	CNV transmitted			CNV not transmitted		
		Z score increase	CNV size	Population prevalence	Fetal Fraction	CNV size	Population prevalence
13	3 / 1,914 (0.16%)	1.0 (43-fold)	723 kb	8 / 19,584 (0.041%)	5%	762 kb	7 / 19,584 (0.036%)
					10%	804 kb	7 / 19,584 (0.036%)
					15%	851 kb	7 / 19,584 (0.036%)
					20%	904 kb	6 / 19,584 (0.031%)
		2.0 (718-fold)	1.45 Mb	1 / 19,584 (0.005%)	5%	1.52 Mb	1 / 19,584 (0.005%)
					10%	1.61 Mb	1 / 19,584 (0.005%)
					15%	1.70 Mb	1 / 19,584 (0.005%)
					20%	1.81 Mb	1 / 19,584 (0.005%)

Continued on Next Page...

Table 3.5 – Continued

Chrom	False-positive rate	CNV transmitted			CNV not transmitted		
		Z score increase	CNV size	Population prevalence	Fetal Fraction	CNV size	Population prevalence
18	3 / 1,905 (0.16%)	1.0 (43-fold)	288 kb	58 / 19,584 (0.296%)	5%	303 kb	58 / 19,584 (0.296%)
					10%	320 kb	51 / 19,584 (0.260%)
					15%	339 kb	45 / 19,584 (0.230%)
					20%	360 kb	45 / 19,584 (0.230%)
		2.0 (718-fold)	577 kb	14 / 19,584 (0.071%)	5%	607 kb	13 / 19,584 (0.066%)
					10%	641 kb	10 / 19,584 (0.051%)
					15%	678 kb	9 / 19,584 (0.046%)
					20%	721 kb	5 / 19,584 (0.026%)

Continued on Next Page…

Table 3.5 – Continued

Chrom	False-positive rate	CNV transmitted			CNV not transmitted		
		Z score increase	CNV size	Population prevalence	Fetal Fraction	CNV size	Population prevalence
21	6 / 1,909 (0.31%)	1.0 (43-fold)	246 kb	24 / 19,584 (0.123%)	5%	259 kb	22 / 19,584 (0.112%)
					10%	273 kb	17 / 19,584 (0.087%)
					15%	290 kb	16 / 19,584 (0.082%)
					20%	308 kb	15 / 19,584 (0.077%)
		2.0 (718-fold)	492 kb	8 / 19,584 (0.041%)	5%	518 kb	5 / 19,584 (0.026%)
					10%	547 kb	4 / 19,584 (0.020%)
					15%	579 kb	3 / 19,584 (0.015%)
					20%	615 kb	3 / 19,584 (0.015%)

Chapter 4

CELL-FREE DNA COMPRISES AN *IN VIVO* NUCLEOSOME FOOTPRINT THAT INFORMS ITS TISSUES-OF-ORIGIN

This chapter has been adapted with minor changes from: Snyder, MW; Kircher M; Hill, AJ; Daza, RM; and Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues of origin. *Cell* 164:57-68 (2016).

In the published manuscript, I share first author credit with Dr. Kircher. My specific contributions include the conceptualization of the project, processing of biological samples, preparation and sequencing of cfDNA libraries, data analysis, preparation of many of the figures and tables, and significant effort towards the first draft of the manuscript.

Abstract

Nucleosomes are the basic unit of chromatin packaging, and nucleosome positioning varies between cell types. We deeply sequenced plasma-borne cell-free DNA (cfDNA) to generate a genome-wide map of *in vivo* nucleosome occupancy, while also finding that short cfDNA fragments footprint transcription factors. The nucleosome map, which contains 13M positions and spans 2.5 gigabases of the human genome, correlates well with nuclear architecture, gene structure and expression. Nucleosome spacing inferred from cfDNA in healthy individuals correlates most strongly with epigenetic features of lymphoid and myeloid cells, consistent with hematopoietic cell death as the normal source of cfDNA. We build on this observation to show how nucleosome footprints can be used to infer cell types contributing to cfDNA in pathological states such as cancer. Because it does not rely on genotypic differences, this strategy may enable the noninvasive cfDNA-based monitoring of a much broader set of clinical conditions than is currently possible.

4.1 Introduction

Cell-free DNA (cfDNA) is present in the circulating plasma, urine, and other bodily fluids of humans (Chan et al., 2003). The cfDNA comprises double-stranded DNA fragments that are overwhelmingly short (<200 base-pairs (bp)) and normally at a low concentration (Fleischhacker and Schmidt, 2007). In healthy individuals, plasma cfDNA is believed to derive primarily from apoptosis of normal cells of the hematopoietic lineage, with minimal contributions from other tissues (Lui et al., 2002). The short half-life of cfDNA in the circulation (Lo et al., 1999) suggests a model of ongoing release from apoptotic cells and rapid degradation or filtration. The size distribution of cfDNA fragments bears correspondence with these origins – specifically, peaks corresponding to nucleosomes (~147 bp) and chromatosomes (nucleosome + linker histone; ~167 bp) have been noted (Fan et al., 2008; Lo et al., 2010) – and some proportion of cfDNA may circulate as nucleosomes or chromatosomes, rather than as free DNA (Holdenrieder et al., 2005; Wimberger et al., 2010)).

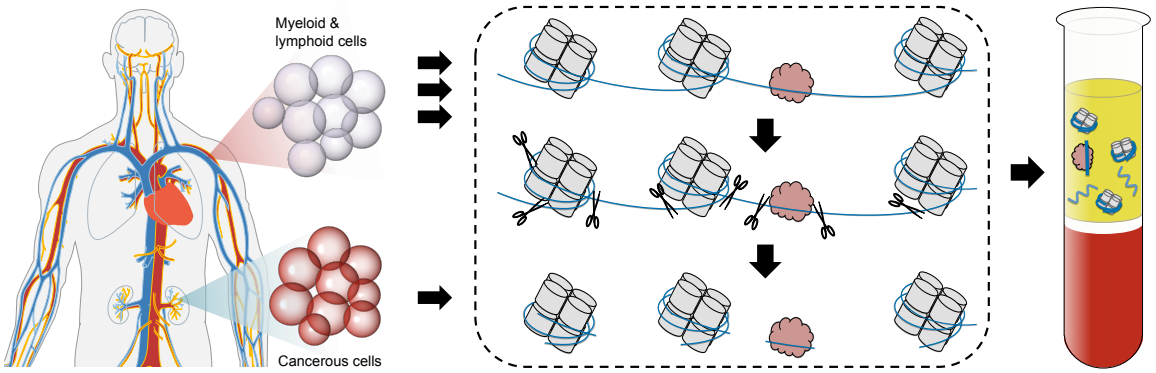


Figure 4.1: **Schematic overview of cfDNA fragmentation**. Schematic overview of cfDNA fragmentation. Apoptotic or necrotic cell death results in near-complete digestion of native chromatin. Protein-bound DNA fragments, typically associated with histones or TFs, preferentially survive digestion and are released into the circulation, while naked DNA is lost. Fragments can be recovered from peripheral blood plasma following proteinase treatment. In healthy individuals, cfDNA is primarily derived from myeloid and lymphoid cell lineages, but contributions from one or more additional tissues may be present in certain medical conditions.

In the context of specific physiological conditions or disease processes, a substantial proportion of cfDNA may be derived from a different distribution of tissues than during the typical, healthy state. This fact has been exploited in recent years to achieve non-invasive diagnostics based on cfDNA composition. In pregnant women, ~10-15% of cfDNA originates from placental trophoblasts, and cfDNA-based screening for fetal genetic abnormalities is now common in high-risk pregnancies (Chiu et al., 2008; Fan et al., 2008). In oncology, the monitoring of advanced cancers by quantifying mutations or aneuploidy in tumor-shed cfDNA is gaining traction (Diaz and Bardelli, 2014). In transplant medicine, allograft rejection events can be correlated with abnormally high levels of donor-derived cfDNA fragments contributed by the transplanted solid organ (Snyder et al., 2011).

Despite these advances, a common limitation is the requirement for genetic differences to distinguish between contributing tissues, e.g. fetus vs. mother, tumor vs. normal, or donor vs. recipient. Conditions such as myocardial infarction (Chang et al., 2003), stroke (Rainer et al., 2003) and autoimmune disorders (Galeazzi et al., 2003) are associated with elevations in cfDNA levels, possibly consequent to tissue damage, but cannot be specifically monitored via cfDNA because of the lack of such genetic differences. Furthermore, even as mutations enable monitoring of tumor-derived cfDNA, they only weakly inform a tumor's tissue-of-origin.

We hypothesized that if cfDNA is the detritus of cell death, and if the boundaries of cfDNA fragments are biased by their association with nucleosomes, then the fragmentation patterns observed in an individual's cfDNA might contain evidence of the epigenetic landscape(s) of the cells giving rise to these fragments – and thus, of their tissue(s)-of-origin – i.e., a strategy that does not rely on genotypic differences between contributing cell types.

To evaluate this hypothesis, we first set out to deeply sequence cfDNA to better understand the processes that give rise to it. We use the resulting data to build a map of nucleosome occupancy that approaches saturation of the mappable human genome. By optimizing protocols to recover short fragments, we discover that the *in vivo* occupancies of transcription factors (TFs) such as CTCF are also directly footprinted by cfDNA. Finally,

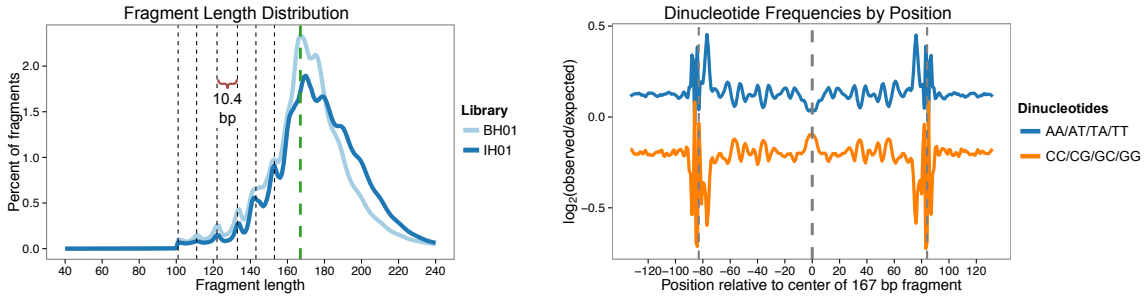


Figure 4.2: **Characteristics of conventional cell-free DNA sequencing libraries**. Characteristics of conventional cell-free DNA sequencing libraries. *(left)* Fragment length of cfDNA observed with conventional sequencing library preparation, inferred from alignment of paired-end reads. A reproducible peak in fragment length at 167 bp (green dashed line) is consistent with association with chromatosomes. Additional peaks evidence ~10.4 bp periodicity, corresponding to the helical pitch of DNA on the nucleosome core. Enzymatic end-repair during library preparation removes 5′ and 3′ overhangs and may obscure true cleavage sites. *(right)* Dinucleotide composition of 167 bp fragments and flanking genomic sequence in conventional libraries. Observed dinucleotide frequencies in the BH01 library were compared to expected frequencies from simulated fragments.

we show that nucleosome spacing in regulatory elements and gene bodies, as revealed by cfDNA sequencing in healthy individuals, correlates most strongly with DNase I hypersensitivity (DHS) and gene expression in lymphoid and myeloid cell lines. To test whether we can infer additional contributing tissues in non-healthy states, we sequenced cfDNA samples from five late-stage cancer patients. The patterns of nucleosome spacing in these samples reveal additional contributions to cfDNA that correlate most strongly with non-hematopoietic tissues or cell lines, often matching the anatomical origin of the patient's cancer.

4.2 Results

4.2.1 cfDNA fragments correspond to chromatosomes and contain substantial DNA damage

We prepared conventional sequencing libraries by end-repair and adaptor ligation to cfDNA fragments purified from plasma pooled from an unknown number of healthy individuals

('BH01') or a single individual ('IH01') (Figure 4.1; Table 4.1). We sequenced these libraries to 96- and 105-fold coverage (1.5 billion (G) and 1.6G fragments). The fragment length distributions have a dominant peak at ~167 bp (coincident with the length of DNA associated with a chromatosome), and ~10.4 bp periodicity in the 100-160 bp range (Figure 4.2). These distributions support a model in which cfDNA fragments are preferentially protected from nuclease cleavage by association with proteins – in this case, by the nucleosome core particle (NCP) and linker histone – but where some degree of additional nicking or cleavage occurs in relation to the helical pitch of nucleosome-bound DNA (Fan et al., 2008; Lo et al., 2010). Further supporting this model is the dinucleotide composition of these fragments, which recapitulates key features of earlier studies of MNase-derived, nucleosome-associated fragments (e.g. bias against A/T dinucleotides at the dyad) (Gaffney et al., 2012) and supports the notion that the NCP is symmetrically positioned with respect to the chromatosome (Harshman et al., 2013) (Figure 4.2).

A prediction of this model is widespread DNA damage, e.g. single-strand nicks as well as 5' and 3' overhangs. During conventional library preparation, damaged as well as short dsDNA molecules (Mouliere et al., 2014) may be poorly recovered. To address this, we prepared a single-stranded cfDNA library from an additional healthy individual ('IH02') using a protocol adapted from studies of ancient DNA (Figure 4.3; Table 4.2) (Gansauge and Meyer, 2013), and sequenced it to 30-fold coverage (779M fragments).

The fragment length distribution again exhibited a dominant peak at ~167 bp, but was considerably enriched for shorter fragments relative to conventional library preparation (Figure 4.4). Although all libraries exhibit ~10.4 bp periodicity, the fragment sizes are offset by ~3 bp for the two methods, consistent with damaged or non-flush input molecules whose true endpoints are more faithfully represented in single-stranded libraries.

4.2.2 A genome-wide map of *in vivo* nucleosome protection based on deep cfDNA sequencing

We next asked whether the predominant local positions of nucleosomes in tissue(s) contributing to cfDNA could be inferred from the distribution of aligned fragment endpoints.

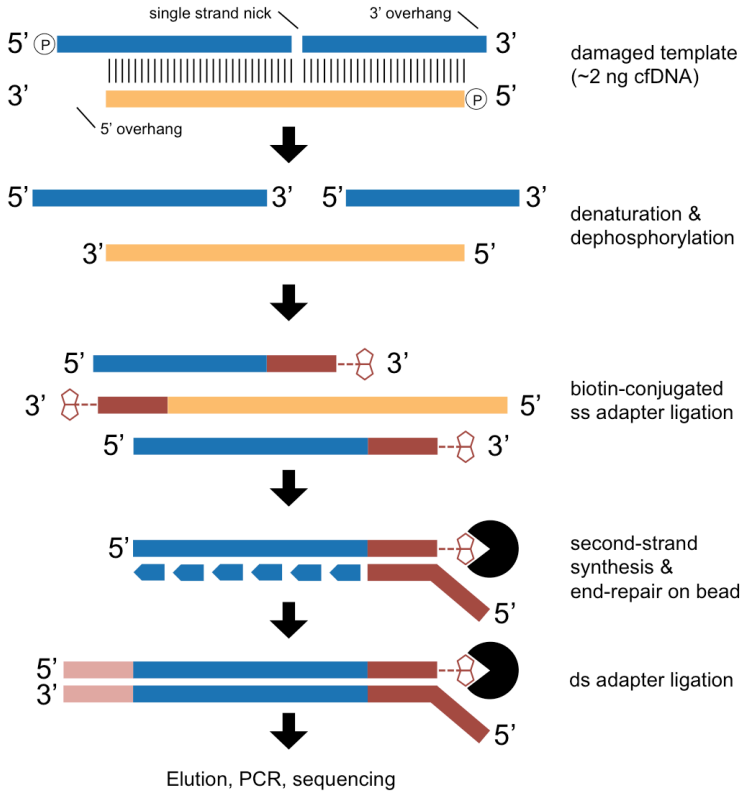


Figure 4.3: **Schematic of single-stranded library preparation protocol**. Input molecules, consisting of approximately 1-10 ng of cfDNA fragments, may have single stranded nicks and/or 5′ or 3′ overhangs *(top, yellow and blue bars)*. Fragments are dephosphorylated and denatured, separating single strands of DNA at nick sites *(second panel)*. After ligation of a 3′ biotin-conjugated single-stranded adapter to each fragment *(third panel, red bars)*, fragments are tightly bound to streptavidin-coated beads for downstream steps. Second-strand synthesis and end polishing *(fourth panel)* enable the ligation of a second, double-stranded adapter *(fifth panel)*. After elution of the newly synthesized second strand, sample indices and flow-cell adapters are added during PCR.

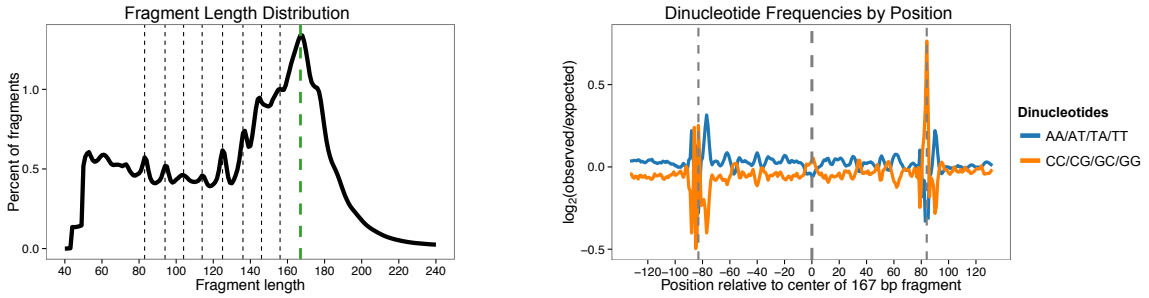


Figure 4.4: **Characteristics of single-stranded cell-free DNA sequencing libraries**. Characteristics of single-stranded cell-free DNA sequencing libraries. *(left)* Fragment length of cfDNA in single-stranded sequencing library preparation. No enzymatic end-repair is performed to template molecules during library preparation. Short fragments of 50-120 bp are highly enriched compared to conventional libraries. While ~10.4 bp periodicity remains, its phase is shifted by ~3 bp. *(right)* Dinucleotide composition of 167 bp fragments and flanking genomic sequence in single-stranded library IH02, calculated as in 4.2. The apparent difference in the background level of bias between BH01 and IH02 relate to differences between the simulations, rather than the real libraries *(data not shown)*.

Specifically, we expect that cfDNA fragment endpoints should cluster adjacent to NCP boundaries, while also being depleted on the NCP itself. To quantify this, we developed a Windowed Protection Score (WPS), which is the number of DNA fragments completely spanning a 120 bp window centered at a given genomic coordinate, minus the number of fragments with an endpoint within that same window (Figure 4.5).

As expected, the WPS correlates with the locations of nucleosomes within strongly positioned arrays, as mapped by other groups with *in vitro* methods (Gaffney et al., 2012; Valouev et al., 2012) or ancient DNA (Pedersen et al., 2014) (Figure 4.6). At other sites, the WPS correlates with genomic features such as DHS sites, e.g. consistent with the repositioning of nucleosomes flanking a distal regulatory element (Figure 4.7).

We applied a heuristic peak-calling algorithm to the genome-wide WPS of the BH01, IH01 and IH02 datasets to identify and score 12.6M, 11.9M, and 9.7M local maxima of nucleosome protection (Figures 4.8 and 4.9). In each sample, the mode distance between adjacent peaks is 185 bp with low variance (Figure 4.9), consistent with previous analyses

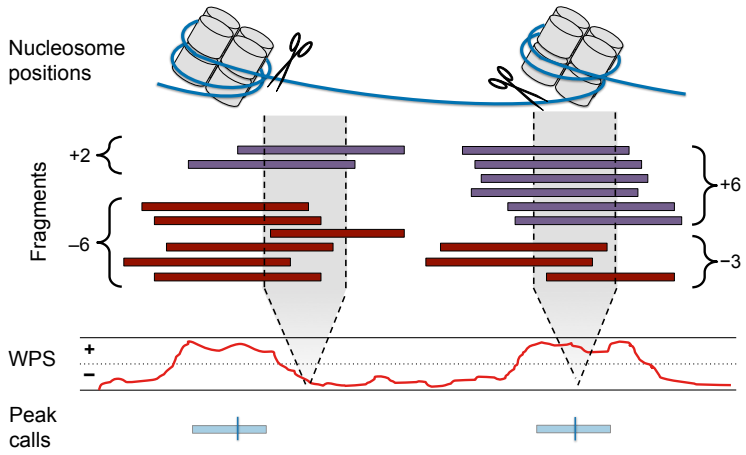


Figure 4.5: **Schematic of inference of nucleosome positioning**. Schematic of inference of nucleosome positioning. A per-base windowed protection score (WPS) is calculated by subtracting the number of fragment endpoints within a 120 bp window from the number of fragments completely spanning the window. High WPS values indicate increased protection of DNA from digestion; low values indicate that DNA is unprotected. Peak calls identify contiguous regions of elevated WPS.

of the nucleosome repeat length in mammalian cells (Teif et al., 2012; Valouev et al., 2012). The positions of peak calls are concordant between samples (Figure 4.10), e.g. the median (absolute) distance from a BH01 peak call to a nearest-neighbor IH01 peak call is 23 bp overall, but <10 bp for the most highly scored peaks (*data not shown*).

As biases introduced by nuclease specificity or library preparation might artifactually contribute to the signal of nucleosome protection, we also simulated fragment endpoints, matching for the depth, size distribution and terminal dinucleotide frequencies. We then calculated genome-wide WPS and called 10.3M, 10.2M, and 8.0M local maxima by the same heuristic, for simulated datasets matched to BH01, IH01 and IH02, respectively. Peaks from simulated datasets are associated with lower scores than peaks from real datasets and do not align well with the locations of peaks called from real datasets (Figure 4.10).

We next pooled and reanalyzed data from BH01, IH01, and IH02 ('CH01'; 231-fold coverage; 3.8G fragments; Table 4.1). The resulting map of *in vivo* nucleosome occupancy

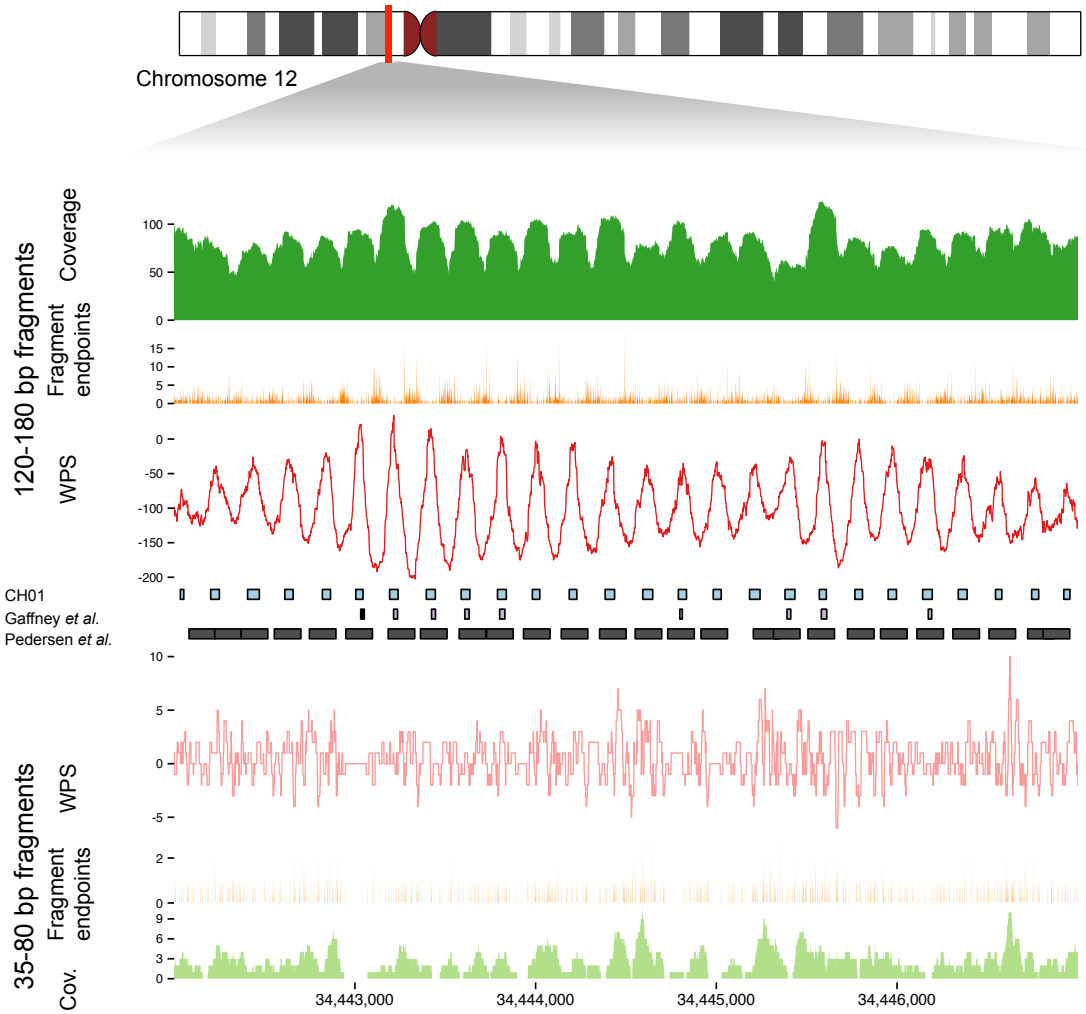


Figure 4.6: **Strongly positioned nucleosomes at a well-studied alpha-satellite array**. Strongly positioned nucleosomes at a well-studied alpha-satellite array. Coverage, fragment endpoints, and WPS values from sample CH01 are shown for long fragment (120 bp window; 120–180 bp fragments) or short fragment (16 bp window; 35–80 bp fragments) bins at a pericentromeric locus on chromosome 12. Nucleosome calls from CH01 *(middle, blue boxes)* are regularly spaced across the locus. Nucleosome calls from two published callsets (Gaffney et al., 2012; Pedersen et al., 2014) *(middle, purple and black boxes)* are also displayed.

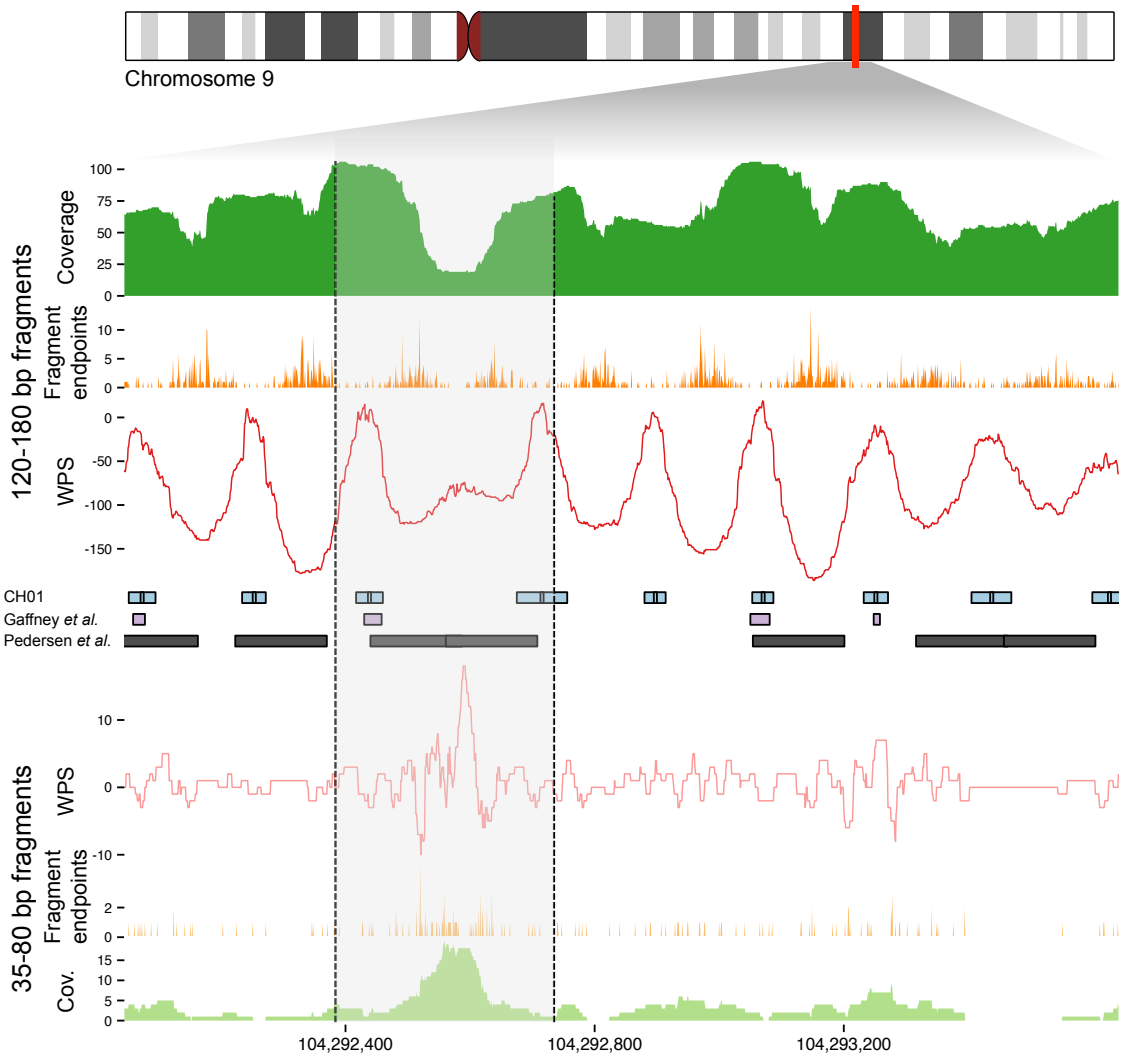


Figure 4.7: **Inferred nucleosome positioning around an example DNase-I Hypersensitive Site**. Inferred nucleosome positioning around an example DNase-I Hypersensitive Site. Coverage, fragment endpoints, WPS values, and nucleosome calls are shown as in Figure 4.6. The hypersensitive region (gray shading), is marked by reduced coverage in the long fragment bin. Nucleosome calls adjacent to the DHS site are spaced more widely than typical adjacent pairs, consistent with accessibility of the intervening sequence to regulatory proteins including TFs. Coverage of short fragments, which may be associated with such proteins, is increased at the DHS site, which overlaps with several annotated TFBSs *(not shown)*.

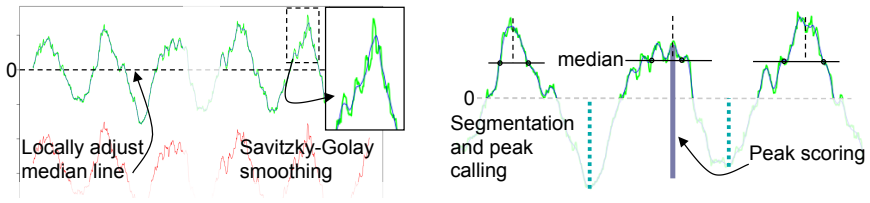


Figure 4.8: **Schematic of peak calling and scoring**. Nucleosomes are called from the long fragment WPS values after local adjustment to a running median of zero (1 kb window) and smoothing with a Savitzky-Golay filter. The WPS track is segmented into above-zero regions, allowing up to 5 consecutive positions below zero. The median WPS value for each region is calculated, and the maximum-sum contiguous window above the median is identified. A single nucleosome call consists of the start, end and center coordinates of this window, and an associated score. The score of the call is defined as the distance between maximum value in the window and the average of the two adjacent WPS minima neighboring the region.

comprises 12.9M peaks, with higher scores and approaching saturation. Considering all peak-to-peak distances below 500 bp (Figure 4.11), the CH01 peaks span 2.53 gigabases.

Nucleosomes are known to be well-positioned in relation to landmarks of gene regulation, e.g. transcriptional start sites (Pedersen et al., 2014) and exon-intron boundaries (Andersson et al., 2009; Chodavarapu et al., 2010). We observe such positioning in our data as well, in relation to landmarks of transcription, translation and splicing (Figure 4.12A-D). We further examined the median peak-to-peak spacing within 100 kilobase (kb) windows that had been assigned to compartment A (enriched for open chromatin) or compartment B (enriched for closed chromatin) on the basis of chromatin contact maps in a lymphoblastoid cell line (Rao et al., 2014). Nucleosomes in compartment A exhibit tighter spacing than nucleosomes in compartment B (median 187 bp (A) vs. 190 bp (B)), with further differences between subcompartments (Figure 4.13). Along the length of chromosomes, no general pattern is seen, except that median nucleosome spacing drops sharply in pericentromeric regions, presumably driven by strong positioning across arrays of alpha satellites (Figures 4.6 and 4.13).

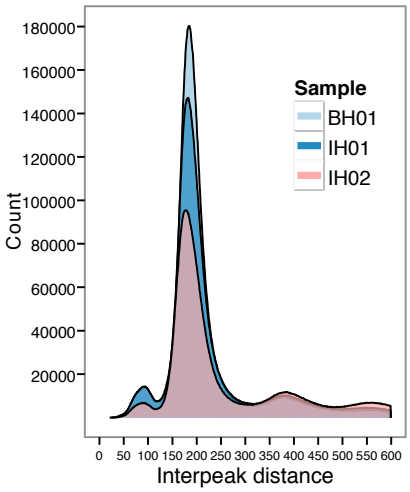


Figure 4.9: **Distances between adjacent peaks, by sample**. Distances between adjacent peaks by sample. Distances are measured between adjacent peak centers.

4.2.3 Short cfDNA fragments directly footprint CTCF and other TFs

Previous studies of DNase I cleavage patterns identified two dominant classes of fragments: longer fragments associated with cleavage between nucleosomes, and shorter fragments associated with cleavage adjacent to transcription factor binding sites (TFBS) (Vierstra et al., 2013). To ask whether *in vivo*-derived cfDNA fragments also result from two classes of sensitivity to nuclease cleavage, we partitioned sequence reads (CH01) on the basis of inferred fragment length and recalculated the WPS using long fragments (120-180 bp; 120 bp window; the same as the WPS described for nucleosome calling) or short fragments (35-80 bp; 16 bp window) separately (Figures 4.6 and 4.7). To obtain a set of well-defined TFBSs enriched for actively bound sites in our data, we intersected clustered FIMO predictions (Grant et al., 2011; Maurano et al., 2012) with a unified set of ChIP-seq peaks from ENCODE for each TF.

Consistent with observations by others (Fu et al., 2008; Pedersen et al., 2014; Teif et al., 2012), the long fraction WPS (L-WPS) supports strong organization of nucleosomes near

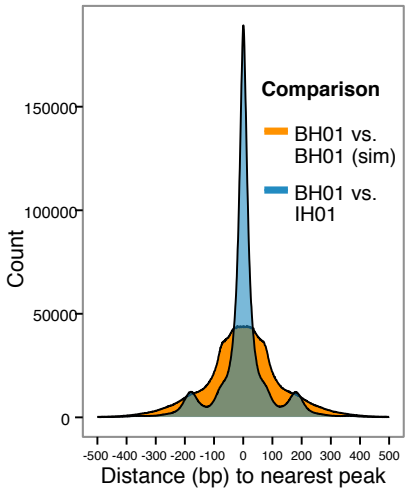


Figure 4.10: **Comparison of peak calls between samples**. Comparison of peak calls between samples. For each pair of samples, the distances between each peak call in the sample with fewer peaks and the nearest peak call in the other sample are shown. Negative and positive numbers indicate the nearest peak is upstream or downstream, respectively.

CTCF binding sites (Figure 4.14A). However, we also observe a strong signal in the short fraction WPS (S-WPS) coincident with the CTCF site itself (Ong and Corces, 2014) (Figures 4.14A and 4.14B). We stratified CTCF sites based on our confidence that they are bound *in vivo*. Experimentally well-supported CTCF sites exhibit substantially broader spacing between the flanking −1 and +1 nucleosomes based on the L-WPS, consistent with their repositioning upon CTCF binding (~190 bp → ~260 bp; Figure 4.14C). Experimentally well-supported CTCF sites also exhibit a much stronger S-WPS signal over the CTCF binding site itself (Figure 4.14D).

We performed similar analyses for additional TFs for which both FIMO predictions and ENCODE ChIP-seq data were available. For many of these, e.g. ETS and MAFK (Figures 4.14E and 4.14F), we observe a short fraction footprint that is accompanied by periodic signal in the L-WPS, consistent with strong positioning of nucleosomes surrounding bound TFBS. Overall, these data support the view that short cfDNA fragments, which are much better recovered by the single-stranded protocol (Figures 4.2 and 4.4), directly footprint

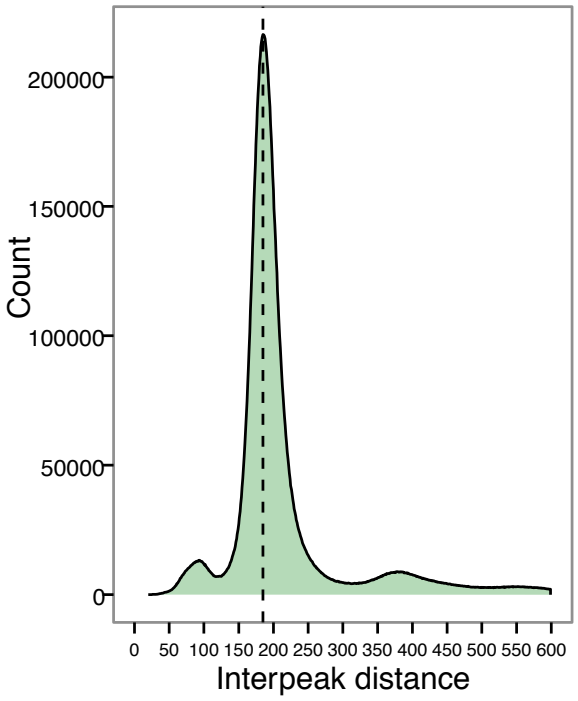


Figure 4.11: **Distances between adjacent peaks, samples CH01**. Distances between adjacent peaks, sample CH01. The dotted black line indicates the mode of the distribution (185 bp).

the *in vivo* occupancy of DNA-bound TFs including CTCF and others.

4.2.4 Nucleosome spacing patterns inform cfDNA tissues-of-origin

We next asked whether *in vivo* nucleosome protection, as measured through cfDNA, could be used to infer the cell types contributing to cfDNA in healthy individuals. We examined the peak-to-peak spacing of nucleosome calls within DHS sites defined in 116 diverse biological samples (Maurano et al., 2012). Similar to bound CTCF sites (Figure 4.14C), we observe substantially broader spacing for nucleosome pairs within a subset of DHS sites, plausibly corresponding to sites at which the nucleosomes are repositioned by intervening TF binding in the cell type(s) giving rise to cfDNA (~190 bp → ~260 bp; Figure 4.15). Indeed, the proportion of widened nucleosome spacing (~260 bp) varies considerably de-

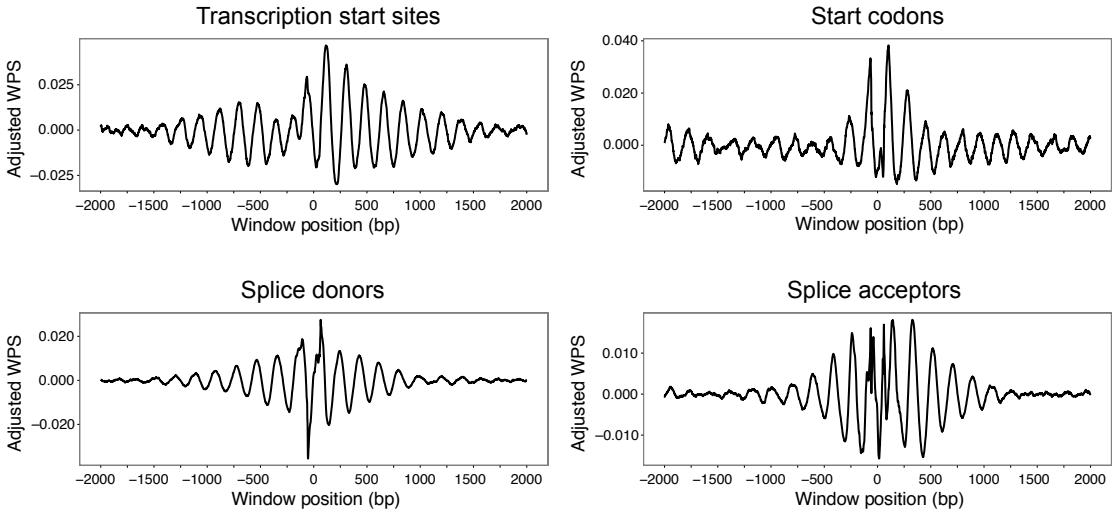


Figure 4.12: **Nucleosome positioning and spacing correlates with genomic features**. Nucleosome positioning and spacing correlates with genomic features. *(top left)* Aggregate, adjusted windowed protection scores (WPS; 120 bp window) around 22,626 transcription start sites (TSS). TSS are aligned at the 0 position after adjusting for strand and direction of transcription. Aggregate WPS is tabulated for both real data and simulated data by summing per-TSS WPS at each position relative to the centered TSS. The values plotted represent the difference between the real and simulated aggregate WPS (see Methods for details). *(top right)* Aggregate, adjusted WPS around 22,626 start codons. *(bottom left and bottom right)* Aggregate, adjusted WPS around 224,910 splice donor and 224,910 splice acceptor sites.

pending on which cell type's DHS sites are used. However, all of the cell types for which this proportion is highest are lymphoid or myeloid in origin (e.g. CD3_CB-DS17706, etc. in Figure 4.15), consistent with hematopoietic cell death as the dominant source of cfDNA in healthy individuals (Lui et al., 2002).

We next re-examined the signal of nucleosome protection in the vicinity of transcriptional start sites (TSS) (Figure 4.12). If we stratify based on gene expression in a lymphoid cell line, NB-4, we observe strong differences in the patterns of nucleosome protection in relation to the TSS, in highly vs. lowly expressed genes (Figure 4.16A). Furthermore, if we examine the S-WPS, we observe a clear footprint immediately upstream of the TSS whose intensity also strongly correlates with expression (Figure 4.16B). This plausibly re-

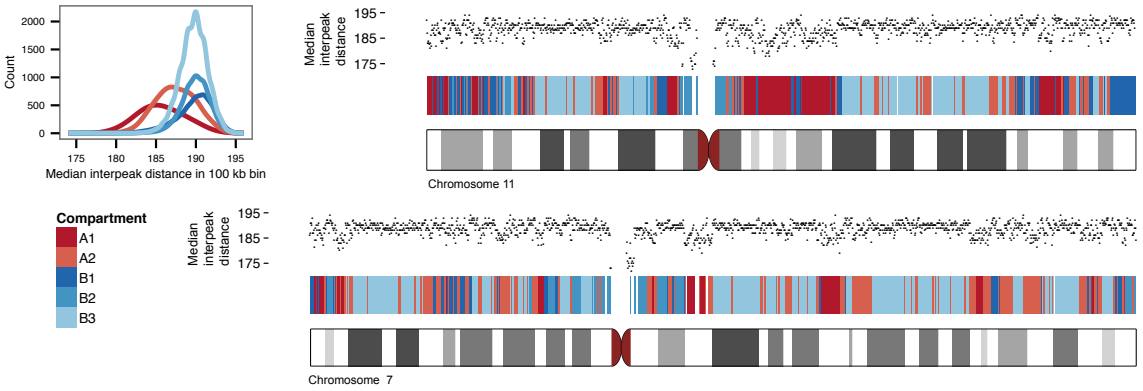


Figure 4.13: **Nucleosome spacing in A/B compartments**. Nucleosome spacing in A/B compartments. *(top left)* Median nucleosome spacing in non-overlapping 100 kb bins, each containing 500 nucleosome calls, is calculated genome-wide. A/B compartment predictions, also with 100 kb resolution, are shown for GM12878. Compartments A and B are associated with open and closed chromatin, respectively. *(top right and bottom right)*Nucleosome spacing and A/B compartments on chromosomes 7 and 11. A/B segmentation (red and blue bars) largely recapitulates chromosomal G-banding (ideograms, gray bars). Median nucleosome spacing (black dots) is calculated in 100 kb bins.

flects footprinting of the transcription pre-initiation complex, or some component thereof, at transcriptionally active genes.

These observations support our thesis that cfDNA fragmentation patterns indeed contain signal that might be used to infer the tissue(s) or cell-type(s) giving rise to cfDNA. However, a challenge is that relatively few reads in a genome-wide cfDNA library directly overlap DHS sites and TSSs.

It was previously observed that nucleosome spacing varies between cell types as a function of chromatin state and gene expression (Teif et al., 2012; Valouev et al., 2012). In general, open chromatin and transcription are associated with a shorter nucleosome repeat length, consistent with our analyses of compartment A vs. B (Figure 4.13). In our peak calls, we also observe a correlation between nucleosome spacing across gene bodies and their expression levels, with tighter spacing associated with higher expression (Figure 4.17A; ρ = -0.17; n = 19,677 genes). The correlation is highest for the gene body itself, rel-

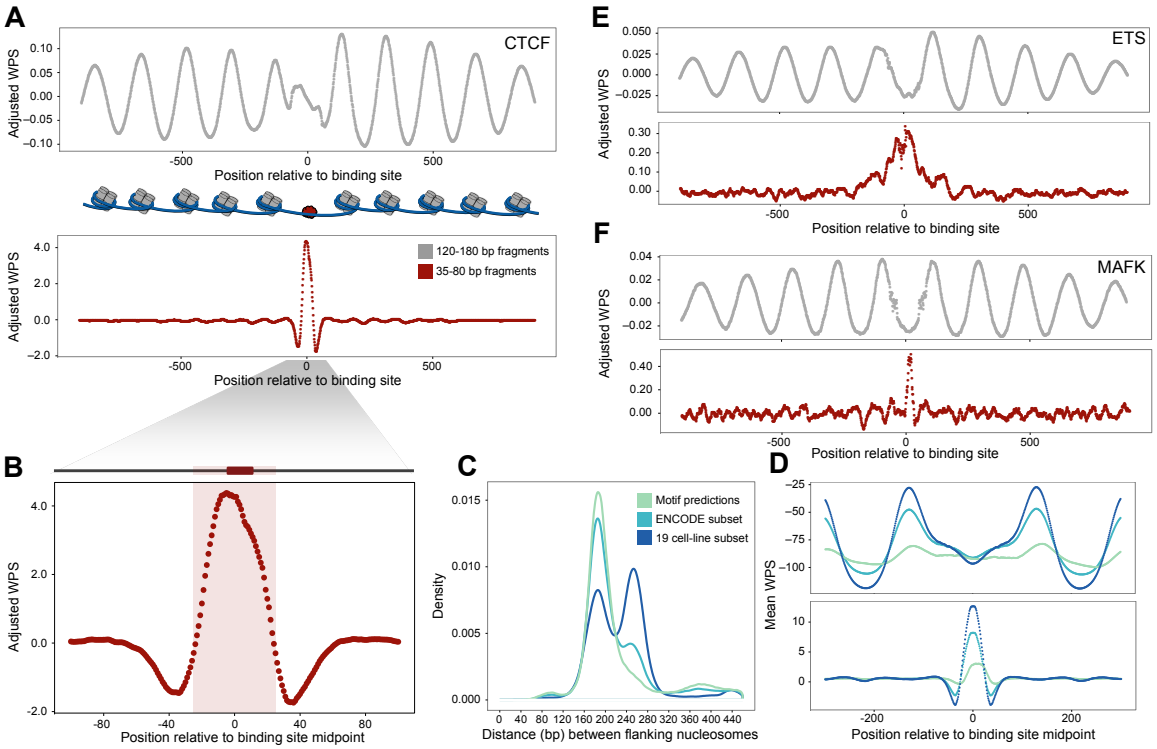


Figure 4.14: **Short cfDNA fragments footprint CTCF and other TF binding sites**. Clustered FIMO predictions were intersected with ChIP-seq data to obtain confident sets of binding site predictions for various TFs. Aggregate, adjusted WPS was calculated for both the long (120-180 bp) and short (35-80 bp) fractions of cfDNA fragments. Higher WPS values indicate greater nucleosome or TF protection, respectively. *(A)* Aggregate, adjusted WPS for 518,632 predicted CTCF binding sites for the long *(top)* and short *(bottom)* cfDNA fractions. Binding of CTCF results in strong positioning of neighboring nucleosomes. *(B)* Aggregate, adjusted WPS, calculated for 518,632 predicted CTCF sites as in *A* and magnified for detail, for 35–80 bp cfDNA fragments. The pink shading indicates the larger 52 bp CTCF binding motif, and the black box shows the location of the 17 bp motif used for FIMO predictions. *(C)* Density of -1 to +1 nucleosome spacing around CTCF sites derived from clustered FIMO predictions (purely motif-based: 518,632 sites), a subset of these predictions overlapping with ENCODE ChIP-seq peaks (93,530 sites), and a further subset active across 19 cell lines (23,723 sites). Flanking nucleosome spacing at the least stringent set of sites (motif-based) mirrors the genome-wide average (~185 bp), while spacing at the most stringent set of sites is highly enriched for greater distances (~260 bp), consistent with active CTCF binding and repositioning of adjacent nucleosomes. *(D)* Mean WPS calculated for the long *(top)* and short *(bottom)* cfDNA fractions at the sets of CTCF sites in *C*.*(E and F)* Aggregate, adjusted WPS calculated for both long *(top)* and short *(bottom)* cfDNA fractions at predicted binding sites for ETS (210,798 sites) *(E)* and MAFK (32,159 sites) *(F)*. For both factors, short fraction WPS is consistent with TF-conferred protection of the binding site, whereas long fraction WPS evidences regular, local positioning of surrounding nucleosomes.

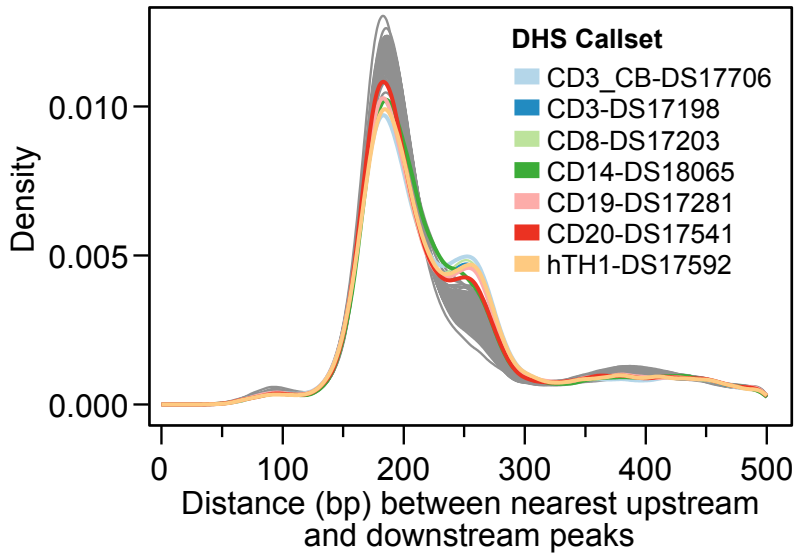


Figure 4.15: **Nucleosome spacing around DHSes in 116 callsets**. Nucleosome spacing around DHSes in 116 callsets. The distribution of nucleosome spacing for peaks flanking DHS sites is bimodal, plausibly corresponding to widened nucleosome spacing at active DHS sites due to intervening TF binding (~190 bp → ~260 bp). Lymphoid or myeloid callsets have the largest proportions of DHS sites with widened nucleosome spacing, consistent with hematopoietic cell death as the dominant source of cfDNA in healthy individuals.

ative to adjacent regions (upstream 10 kb ρ = -0.08; downstream 10 kb ρ = -0.01). If we limit this analysis to gene bodies that span at least 60 nucleosome calls, the correlation is much stronger (ρ = -0.50; n = 12,344 genes).

An advantage of exploiting signals such as nucleosome spacing across gene bodies or other domains is that a much larger proportion of cfDNA fragments will be informative, and moreover we might be able to detect mixtures of signals resulting from multiple cell types contributing to cfDNA. To test this, we performed fast Fourier transformation (FFT) on the L-WPS across the first 10 kb of gene bodies and on a gene-by-gene basis. The intensity of the FFT signal is correlated with gene expression at specific frequency ranges, with a maximum at 177-180 bp for positive correlation and a minimum at ~199 bp for negative correlation (Figure 4.17B). In performing this analysis against a dataset of 76 expression datasets for

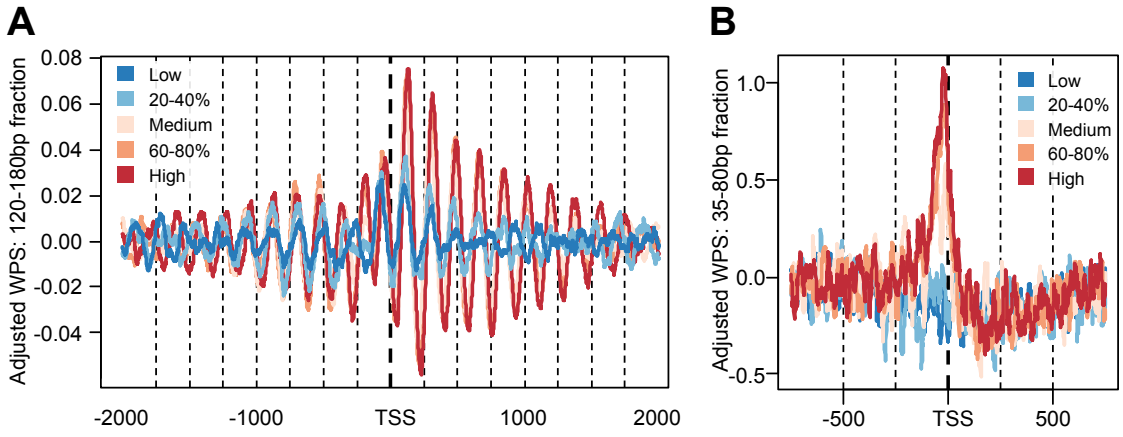


Figure 4.16: **WPS profiles stratified by gene expression**. WPS profiles stratified by gene expression. Partitioning adjusted WPS scores around TSS into five gene expression bins (quintiles) defined for NB-4 (an acute promyelocytic leukemia cell line) reveals differential nucleosome spacing and positioning. *(A)* Highly expressed genes show strong nucleosome phasing within the transcript body. Upstream of the TSS, -1 nucleosomes are well-positioned across expression bins, but -2 and -3 nucleosomes are well-positioned only for medium to highly expressed genes. *(B)* For medium to highly expressed genes, a short fragment WPS peak is observed between the TSS and the -1 nucleosome, plausibly footprinting some or all of the transcription preinitiation complex at transcriptionally active genes.

human cell lines and primary tissues (Uhlén et al., 2015), we observe that the strongest correlations are with hematopoietic lineages (Figure 4.17B). For example, the most highly ranked negative correlations with average intensity in the 193-199 bp frequency range for each of three healthy samples (BH01, IH01, IH02) are all to lymphoid cell lines, myeloid cell lines, or bone marrow tissue (Figure 4.18; Table 4.3). These top correlation ranks are robust to downsampling (Figure 4.19).

4.2.5 Nucleosome spacing in cancer patients' cfDNA identifies non-hematopoietic contributions

We next sought to ask whether we could detect signatures of non-hematopoietic cell types contributing to circulating cfDNA in non-healthy states. We first screened 44 plasma samples from individuals with clinical diagnoses of a variety of Stage IV cancers with light se-

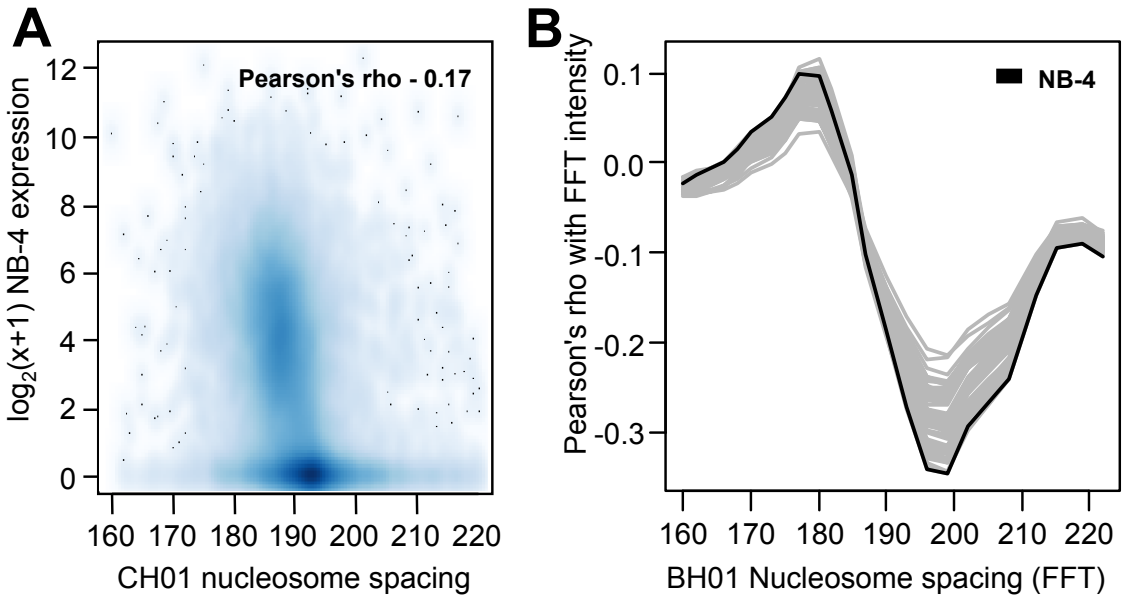
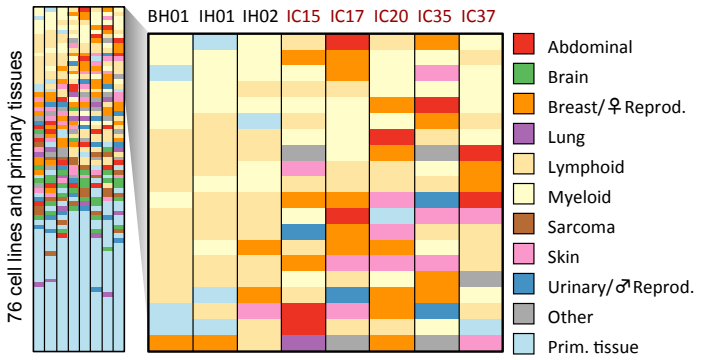


Figure 4.17: **Correlation of gene expression and inferred nucleosome spacing**. Correlation of gene expression and inferred nucleosome spacing. *(A)* Median nucleosome spacing in the transcript body is negatively correlated with gene expression in NB-4 (ρ = -0.17, n = 19,677 genes). Spacing in genes with low or no expression is 193 bp, while spacing in expressed genes ranges from 186 to 193 bp. *(B)* To deconvolve multiple contributions, intensities from fast Fourier transformation (FFT) quantified the specific frequency contributions in the long fragment WPS for 10 kb windows downstream of each TSS. Shown are correlation trajectories for RNA expression in 76 cell lines and primary tissues at different frequencies. Correlations are strongest for intensities in the 193-199 bp frequency range.

quencing of single-stranded libraries prepared from cfDNA (Table 4.4; median 2.2-fold coverage; of note, with same protocol and many in the same batch as IH02). Because matched tumor genotypes were not available, we scored each sample on two metrics of aneuploidy to identify a subset likely to contain a high proportion of tumor-derived cfDNA: first, the deviation from the expected proportion of reads derived from each chromosome (Leary et al., 2012); and second, the per-chromosome allele balance profile for a panel of common single nucleotide polymorphisms. Based on these metrics, we sequenced single-stranded libraries derived from five individuals (with a small cell lung cancer, a squamous cell lung cancer,



Sample	Cancer	Top 3 rank increases
IC15 ♂	Lung (SCLC)	SCLC-21H (31), SH-SY5Y (25), HEK 293 (17)
IC17 ♂	Liver (HCC)	kidney (21), CAPAN-2 (19), BEWO (19)
IC20 ♂	Lung (SCC)	SK-BR-3 (21), Hep G2 (18), HaCaT (18)
IC35 ♀	Breast (DC)	BEWO (27), SiHa (27), CAPAN-2 (25)
IC37 ♀	Colorectal (AC)	Hep G2 (24), SK-BR-3 (22), EFO-21 (20)
Control	-	duodenum (10), U-251 MG (9), small intestine (9)

Figure 4.18: **Inference of mixtures of cell types contributing to cell-free DNA in healthy states and cancer**. Inference of mixtures of cell types contributing to cell-free DNA in healthy states and cancer. The ranks of correlation for 76 RNA expression datasets with average intensity in the 193-199 bp frequency range for various cfDNA libraries are shown, categorized by type and listed from highest *(top row)* to lowest rank *(bottom row)*. Correlation values and full cell line or tissue names are provided in Table 4.3. All of the strongest correlations for all three healthy samples (BH01, IH01 and IH02; first three columns) are with lymphoid and myeloid cell lines or with bone marrow. In contrast, cfDNA samples obtained from stage IV cancer patients (IC15, IC17, IC20, IC35, IC37; last five columns) show top correlations with various cancer cell lines, e.g. IC17 (hepatocellular carcinoma, HCC) showing highest correlations with HepG2 (HCC cell line), and IC35 (breast ductal carcinoma, DC) with MCF7 (metastatic breast adenocarcinoma cell line). When comparing cell line/tissue ranks observed for the cancer samples to each of the three healthy samples and averaging the rank changes, maximum rank changes are over two-fold higher than those observed from comparing the three healthy samples with each other and averaging rank changes ('Control'). For example, for IC15 (small cell lung carcinoma, SCLC) the rank of SCLC-21H (SCLC cell line) increased by an average of 31 positions, for IC20 (squamous cell lung carcinoma, SCC) SK-BR-3 (metastatic breast adenocarcinoma cell line) increased by an average rank of 21, and for IC37 (colorectal adenocarcinoma, AC) HepG2 increased by 24 ranks.

a colorectal adenocarcinoma, a hepatocellular carcinoma, and a ductal carcinoma *in situ* breast cancer) to a depth similar to that of IH02 (Table 4.5).

We again performed FFT on the L-WPS across gene bodies and correlated the average intensity in the 193-199 bp frequency range against the same 76 expression datasets for human cell lines and primary tissues (Uhlén et al., 2015). In contrast with the three samples from healthy individuals (where all of the top 10, and nearly all of the top 20, correlations were to lymphoid or myeloid lineages), we observe that many of the most highly ranked

cell lines or tissues represent non-hematopoietic lineages, in some cases aligning with the cancer type (Figure 4.18). For example, for IC17, where the patient had a hepatocellular carcinoma, the top-ranked correlation is with HepG2, a hepatocellular carcinoma cell line. For IC35, where the patient had a ductal carcinoma *in situ* breast cancer, the top-ranked correlation is with MCF7, a metastatic breast adenocarcinoma cell line. In other cases, the cell lines or primary tissues that exhibit the greatest change in correlation rank align with the cancer type. For example, for IC15, where the patient had small-cell lung cancer, the largest change in correlation rank (-31) is for a small-cell lung cancer cell line (SCLC-21H), and the second largest change (-25) is for a neuroblastoma cancer cell line (SH.SY5Y). For IC20 (a lung squamous cell carcinoma) and IC35 (a colorectal adenocarcinoma), there are many non-hematopoietic cancer cell lines displacing the lymphoid/myeloid cell lines in terms of correlation rank, but the alignment of these to the specific cancer type is less clear. It is possible that the molecular profile of these patients' cancers is not well-represented amongst our 76 expression datasets (none are lung squamous cell carcinomas; CACO-2 is a cell line derived from a colorectal adenocarcinoma, but is highly heterogeneous (Sambuy et al., 2005)). As with samples from healthy individuals, the top correlation ranks associated with the samples from cancer patients are robust to downsampling (Figure 4.19). However, *in silico* "dilution" of samples from cancer patients with samples from healthy patients results in proportionally lower ranks for non-hematopoietic cell lines, consistent with expectation (Figure 4.20).

4.3 Discussion

We present a dense, genome-wide map of *in vivo* nucleosome protection inferred from plasma-borne cfDNA fragments. Although the number of peaks is essentially saturated in CH01, other metrics of quality continued to be a function of sequencing depth (Figure 4.7). We therefore constructed an additional genome-wide nucleosome map based on all of the cfDNA sequencing that we have performed to date ('CA01', 14.5G fragments; 700-fold coverage; 13.0M peaks). Although this map exhibits even more uniform spacing (Figure 4.21)

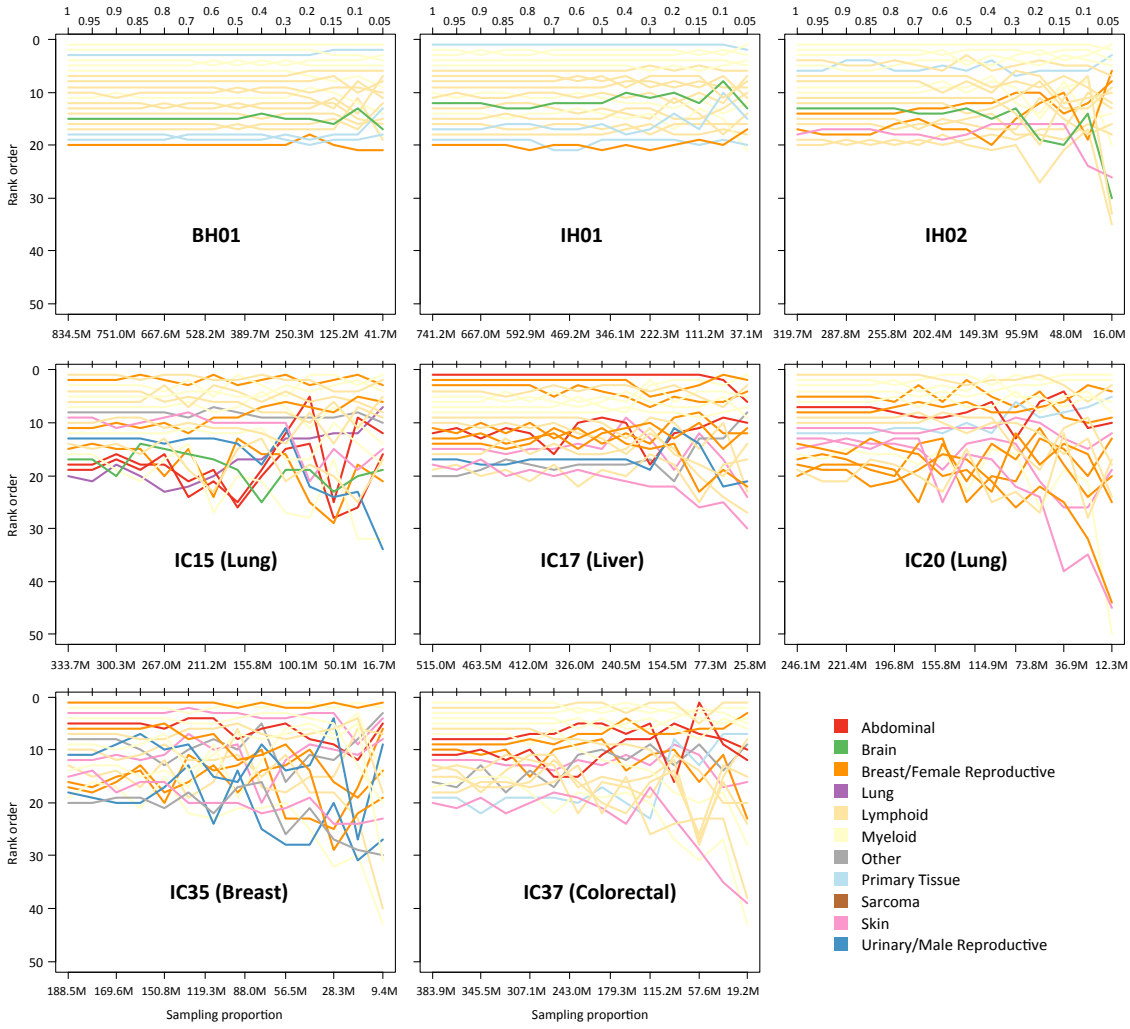


Figure 4.19: **Stability of inference of cfDNA tissues-of-origin to downsampling**. Impact of reduced sequencing coverage. Each of the deeply sequenced cfDNA libraries, including three libraries derived from healthy individuals (BH01, IH01, IH02) and five libraries derived from cancer patients (IC15, IC17, IC20, IC35, IC37), were downsampled in fixed proportions *(top horizontal axis)* to reduce the effective sequencing coverage. Downsampling was performed by randomly removing examined fragments until the desired target coverage was reached *(bottom horizontal axis)*. After reducing coverage, a WPS track was generated for each downsampled library. Each WPS track was analyzed in the same manner as the samples with full coverage. The rank order of contributing tissues and cell lines in each downsampled library is compared to the rank order observed in the full coverage library. The trajectories of the ranks of each tissue or cell line falling within the top 20 highest negative correlations in the full coverage library are plotted for each sampling proportion *(vertical axis)*.

and more highly supported peak calls *(not shown)*, we caution that it is based on cfDNA from both healthy and non-healthy individuals (Tables 4.1 and 4.5).

Our work builds directly on previous efforts to map nucleosome occupancy in human cells genome-wide (Gaffney et al., 2012; Pedersen et al., 2014; Schones et al., 2008; Teif et al., 2012; Valouev et al., 2012), but our callset is substantially more complete and uniform (Figure 4.21). The fragments that we observe are generated by endogenous physiological processes, avoiding the technical variation associated with *in vitro* MNase digestion. A limitation of our map is that the cell types that give rise to cfDNA are inevitably heterogeneous (e.g. a mixture of lymphoid and myeloid cell types in healthy individuals). Nonetheless, the map's relative completeness may facilitate a deeper understanding of the interplay of nucleosome positioning and spacing with primary sequence, epigenetic regulation, transcriptional output, and nuclear architecture.

A second goal of this study was to explore whether the nucleosome footprints contained in cfDNA fragments can be used to infer contributing cell types. Through comparisons with gene expression and regulatory site profiles, we identify the epigenetic signature of hematopoietic lineages contributing to cfDNA in healthy individuals, with plausible additional contributions from one or more non-hematopoietic tissues in a small panel of individuals with advanced cancers. For this proof-of-concept, we stacked the odds in our favor by focusing individuals that appeared to have large burdens of tumor-derived DNA. However, it should be emphasized that in the context of cancer, our goal is not necessarily to outperform the sensitivity of mutation-based monitoring of circulating tumor DNA. Rather, we envision that a unique application of this approach may be to non-invasively classify cancers at time-of-diagnosis by matching the epigenetic signature of cfDNA fragmentation patterns against reference datasets corresponding to diverse cancer types. For example, this may have value for non-invasively and molecularly classifying "cancers of unknown primary", which comprise 4-5% of all invasive cancers (Greco and Hainsworth, 2009), as well as cancers where invasive biopsies are currently required for definitive diagnosis and/or for subtyping (e.g. lung cancer).

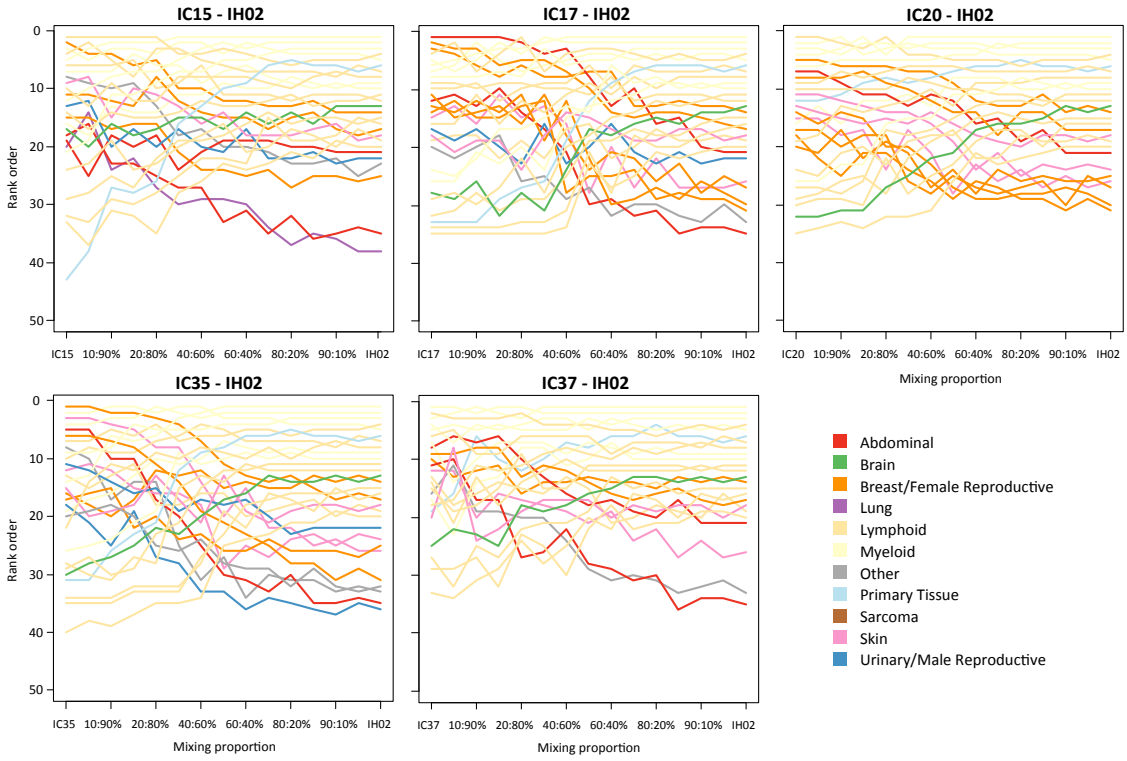


Figure 4.20: **Stability of inference of cfDNA tissues-of-origin to dilution**. *In silico* dilutions of cancer samples. Each of the five cfDNA samples from cancer patients sequenced to higher coverage was diluted by the addition of sequenced fragments derived from a sample from a healthy individual (IH02) at fixed proportions *(horizontal axis)*. The total number of fragments examined was held constant at each mixture proportion. After dilution, a WPS track was generated for each mixture proportion in each sample. Each WPS track was analyzed in the same manner as the undiluted samples. The rank order of contributing tissues and cell lines in each mixed sample is compared to the order observed in the original, undiluted sample. The trajectories of the ranks of each tissue or cell line falling within the top 20 highest negative correlations in either the undiluted sample or IH02 are plotted *(vertical axis)*. The IH02 library and each of the cancer sample-derived libraries were constructed with the single-stranded library preparation protocol.

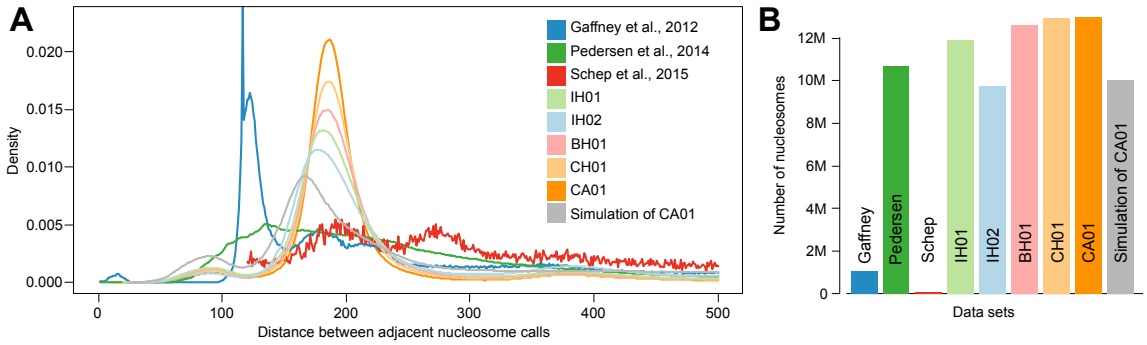


Figure 4.21: **Comparison of uniformity and completeness of nucleosome callsets**. Comparison of uniformity and completeness of nucleosome callsets. (A) Distance between nucleosome peak calls across three published data sets (Gaffney et al., 2012; Pedersen et al., 2014; Schep et al., 2015) and calls produced in this study. Previously published callsets lack one defined mode at the canonical 185 bp nucleosome spacing, possibly due to sparse sampling or wide call ranges. In contrast, all the nucleosome calls from cfDNA show one well-defined mode, the magnitude of which increases with the number of fragments examined. The callset produced from simulation has a lower mode (166 bp) and a wider distribution. (B) Number of calls in each set. The densest cfDNA-derived callset contains nearly 13M nucleosome calls.

In addition, there are a range of non-malignant conditions for which it may be valuable to explore the nucleosome and TF footprints contained in cfDNA as markers for acute or chronic tissue damage, e.g. myocardial infarction (Chang et al., 2003), stroke (Rainer et al., 2003) and autoimmune disorders (Galeazzi et al., 2003). Contributions from these tissues to cfDNA cannot be readily detected under the current paradigm of analyzing genotypic differences, which are effectively non-existent in these conditions. By contrast, the approach presented here should generalize to detecting contributions to cfDNA from any non-hematopoietic cell lineage (and, possibly, grossly aberrant contributions from hematopoietic cell lineages).

Alternative "genotype-independent" approaches for using circulating nucleic acids as markers for disease include cell-free RNA (Koh et al., 2014) and DNA methylation (Sun et al., 2015). Although these merit exploration, tumor-derived cell-free RNA appears to be much less stable than nucleosome-bound cfDNA (García-Olmo et al., 2013), while bisulfite

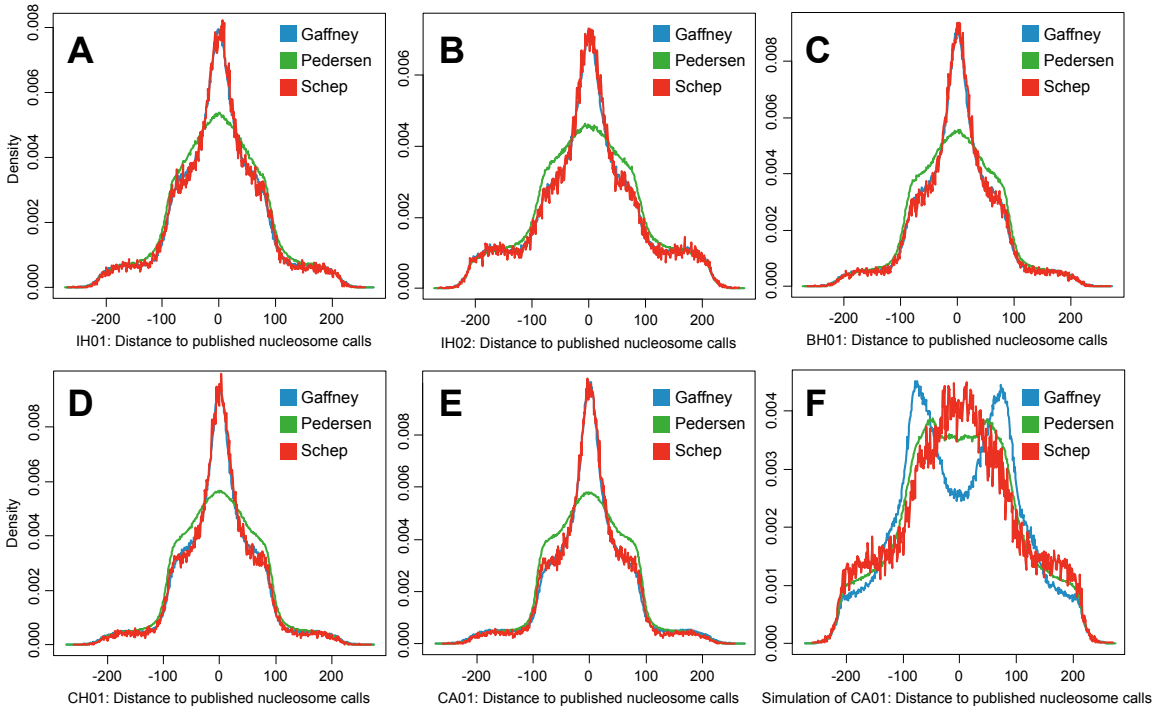


Figure 4.22: **Comparison of peak locations between samples and call sets**. Comparison of peak locations between samples and call sets. For each pair of samples, the distribution of distances between each peak call in the sample with fewer peaks and the nearest peak call in the other sample is shown. Negative numbers indicate the nearest peak is upstream; positive numbers indicate the nearest peak is downstream. Concordance between callsets increases with the number of cfNDA fragments examined.

sequencing libraries are challenging to robustly construct from small amounts of starting material.

A limitation of this study is the small number of samples studied (n = 8) and the relatively small size of the reference dataset of cell lines and tissues against which these samples were compared (n = 76). We anticipate that increasing the number of samples studied, as well as the range of physiological states and diseases with which these samples are associated, is necessary to fully evaluate the potential and limitations of this approach. Furthermore, expanding the breadth and quality of the reference datasets against which these sam-

ples are compared (e.g., directly comparing to cell-type-specific nucleosome maps, rather than to expression profiles), is likely to improve the ability to robustly assign and quantify contributing cell types.

Cell-free DNA has tremendous potential as a “liquid biopsy”, and indeed its use in non-invasive prenatal screening for fetal trisomies has vastly outpaced all other applications of DNA sequencing in terms of clinical uptake. In contrast with current paradigms for analyzing cfDNA, we show how the information contained in cfDNA fragmentation patterns, effectively the footprints of protein-DNA interactions, can be used to infer contributing cell types without relying on genotypic differences. To the extent that cfDNA composition is impacted by cell death consequent to malignancy, acute or chronic tissue damage, or other conditions, this method may substantially expand the range of clinical scenarios in which cfDNA sequences comprise a clinically useful biomarker.

4.4 Methods

4.4.1 Plasma Samples

Bulk human peripheral blood plasma, containing contributions from an unknown number of healthy individuals, was obtained from STEMCELL Technologies (Vancouver, British Columbia, Canada). Anonymous, individual human peripheral blood plasma from healthy donors, donors with clinical diagnosis of Stage IV cancer, and donors with clinical diagnosis with autoimmune disease (Tables 4.1, 4.4, and 4.5) was obtained from Conversant Bio (Huntsville, Alabama, USA) or PlasmaLab International (Everett, Washington, USA). Plasma was stored at -80°C and thawed on the bench-top immediately before use. Cell-free DNA was purified from each sample with the QiaAMP Circulating Nucleic Acids kit (Qiagen) as per the manufacturer's protocol. DNA was quantified with a Qubit fluorometer (Invitrogen).

4.4.2 Preparation of double-stranded sequencing libraries

Conventional, double-stranded sequencing libraries were prepared with the ThruPLEX-FD or ThruPLEX DNA-seq kits (Rubicon Genomics), comprising a proprietary series of end-repair, ligation, and amplification reactions. Libraries were prepared with 0.5-30.0 ng of cfDNA input and individually barcoded. Library amplification was monitored by real-time PCR and was typically terminated after 4-6 cycles.

4.4.3 Preparation of single-stranded sequencing libraries

Single-stranded sequencing libraries were prepared with a protocol adapted from Gansauge and Meyer (2013) as follows: Adapter 2 was prepared by combining 4.5 ul TE (pH 8), 0.5 ul 1M NaCl, 10 uL 500 uM oligo Adapter2.1, and 10 ul 500 uM oligo Adapter2.2, incubating at 95°C for 10 seconds, and ramping to 14°C at a rate of 0.1°C/s. Purified cfDNA fragments were dephosphorylated by combining 2X CircLigase II buffer (Epicentre), 5 mM MnCl2, and 1U FastAP (Thermo Fisher) with 0.5-10 ng fragments in 20 ul reaction volume and in-

cubating at 37°C for 30 minutes. Fragments were then denatured by heating to 95°C for 3 minutes, and were immediately transferred to an ice bath. The reaction was supplemented with biotin-conjugated adapter oligo CL78 (5 pmol), 20% PEG-6000 (w/v), and 200U CircLigase II (Epicentre) for a total volume of 40 ul, and was incubated overnight with rotation at 60°C, heated to 95°C for 3 minutes, and placed in an ice bath. For each sample, 20 ul MyOne C1 beads (Life Technologies) were twice washed in bead binding buffer (BBB) (10 mM Tris-HCl [pH 8], 1M NaCl, 1 mM EDTA [pH 8], 0.05% Tween-20, and 0.5% SDS), and resuspended in 250 ul BBB. Adapter-ligated fragments were bound to the beads by rotating for 60 minutes at room temperature. Beads were collected on a magnetic rack and the supernatant was discarded. Beads were washed once with 500 ul wash buffer A (WBA) (10 mM Tris-HCl [pH 8], 1 mM EDTA [pH 8], 0.05% Tween-20, 100 mM NaCl, 0.5% SDS) and once with 500 ul wash buffer B (WBB) (10 mM Tris-HCl [pH 8], 1 mM EDTA [pH 8], 0.05% Tween-20, 100 mM NaCl). Beads were combined with 1X Isothermal Amplification Buffer (NEB), 2.5 uM oligo CL9, 250 uM (each) dNTPs, and 24U Bst 2.0 DNA Polymerase (NEB) in a reaction volume of 50 ul, incubated with gentle shaking by ramping temperature from 15°C to 37°C at 1°C/minute, and held at 37°C for 10 minutes. After collection on a magnetic rack, beads were washed once with 200 ul WBA, resuspended in 200 ul of stringency wash buffer (SWB) (0.1X SSC, 0.1% SDS), and incubated at 45°C for 3 minutes. Beads were again collected and washed once with 200 ul WBB. Beads were then combined with 1X CutSmart Buffer (NEB), 0.025% Tween-20, 100 uM (each) dNTPs, and 5U T4 DNA Polymerase (NEB) and incubated with gentle shaking for 30 minutes at room temperature. Beads were washed once with each of WBA, SWB, and WBB as described above. Beads were then mixed with 1X CutSmart Buffer (NEB), 5% PEG-6000, 0.025% Tween-20, 2 uM double-stranded Adapter 2, and 10U T4 DNA Ligase (NEB), and incubated with gentle shaking for 2 hours at room temperature. Beads were washed once with each of WBA, SWB, and WBB as described above, and resuspended in 25 ul TET buffer (10 mM Tris-HCl [pH 8], 1 mM EDTA [pH 8], 0.05% Tween-20). Second strands were eluted from beads by heating to 95°C, collecting beads on a magnetic rack, and transferring the supernatant to a new tube. Library amplifi-

cation was monitored by real-time PCR, requiring an average of 4-6 cycles per library.

4.4.4 Sequencing and primary data processing

All libraries were sequenced on HiSeq 2000 or NextSeq 500 instruments (Illumina). Details of sequencing are provided in Tables 4.1 and 4.5. Barcoded paired-end (PE) sequencing data was split allowing up to one substitution in the barcode sequence. Fragments shorter than or equal to the read length were consensus-called and adapter-trimmed. Remaining consensus single-end reads (SR) and the individual PE reads were aligned to the human reference genome (GRCh37, 1000 Genomes phase 2 technical reference, `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/`) using the ALN algorithm in BWA v0.7.10 (Li and Durbin, 2010). PE reads were further processed with BWA SAMPE to resolve ambiguous placement of read pairs or to rescue missing alignments by a more sensitive alignment step around the location of one placed read end. Aligned SR and PE data were stored in BAM format using the samtools API (Li et al., 2009). BAM files for each sample were merged across lanes and sequencing runs.

4.4.5 Simulated reads and nucleotide frequences

Sequencing data was simulated procedurally to mimic observed cleavage and ligation biases and length distributions. Specifically, aligned sequencing data was simulated (SR if shorter than 45 bp, PE 45 bp otherwise) for all major chromosomes of the human reference (GRC37h). Dinucleotide frequencies were determined from real data on both fragment ends and both strand orientations, and for the reference genome on both strands. The insert size distribution of the real data was extracted for the 1-500 bp range. Reads were simulated procedurally: at each step (i.e., at least once at each genomic coordinate, depending on desired coverage), (1) the strand is randomly chosen, (2) the ratio of the dinucleotide frequency in the real data to that in the reference sequence is used to randomly decide whether the initiating dinucleotide is considered, (3) an length is sampled from the

insert size distribution, and (4) the frequency ratio of the terminal dinucleotide is used to randomly decide whether the generated alignment is reported. The simulated coverage was matched to that of the original data after PCR duplicate removal.

4.4.6 Analysis of nucleotide composition of 167 bp fragments

Fragments with inferred lengths of exactly 167 bp were filtered within samples to remove duplicates. Dinucleotide frequencies were calculated in a strand-aware manner, using a sliding 2 bp window and reference alleles at each position, beginning 50 bp upstream of one fragment endpoint and ending 50 bp downstream of the other endpoint. Observed dinucleotide frequencies at each position were compared to expected dinucleotide frequencies determined from a set of simulated reads reflecting the same cleavage biases calculated in a library-specific manner.

4.4.7 Coverage, fragment endpoints, and windowed protection scores

Fragment endpoint coordinates were extracted from BAM files with the SAMtools API (Li et al., 2009). Both outer alignment coordinates of PE data were extracted for properly paired reads. Both end coordinates of SR alignments were extracted when PE data was collapsed to SR data by adapter trimming. A fragment's coverage is defined as all positions between the two (inferred) fragment ends, inclusive of endpoints. We define the Windowed Protection Score (WPS) of a window of size k as the number of molecules spanning the window minus those with an endpoint within the window. We assign the determined WPS to the center of the window. For 35-80 bp fragments (short fraction, S-WPS), k=16; for 120-180 bp fragments (long fraction, L-WPS), k=120.

4.4.8 Nucleosome peak calling

L-WPS is locally adjusted to a running median of zero in 1 kb windows and smoothed using a Savitzky-Golay filter (Savitzky and Golay, 1964) (window size 21, 2nd order polynomial). The L-WPS track is then segmented into above-zero regions (allowing up to 5 consecutive

positions below zero). If the resulting region is 50-150 bp, we identify the median L-WPS value of that region and search for the maximum-sum contiguous window above the median. We report the start, end and center coordinates of this window as the "peak," or local maximum of nucleosome protection. All calculations involving distances between peaks are based on these center coordinates. A score for each peak is determined as the distance between maximum value in the window and the average of the two adjacent L-WPS minima neighboring the region. If the identified region is 150-450 bp, we apply the same above median contiguous window approach, but only report those windows that are 50-150 bp. For score calculation of multiple windows derived from 150-450 bp regions, we set the neighboring minima to zero. We discard regions <50 bp or >450 bp.

4.4.9 Analysis of TFBS, DHS sites, and genic features

Features were aggregated and aligned at starting coordinates while adjusting for strand and direction of transcription. TFBS sets were obtained by filtering motif predictions with ChIP-seq peaks. For most features, L-WPS values were adjusted to account for signal observed in matched simulations.

For TFBS, we started with clustered FIMO (motif-based) intervals (Grant et al., 2011; Maurano et al., 2012) defining a set of computationally predicted binding sites. For a subset of clustered TFs (AP-2-2, AP-2, CTCF_Core-2, E2F-2, EBF1, Ebox-CACCTG, Ebox, ESR1, ETS, MAFK, MEF2A-2, MEF2A, MYC-MAX, PAX5-2, RUNX2, RUNX-AML, STAF-2, TCF-LEF, YY1), we retained only predicted TFBSs that overlap with ENCODE ChIP-seq peaks (TfbsClusteredV3 set downloaded from
`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/`).

WPS values for CH01 and the corresponding simulation were extracted for each position in a 5 kb window around the start coordinate of each TFBS, and were aggregated within each TF cluster. The mean WPS of the first and last 500 bp (which is predominantly flat and represents a mean offset) of the 5 kb window was subtracted from the original WPS at each

position. For L-WPS only, a sliding window mean is calculated using a 200 bp window and subtracted from the original signal. Finally, the corrected WPS profile for the simulation is subtracted from the corrected WPS profile for CH01 to correct for signal that is a product of fragment length and ligation bias. This final profile is plotted and termed the Adjusted WPS. In figures, CTCF binding sites are shifted such that the zero coordinate on the x-axis is the center of its 52 bp binding footprint (Ong and Corces, 2014). Genomic coordinates of transcription start sites, transcription end sites, start codons, and splice donor and acceptor sites were obtained from Ensembl Build version 75. Adjusted WPS surrounding these features was calculated as for TFBSs.

CTCF sites first included clustered FIMO binding site predictions (described above). This set was intersected with ENCODE ChIP-seq peaks (TfbsClusteredV3, described above), and then further intersected with a set of CTCF binding sites experimentally observed to be active across 19 tissues (Wang et al., 2012), to produce three increasingly stringent sets. For each CTCF site, distances between each of 20 flanking nucleosomes (10 upstream and 10 downstream) were calculated. The mean S-WPS and L-WPS at each position relative to the center of the CTCF binding motif were also calculated within bins defined by spacing between -1 and +1 nucleosomes (>160 bp, 161-200 bp, 201-230 bp, 231-270 bp, 271-420 bp, 421-460 bp, and >460 bp).

DHS peaks for 349 primary tissue and cell line samples were downloaded from `http://www.uwencode.org/proj/Science_Maurano_Humbert_et_al/data/all_fdr0.05_hot.tgz`. Samples derived from fetal tissues (233 of 349) were removed from the analysis as they behaved inconsistently within tissue type, possibly because of unequal representation of cell types within each tissue sample. 116 DHS callsets from a variety of cell lineages were retained for analysis. For the midpoint of each DHS peak in a particular set, the nearest upstream and downstream calls in the CH01 callset were identified, and the distance between the centers of those two calls was calculated. The distribution of all such distances was visualized for each DHS peak callset using a smoothed density estimate calculated for distances between 0 and 500 bp.

4.4.10 Gene expression analysis

FPKM gene expression (GE) values measured for 20,344 Ensembl gene identifiers in 44 human cell lines and 32 primary tissues by the Human Protein Atlas (Uhlén et al., 2015) was downloaded from `http://www.proteinatlas.org/download/rna.csv.zip`. Genes with 3 or more non-zero expression values were retained (n=19,378 genes). The GE data set is provided with one decimal precision for the FPKM values. Thus, a zero GE value (0.0) indicates expression in the interval [0, 0.05) Unless otherwise noted, we set the minimum GE value to 0.04 FPKM before log2-transformation.

4.4.11 Fourier transformation and correlation with expression

L-WPS was used to calculate periodograms of genomic regions using Fast Fourier Transform (FFT, spec.pgram in R) with frequencies between 1/500 and 1/100 bases. See Supplemental Experimental Procedures for details. Intensity values for the 120-280 bp frequency range were determined from smooth FFT periodograms. S-shaped Pearson correlation between GE values and FFT intensities was observed around the major inter-nucleosome distance peak, along with a pronounced negative correlation in the 193 - 199 bp frequency range. The mean intensity in this frequency range was correlated with the average intensity with log2-transformed GE values for downstream analysis.

4.4.12 Fourier transformation and smoothing of trajectories

We use parameters to smooth (3 bp Daniell smoother; moving average giving half weight to the end values) and de-trend the data (i.e. subtract the mean of the series and remove a linear trend). A recursive time series filter implemented in R was used to remove high frequency variation from trajectories. 24 filter frequencies (1/seq(5,100,4)) were used, and the first 24 values of the trajectory were taken as init values. The 24-value shift in the resulting trajectories was corrected by repeating the last 24 values of the trajectory.

Acknowledgements

We thank D. May, J. Vierstra, M. Maurano, and members of the Shendure lab for helpful discussions. This work was funded in part by an NIH Director's Pioneer Award (1DP1HG007811 to J.S.). J.S. is an investigator of the Howard Hughes Medical Institute. A patent application has been filed for aspects of the methods disclosed here (M.W.S., M.K., and J.S.: "Methods of determining tissues and/or cell types giving rise to cell-free DNA, and methods of identifying a disease or disorder using same"; PCT/US2015/042310).

Table 4.1: **Sequencing statistics for samples BH01, IH01, IH02, and CH01**. For each sample, sequencing-related statistics, including the total number of fragments sequenced, read lengths, the percentage of such fragments aligning to the reference with and without a mapping quality threshold, mean coverage, duplication rate, and the proportion of sequenced fragments in two length bins, are tabulated. Fragment length is inferred from alignment of paired-end reads. Due to the short read lengths, coverage is calculated by assuming the entire fragment had been read. The estimated number of duplicate fragments is based on fragment endpoints, which may overestimate the true duplication rate in the presence of highly stereotyped cleavage.
SSP, single-stranded library preparation protocol. DSP, double-stranded library preparation protocol. .

Sample	Library type	Reads	Fragments sequenced	Aligned	Aligned Q30	Coverage	% Duplicates	35-80 bp	120-180 bp
BH01	DSP	2x101	1489569204	97.20%	88.85%	96.32	6.00%	0.65%	57.64%
IH01	DSP	2x101	1572050374	98.58%	90.60%	104.92	21.00%	0.77%	47.83%
IH02	SSP	2x50, 43/42	779794090	93.19%	75.27%	30.08	20.05%	21.83%	44.00%
CH01	–	–	3841413668	96.95%	86.81%	231.32	14.99%	5.00%	50.85%

Table 4.2: **Synthetic oligos used in preparation of single stranded sequencing libraries**. Sequences adapted from Gansauge et al., 2013. *, phosphorothioate bond. /5Phos/, 5′ phosphorylation. /ddT/, dideoxythymidine. /iSpC3/, C3 spacer. /3BioTEG/, 3′ biotin-TEG.

Oligo name	Sequence (5′-3′)	Notes
CL9	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	HPLC purification
Adapter2.1	CGACGCTCTTCCGATC/ddT/	HPLC purification
Adapter2.2	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAG*T*G*T*A	HPLC purification
CL78	/5Phos/AGATCGGAAG/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/ iSpC3/iSpC3/iSpC3/iSpC3/3BioTEG/	Dual HPLC purification

Table 4.3: **Correlation of WPS FFT intensities with gene expression datasets**. Correlation values between average FFT (fast Fourier Transformation) intensities for the 193-199bp frequencies in the first 10 kb downstream of the transcriptional start site with FPKM expression values measured for 19,378 Ensembl gene identifiers in 44 human cell lines and 32 primary tissues by the Human Protein Atlas (Uhlén et al., 2015). This table also contains brief descriptions for each of the expression samples as provided by the Protein Atlas as well as rank transformations and rank differences to the IH01, IH02 and BH01 samples.

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
A-431	Skin	Skin cancer (Squamous cells)	Epidermoid carcinoma cell line	-0.2977	-0.1881	-0.1491	-0.2004	-0.1402	-0.1757	-0.1950	-0.1782	2	3	-9	-9	-12	-21
A549	Lung	Lung carcinoma	Lung carcinoma cell line	-0.2890	-0.1854	-0.1441	-0.2015	-0.1390	-0.1717	-0.1877	-0.1698	3	-14	-12	-9	-2	-13
adipose tissue	Primary Tissue	Adipose tissue	Primary tissue	-0.2698	-0.1694	-0.1369	-0.1686	-0.1207	-0.1532	-0.1656	-0.1482	1	12	5	0	14	12
adrenal gland	Primary Tissue	Adrenal gland	Primary tissue	-0.2572	-0.1576	-0.1306	-0.1731	-0.1177	-0.1448	-0.1607	-0.1379	-2	-11	-5	1	5	8
AN3-CA	Breast/Female Reproductive	Uterine cancer	Metastatic endometrial adenocarcinoma cell line	-0.3035	-0.1943	-0.1568	-0.2135	-0.1470	-0.1826	-0.1951	-0.1709	-4	-16	-13	-15	-8	-2
appendix	Primary Tissue	Appendix	Primary tissue	-0.2871	-0.1855	-0.1374	-0.1681	-0.1185	-0.1476	-0.1710	-0.1518	6	24	20	23	8	9
BEWO	Other	Uterine cancer	Metastatic choriocarcinoma cell line	-0.2845	-0.1836	-0.1465	-0.1930	-0.1385	-0.1734	-0.1941	-0.1732	-5	3	-12	-15	-19	-27
bone marrow	Primary Tissue	Bone marrow	Primary tissue	-0.3435	-0.2301	-0.1649	-0.1922	-0.1415	-0.1673	-0.1933	-0.1654	2	40	9	30	16	28
CACO-2	Abdominal	Colon adeno-carcinoma	Colon adenocarcinoma cell line	-0.2814	-0.1770	-0.1366	-0.1925	-0.1278	-0.1694	-0.1842	-0.1637	5	-5	-5	-14	-10	-9
CAPAN-2	Abdominal	Pancreas ade-nocarcinoma	Pancreas adenocarcinoma cell line	-0.2910	-0.1870	-0.1452	-0.2021	-0.1365	-0.1761	-0.1951	-0.1751	3	-12	-2	-18	-19	-25
cerebral cortex	Primary Tissue	Cerebral cortex	Primary tissue	-0.2250	-0.1364	-0.1203	-0.1676	-0.1077	-0.1342	-0.1419	-0.1254	-1	-9	-3	0	0	0
colon	Primary Tissue	Colon	Primary tissue	-0.2608	-0.1622	-0.1240	-0.1644	-0.1106	-0.1446	-0.1679	-0.1477	7	8	8	6	-7	1

Continued on Next Page...

Table 4.3 – Continued

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
Daudi	Lymphoid	Human Burkitt lymphoma	Human Burkitt lymphoma cell line	-0.3213	-0.2061	-0.1533	-0.1949	-0.1329	-0.1654	-0.1889	-0.1599	4	17	19	19	13	24
duode-num	Primary Tissue	Duodenum	Primary tissue	-0.2613	-0.1640	-0.1221	-0.1586	-0.1093	-0.1437	-0.1664	-0.1436	10	10	10	7	-4	7
EFO-21	Breast/Female Reproductive	Ovarian cancer	Metastatic ovarian serous cystadeno-carcinoma cell line	-0.2872	-0.1858	-0.1490	-0.2013	-0.1395	-0.1761	-0.1882	-0.1691	-7	-9	-14	-20	-1	-8
en-dometrium	Primary Tissue	Endometrium	Primary tissue	-0.2574	-0.1584	-0.1324	-0.1776	-0.1186	-0.1507	-0.1658	-0.1512	-3	-11	-4	-8	-3	-12
esophagus	Primary Tissue	Esophagus	Primary tissue	-0.2365	-0.1470	-0.1236	-0.1556	-0.1161	-0.1407	-0.1577	-0.1452	-3	1	-7	0	0	-7
fallopian tube	Primary Tissue	Fallopian tube	Primary tissue	-0.2471	-0.1565	-0.1285	-0.1708	-0.1145	-0.1453	-0.1606	-0.1452	-4	-13	-2	-3	3	-2
gallblad-der	Primary Tissue	Gallbladder	Primary tissue	-0.2489	-0.1556	-0.1193	-0.1530	-0.1029	-0.1378	-0.1540	-0.1406	4	4	4	3	4	1
HaCaT	Skin	Keratinocyte cell line	Keratinocyte cell line	-0.2899	-0.1857	-0.1495	-0.1935	-0.1422	-0.1729	-0.1929	-0.1727	-5	7	-18	-8	-8	-17
HDLM-2	Lymphoid	Hodgkin lymphoma	Hodgkin lymphoma cell line	-0.3155	-0.1999	-0.1535	-0.2010	-0.1365	-0.1734	-0.1947	-0.1713	1	6	11	1	-5	-5
heart muscle	Primary Tissue	Heart muscle	Primary tissue	-0.2463	-0.1493	-0.1257	-0.1662	-0.1130	-0.1410	-0.1547	-0.1396	-3	-3	-3	0	3	2
HEK 293	Other	Kidney adrenal precursor cell line	“Embryonal kidney cell line, transformed by adenovirus type 5”	-0.2920	-0.1873	-0.1501	-0.2085	-0.1390	-0.1720	-0.1894	-0.1681	-4	-17	-4	-2	3	0
HEL	Myeloid	Ery-throleukemia	Erythroleukemia cell line (AML M6 in relapse after treatment for Hodgkin’s disease)	-0.3244	-0.2048	-0.1606	-0.2104	-0.1401	-0.1720	-0.1941	-0.1676	-1	-5	4	12	5	14
HeLa	Breast/Female Reproductive	Cervical cancer	Cervical epithelial adenocarcinoma cell line	-0.2959	-0.1862	-0.1494	-0.2028	-0.1393	-0.1724	-0.1898	-0.1709	1	-10	-5	-3	1	-8
Hep G2	Abdominal	Hepatocellular carcinoma	Hepatocellular carcinoma cell line	-0.2938	-0.1860	-0.1523	-0.2016	-0.1450	-0.1859	-0.1964	-0.1669	-4	-6	-18	-24	-17	2

Continued on Next Page…

Table 4.3 – Continued

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
HL-60	Myeloid	Promyelocytic leukemia	Acute promyelocytic leukemia (APL) cell line	-0.3319	-0.2077	-0.1610	-0.2022	-0.1371	-0.1712	-0.1972	-0.1703	2	8	18	18	-1	11
HMC-1	Myeloid	Mastcell leukemia	Mastcell leukemia cell line	-0.3367	-0.2283	-0.1654	-0.2119	-0.1491	-0.1809	-0.1990	-0.1797	0	-1	-2	3	0	-2
K-562	Lymphoid	Leukemia	Chronic myeloid leukemia (CML) cell line	-0.3173	-0.2023	-0.1580	-0.2110	-0.1432	-0.1776	-0.1945	-0.1657	-3	-9	-5	-6	-1	13
Karpas-707	Lymphoid	Multiple myeloma	Multiple myeloma cell line	-0.3250	-0.2102	-0.1549	-0.1950	-0.1363	-0.1667	-0.1876	-0.1641	4	20	18	22	21	22
kidney	Primary Tissue	Kidney	Primary tissue	-0.2449	-0.1501	-0.1305	-0.1680	-0.1195	-0.1530	-0.1706	-0.1469	-7	-4	-12	-19	-21	-6
liver	Primary Tissue	Liver	Primary tissue	-0.2480	-0.1479	-0.1224	-0.1496	-0.1104	-0.1499	-0.1640	-0.1376	1	4	-1	-13	-4	3
lung	Primary Tissue	Lung	Primary tissue	-0.2638	-0.1695	-0.1329	-0.1702	-0.1213	-0.1482	-0.1672	-0.1489	3	4	0	7	3	6
lymph node	Primary Tissue	Lymph node	Primary tissue	-0.3075	-0.1955	-0.1484	-0.1817	-0.1283	-0.1552	-0.1808	-0.1561	7	24	17	25	14	22
MCF7	Breast/Female Reproductive	Breast cancer	Metastatic breast adenocarcinoma cell line	-0.2979	-0.1948	-0.1545	-0.2070	-0.1447	-0.1830	-0.1958	-0.1815	-3	-9	-12	-18	-11	-19
MOLT-4	Lymphoid	Leukemia (ALL)	Acute lymphoblastic leukemia (T-ALL) cell line	-0.3226	-0.2043	-0.1632	-0.2117	-0.1444	-0.1766	-0.1972	-0.1727	-3	-7	-2	-1	-5	-1
NB-4	Myeloid	Promyelocytic leukemia	Acute promyelocytic leukemia (APL) cell line	-0.3482	-0.2284	-0.1717	-0.2114	-0.1475	-0.1819	-0.2024	-0.1712	0	4	3	5	2	13
NTERA-2	Uri-nary/Male Reproductive	Urinary cancer	“Metastatic embryonal carcinoma cell line, cloned from TERA-2”	-0.2687	-0.1699	-0.1374	-0.1926	-0.1174	-0.1568	-0.1695	-0.1532	-2	-8	16	-2	5	0
ovary	Primary Tissue	Ovary	Primary tissue	-0.2659	-0.1622	-0.1352	-0.1810	-0.1203	-0.1517	-0.1658	-0.1509	1	-7	2	-2	6	-1

Continued on Next Page…

Table 4.3 – Continued

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
pancreas	Primary Tissue	Pancreas	Primary tissue	-0.2503	-0.1585	-0.1317	-0.1698	-0.1161	-0.1501	-0.1660	-0.1504	-5	-5	1	-6	-5	-7
PC-3	Uri-nary/Male Reproductive	Prostate cancer	Metastatic poorly differentiated prostate adenocarcinoma cell line	-0.2946	-0.1897	-0.1514	-0.2040	-0.1380	-0.1743	-0.1881	-0.1727	-3	-10	2	-6	8	-12
placenta	Primary Tissue	Placenta	Primary tissue	-0.2663	-0.1661	-0.1342	-0.1683	-0.1255	-0.1515	-0.1658	-0.1498	3	10	-7	1	9	6
prostate	Primary Tissue	Prostate	Primary tissue	-0.2480	-0.1608	-0.1329	-0.1752	-0.1225	-0.1505	-0.1648	-0.1510	-8	-10	-11	-8	1	-12
rectum	Primary Tissue	Rectum	Primary tissue	-0.2545	-0.1536	-0.1166	-0.1588	-0.1020	-0.1361	-0.1613	-0.1421	6	0	5	4	-2	0
REH	Lymphoid	Leukemia (ALL)	“Pre-B cell leukemia cell line (ALL, first relapse)”	-0.3299	-0.2161	-0.1655	-0.2137	-0.1501	-0.1824	-0.2035	-0.1738	-2	-5	-5	-2	-4	1
RH-30	Sarcoma	Rhabdomyosar-coma	Metastatic rhab-domyosarcoma cell line	-0.2798	-0.1651	-0.1371	-0.1944	-0.1252	-0.1575	-0.1749	-0.1578	2	-14	-3	-7	-7	-7
RPMI-8226	Lymphoid	Multiple Myeloma	Multiple myeloma cell line	-0.3220	-0.2075	-0.1551	-0.1975	-0.1381	-0.1692	-0.1897	-0.1637	1	16	11	19	14	22
RT4	Uri-nary/Male Reproductive	Bladder cancer	Urinary bladder transitional cell carcinoma cell line	-0.2817	-0.1677	-0.1450	-0.1923	-0.1355	-0.1705	-0.1907	-0.1706	-5	-1	-12	-16	-19	-25
salivary gland	Primary Tissue	Salivary gland	Primary tissue	-0.2623	-0.1661	-0.1378	-0.1765	-0.1278	-0.1542	-0.1725	-0.1551	-7	2	-9	-2	-5	-5
SCLC-21H	Lung	Small cell lung carcinoma	Small cell lung carcinoma cell line	-0.2588	-0.1599	-0.1378	-0.2008	-0.1229	-0.1569	-0.1722	-0.1464	-11	-31	-5	-12	-10	8
SH-SY5Y	Brain	Neuroblastoma	“Metastatic neuroblastoma, clonal subline of neuroepithelioma cell line SK-N-SH”	-0.2714	-0.1701	-0.1369	-0.2011	-0.1241	-0.1574	-0.1705	-0.1508	2	-25	2	-5	1	6

Continued on Next Page…

Table 4.3 – Continued

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
SiHa	Breast/Female Reproductive	Cervical cancer	"Cervical squamous cell carcinoma cell line, integrated 1-2 copies of HPV16"	-0.2884	-0.1804	-0.1479	-0.2005	-0.1395	-0.1760	-0.1928	-0.1746	-2	-7	-15	-19	-11	-27
SK-BR-3	Breast/Female Reproductive	Breast cancer	Metastatic breast adenocarcinoma cell line	-0.2877	-0.1761	-0.1479	-0.1951	-0.1404	-0.1760	-0.1914	-0.1690	-3	-4	-21	-22	-12	-11
SK-MEL-30	Primary Tissue	Melanoma	Metastatic malignant melanoma cell line	-0.3014	-0.1873	-0.1544	-0.2080	-0.1407	-0.1739	-0.1929	-0.1709	-2	-12	-8	-3	-1	-6
skeletal muscle	Primary Tissue	Skeletal muscle	Primary tissue	-0.2609	-0.1659	-0.1340	-0.1792	-0.1246	-0.1501	-0.1644	-0.1452	-1	-7	-7	0	9	11
skin	Skin	Skin	Primary tissue	-0.2589	-0.1661	-0.1340	-0.1681	-0.1270	-0.1479	-0.1672	-0.1510	-4	8	-14	5	-1	-4
small intestine	Primary Tissue	Small intestine	Primary tissue	-0.2603	-0.1643	-0.1212	-0.1579	-0.1074	-0.1410	-0.1657	-0.1424	9	10	11	9	0	7
smooth muscle	Primary Tissue	Smooth muscle	Primary tissue	-0.2588	-0.1580	-0.1273	-0.1693	-0.1126	-0.1437	-0.1609	-0.1489	2	-6	3	4	4	-5
spleen	Primary Tissue	Spleen	Primary tissue	-0.3078	-0.2017	-0.1484	-0.1803	-0.1305	-0.1553	-0.1766	-0.1538	7	27	15	25	20	25
stomach	Primary Tissue	Stomach	Primary tissue	-0.2638	-0.1701	-0.1315	-0.1702	-0.1175	-0.1494	-0.1691	-0.1507	6	3	9	6	0	2
testis	Primary Tissue	Testis	Primary tissue	-0.2154	-0.1422	-0.1092	-0.1471	-0.0931	-0.1263	-0.1326	-0.1235	0	0	0	0	0	0
THP-1	Myeloid	Monocytic leukemia	Acute monocytic leukemia (AML) cell line	-0.3378	-0.2181	-0.1680	-0.2056	-0.1487	-0.1821	-0.2039	-0.1761	-1	8	-1	1	-3	0
thyroid gland	Primary Tissue	Thyroid gland	Primary tissue	-0.2614	-0.1582	-0.1357	-0.1779	-0.1207	-0.1525	-0.1700	-0.1613	-2	-7	-2	-6	-6	-19
TIME	Other	Microvascular endothelial cell line	Telomerase-immortalized human microvascular endothelial cells (pooled)	-0.2955	-0.1803	-0.1473	-0.1982	-0.1337	-0.1697	-0.1859	-0.1702	5	-3	3	-1	3	-11
tonsil	Primary Tissue	Tonsil	Primary tissue	-0.2820	-0.1793	-0.1410	-0.1692	-0.1246	-0.1473	-0.1731	-0.1518	-1	20	8	23	4	9

Continued on Next Page...

Table 4.3 – Continued

Name	Category	Type	Description	Correlations								Rank changes					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	healthy	IC15	IC20	IC17	IC37	IC35
U-138 MG	Brain	Glioblastoma	Glioblastoma cell line	-0.2876	-0.1771	-0.1444	-0.1915	-0.1263	-0.1620	-0.1775	-0.1612	1	8	7	0	2	2
U-2 OS	Sarcoma	Osteosarcoma	Osteosarcoma cell line	-0.2747	-0.1746	-0.1394	-0.1919	-0.1338	-0.1587	-0.1704	-0.1596	-2	0	-11	-3	6	-3
U-2197	Sarcoma	Sarcoma	Malignant fibrous histiocytoma cell line	-0.2902	-0.1808	-0.1458	-0.1949	-0.1286	-0.1639	-0.1798	-0.1650	2	1	5	3	4	0
U-251 MG	Brain	Glioblastoma	Glioblastoma cell line	-0.2923	-0.1782	-0.1396	-0.1966	-0.1247	-0.1596	-0.1773	-0.1653	9	-6	11	4	4	-4
U-266/70	Lymphoid	Multiple Myeloma	"Multiple myeloma cell line (1970, IL-6-dependent)"	-0.3196	-0.2066	-0.1572	-0.2022	-0.1353	-0.1701	-0.1906	-0.1655	-1	4	19	15	12	17
U-266/84	Lymphoid	Multiple Myeloma	"Multiple myeloma cell line (1984, in vitro differentiated)"	-0.3263	-0.2119	-0.1621	-0.2073	-0.1394	-0.1753	-0.1940	-0.1691	-1	2	11	8	10	14
U-698	Lymphoid	B-cell lymphoma	B-cell lymphoma cell line (lymphoblastic lymphosarcoma)	-0.3279	-0.2121	-0.1590	-0.2029	-0.1366	-0.1699	-0.1942	-0.1657	2	5	18	20	6	20
U-87 MG	Brain	"Glioblastoma, astrocytoma"	"Glioblastoma, astrocytoma cell line"	-0.2853	-0.1745	-0.1425	-0.1924	-0.1267	-0.1599	-0.1738	-0.1620	1	0	2	-2	2	-4
U-937	Myeloid	Myelomono-cytic histiocytic lymphoma	Myelomonocytic histiocytic lymphoma cell line	-0.3462	-0.2237	-0.1670	-0.2014	-0.1462	-0.1797	-0.1989	-0.1730	1	18	3	5	2	6
urinary bladder	Primary Tissue	Urinary bladder	Primary tissue	-0.2595	-0.1582	-0.1298	-0.1646	-0.1175	-0.1456	-0.1643	-0.1500	3	5	-2	1	3	-6
WM-115	Skin	Melanoma	Malignant melanoma cell line	-0.2843	-0.1755	-0.1437	-0.1932	-0.1299	-0.1602	-0.1778	-0.1567	-1	-4	-4	-3	-3	2

Table 4.4: **Clinical diagnoses and cfDNA yield for cancer panel samples**. Shown are the clinical and histological diagnoses for 48 patients from whom plasma-borne cfDNA was screened for evidence of high tumor burden, along with total cfDNA yield from 1.0 ml of plasma from each individual and relevant clinical covariates. Of these 48, 44 passed QC and had sufficient material. Of these 44, five were selected for deeper sequencing. cfDNA yield was determined by Qubit Fluorometer 2.0 (Life Technologies).
*: sample was selected for additional sequencing.
**: only 0.5 ml of plasma was available for this sample.
***: sample failed QC and was not used for further analysis.

Sample	Clinical Dx	Stage	cfDNA yield (ng/ml)	Sex
IC01 ***	Kidney cancer (Transitional cell)	IV	242	F
IC02	Ovarian cancer (undefined)	IV	22.5	F
IC03	Skin cancer (Melanoma)	IV	12.0	M
IC04	Breast cancer (Invasive/infiltrating ductal)	IV	12.6	F
IC05	Lung cancer (Adenocarcinoma)	IV	5.4	M
IC06	Lung cancer (Mesothelioma)	IV	11.4	M
IC07 ***	Gastric cancer (undefined)	IV	52.2	M
IC08	Uterine cancer (undefined)	IV	15.0	F
IC09	Ovarian cancer (serous tumors)	IV	8.4	F
IC10	Lung cancer (adenocarcinoma)	IV	11.4	F
IC11	Colorectal cancer (undefined)	IV	11.4	M
IC12	Breast cancer (Invasive/infiltrating lobular)	IV	12.0	F
IC13	Prostate cancer (undefined)	IV	12.3	M
IC14	Head and neck cancer (undefined)	IV	27.0	M
IC15 *	Lung cancer (Small cell)	IV	22.5	M
IC16	Bladder cancer (undefined)	IV	14.1	M
IC17 *	Liver cancer (Hepatocellular carcinoma)	IV	39.0	M
IC18	Kidney cancer (Clear cell)	IV	10.5	F
IC19	Testicular cancer (Seminomatous)	IV	9.6	M
IC20 *	Lung cancer (Squamous cell carcinoma)	IV	21.9	M

Continued on Next Page...

Table 4.4 – Continued

Sample	Clinical Dx	Stage	cfDNA yield (ng/ml)	Sex
IC21	Pancreatic cancer (Ductal adenocarcinoma)	IV	35.4	M
IC22	Lung cancer (Adenocarcinoma)	IV	11.4	F
IC23	Liver cancer (Hepatocellular carcinoma)	IV	17.1	M
IC24	Pancreatic cancer (Ductal adenocarcinoma)	IV	37.2	M
IC25	Pancreatic cancer (Ductal adenocarcinoma)	IV	27.9	M
IC26	Prostate cancer (Adenocarcinoma)	IV	24.6	M
IC27	Uterine cancer (undefined)	IV	19.2	F
IC28	Lung cancer (Squamous cell carcinoma)	IV	33.3	M
IC29	Head and neck cancer (undefined)	IV	14.4	M
IC30	Esophageal cancer (undefined)	IV	10.5	M
IC31 ***	Ovarian cancer (undefined)	IV	334.8	F
IC32	Lung cancer (Small cell)	IV	9.6	F
IC33	Colorectal cancer (Adenocarcinoma)	IV	13.8	M
IC34	Breast cancer (Invasive/infiltrating lobular)	IV	33.6	F
IC35 *	Breast cancer (Ductal carcinoma in situ)	IV	16.2	F
IC36	Liver cancer (undefined)	IV	26.4	M
IC37 *	Colorectal cancer (Adenocarcinoma)	IV	15.9	F
IC38	Bladder cancer (undefined)	IV	6.6	M
IC39	Kidney cancer (undefined)	IV	39.0	M
IC40	Prostate cancer (Adenocarcinoma)	IV	13.8	M
IC41	Testicular cancer (Seminomatous)	IV	16.5	M
IC42	Lung cancer (Adenocarcinoma)	IV	11.4	F
IC43	Skin cancer (Melanoma)	IV	21.9	F
IC44	Esophageal cancer (undefined)	IV	25.8	F
IC45 ***	Colorectal cancer (Adenocarcinoma)	IV	3.0	M
IC46 **	Breast cancer (Ductal carcinoma in situ)	IV	36.6	F
IC47	Pancreatic cancer (Ductal adenocarcinoma)	IV	19.2	F
IC48 **	Breast cancer (Invasive/infiltrating lobular)	IV	13.8	F

Table 4.5: **Sequencing statistics for additional samples included in CA01**. For each sample, sequencing-related statistics, including the total number of fragments sequenced, read lengths, the percentage of such fragments aligning to the reference with and without a mapping quality threshold, mean coverage, duplication rate, and the proportion of sequenced fragments in two length bins, are tabulated. Fragment length is inferred from alignment of paired-end reads. Due to the short read lengths, coverage is calculated by assuming the entire fragment had been read. The estimated number of duplicate fragments is based on fragment endpoints, which may overestimate the true duplication rate in the presence of highly stereotyped cleavage.
SSP, single-stranded library preparation protocol. DSP, double-stranded (conventional) library preparation protocol.
*Sample has been previously published (Kitzman et al., 2012).

Sample	Library type	Reads	Fragments sequenced	Aligned	Aligned Q30	Coverage	% Duplicates	35-80 bp	120-180 bp
IH03	SSP	2x39	53292855	92.66%	72.37%	2.29	15.46%	11.05%	52.34%
IP01 *	DSP	2x101, 2x102	1214536629	97.22%	86.38%	76.11	0.55%	0.08%	62.77%
IP02 *	DSP	2x101, 2x102	855040273	97.16%	87.72%	52.46	0.83%	0.07%	68.10%
IA01	SSP	2x39	53934607	87.42%	68.30%	2.02	22.70%	15.20%	49.77%
IA02	SSP	2x39	42496222	95.42%	76.61%	1.95	4.74%	12.28%	59.00%
IA03	SSP	2x39	51278489	93.12%	71.33%	2.05	25.68%	14.27%	52.57%
IA04	SSP	2x39	50768476	90.30%	70.51%	2.14	7.83%	17.80%	36.76%
IA05	DSP	2x101	194985271	98.80%	90.61%	11.09	12.05%	2.24%	71.67%
IA06	DSP	2x101	171670054	98.90%	90.88%	9.90	5.41%	1.93%	71.26%
IA07	DSP	2x101	208609489	98.67%	90.34%	11.69	11.45%	2.59%	74.84%
IA08	DSP	2x101	193729556	98.81%	90.70%	10.84	11.96%	2.58%	76.24%
IC02	SSP	2x39	57913605	95.07%	75.57%	2.59	5.40%	12.98%	60.00%
IC03	SSP	2x39	63862631	95.78%	75.66%	2.79	8.32%	13.25%	62.20%
IC04	SSP	2x39	55239248	95.47%	76.26%	2.57	8.28%	10.98%	58.48%
IC05	SSP	2x39	39623850	89.80%	69.92%	1.60	9.24%	14.63%	50.33%
IC06	SSP	2x39	59679981	95.57%	74.90%	2.11	3.93%	24.30%	41.46%

Continued on Next Page...

Table 4.5 – Continued

Sample	Library type	Reads	Fragments sequenced	Aligned	Aligned Q30	Coverage	% Duplicates	35-80 bp	120-180 bp
IC08	SSP	2x39	46933688	94.38%	74.21%	1.92	5.92%	16.04%	45.25%
IC09	SSP	2x42	59639583	91.22%	71.15%	2.13	6.69%	21.39%	43.50%
IC10	SSP	2x42	53994406	93.73%	73.40%	1.83	2.00%	27.08%	37.62%
IC11	SSP	2x42	59225460	93.25%	72.51%	2.15	5.26%	21.30%	43.33%
IC12	SSP	2x42	57884742	93.52%	74.33%	2.34	2.66%	18.28%	46.58%
IC13	SSP	2x42	71946779	92.94%	72.47%	2.52	2.18%	23.51%	43.97%
IC14	SSP	2x42	61649203	94.54%	73.47%	2.20	3.23%	22.26%	43.37%
IC15	SSP	2x50, 43/42	908512803	95.49%	76.83%	29.77	10.66%	25.42%	38.47%
IC16	SSP	2x42	62739733	92.81%	72.85%	2.47	2.77%	17.71%	48.04%
IC17	SSP	2x50, 2x39	1072374044	96.02%	76.42%	42.08	12.16%	17.08%	50.02%
IC18	SSP	2x39	59976914	87.91%	68.67%	2.24	4.39%	18.85%	44.44%
IC19	SSP	2x39	51447149	89.38%	69.39%	2.02	8.24%	17.30%	46.33%
IC20	SSP	2x50, 2x39	640838540	96.30%	79.11%	23.38	12.43%	25.72%	39.87%
IC21	SSP	2x39	53000679	94.64%	74.57%	1.79	37.39%	29.89%	43.81%
IC22	SSP	2x39	58102606	94.08%	74.08%	2.51	6.24%	13.65%	58.41%
IC23	SSP	2x39	65859970	95.67%	75.67%	2.94	5.34%	11.09%	60.85%
IC24	SSP	43/42	66344431	94.63%	74.46%	2.48	2.00%	22.46%	46.31%
IC25	SSP	43/42	75066833	93.75%	73.66%	2.86	2.24%	21.30%	46.19%
IC26	SSP	43/42	79180860	92.59%	72.32%	2.97	2.93%	22.34%	40.42%
IC27	SSP	43/42	78037377	88.81%	67.04%	2.20	1.50%	31.31%	30.59%
IC28	SSP	43/42	61402081	95.24%	75.74%	2.60	2.46%	18.71%	46.44%
IC29	SSP	2x39	49989522	94.46%	73.36%	1.75	3.03%	25.82%	36.23%
IC30	SSP	2x39	58439504	93.52%	71.19%	1.75	17.35%	29.58%	30.47%
IC32	SSP	43/42	78233981	87.86%	66.80%	2.25	1.79%	30.12%	31.20%
IC33	SSP	43/42	62196185	87.26%	66.71%	1.93	1.93%	27.44%	36.92%
IC34	SSP	43/42	63572169	95.42%	76.74%	2.53	2.35%	19.64%	48.55%

Continued on Next Page...

Table 4.5 – Continued

Sample	Library type	Reads	Fragments sequenced	Aligned	Aligned Q30	Coverage	% Duplicates	35-80 bp	120-180 bp
IC35	SSP	43/42	618554393	86.47%	65.90%	18.22	5.23%	28.18%	35.24%
IC36	SSP	43/42	54402943	94.62%	74.73%	2.21	3.32%	17.02%	52.42%
IC37	SSP	2x50, 43/42	1175553677	93.00%	74.46%	38.22	10.15%	28.47%	35.11%
IC38	SSP	43/42	47981963	89.35%	69.45%	1.78	6.47%	18.59%	43.03%
IC39	SSP	43/42	61968854	95.29%	75.57%	2.62	2.54%	14.42%	57.28%
IC40	SSP	2x39	53228209	93.54%	71.69%	1.81	8.85%	24.88%	34.95%
IC41	SSP	43/42	78081655	87.11%	65.25%	2.26	1.61%	27.94%	35.21%
IC42	SSP	2x39	53017317	93.59%	74.33%	2.02	10.74%	19.04%	44.12%
IC43	SSP	43/42	76395478	88.41%	67.21%	2.40	1.56%	26.68%	37.76%
IC44	SSP	43/42	61354307	95.15%	74.88%	2.45	4.34%	19.10%	46.39%
IC46	SSP	2x39	60123123	94.51%	72.23%	2.13	10.37%	15.46%	50.93%
IC47	SSP	2x39	59438172	95.58%	73.84%	2.07	9.33%	21.67%	43.34%
IC48	SSP	43/42	55704417	91.35%	72.79%	2.01	13.87%	22.56%	38.68%
IC49	DSP	2x101	170489015	99.02%	90.53%	11.19	5.93%	2.41%	59.93%
IC50	DSP	2x101	203828224	98.72%	90.28%	10.82	2.83%	4.81%	66.23%
IC51	DSP	2x101	200454421	98.63%	90.53%	11.77	9.50%	2.58%	67.04%
IC52	DSP	2x101	186975845	98.97%	91.25%	11.37	2.57%	0.83%	68.96%

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter, I present thoughts on the directions that work in cfDNA will take in the coming years, including opportunities for increased accuracy, precision, and breadth of screening, and technical and ethical challenges to widespread adoption of this methodology. In particular, I discuss the growing application of cfDNA to questions in prenatal medicine, and the nascent but promising use of the "liquid biopsy" in oncology.

5.1 Noninvasive prenatal screening

Portions of this section have been adapted with changes from: Snyder, MW; Simmons, LE; Kitzman, JO; Santillan, DA; Santillan, MK; Gammill, HS; and Shendure, J. Noninvasive fetal genome sequencing: a primer. *Prenatal Diagnosis* 33(6):547-554 (2013).

The impact of noninvasive prenatal screening for aneuploidy on prenatal medicine is hard to overstate. By some reckoning, the testing framework described in Chapter 3 is the fastest growing molecular medical test ever (Chitty and Bianchi, 2015). First aimed at high-risk pregnancies, this class of test is increasingly offered to average-risk pregnancies, both in the US and internationally. Methodological improvements and technological advances in DNA sequencing have boosted test performance metrics to levels that rival those of gold standard, invasive tests like amniocentesis. How much further room for improvement remains?

Future work will likely be geared at increasing the resolution of the tests through the detection of sub-chromosomal genetic insults. Already, some testing methodologies are able to reliably detect specific microdeletions and microduplications linked to a number of diseases, including the 22q11.2 deletion implicated in DiGeorge syndrome. Despite the demonstrated technical possibility of analyzing the fetal genome at the single nucleotide

level, prenatal screening in the near future is likely to remain limited to a discrete number of disorders and diseases that have known molecular causes and are highly penetrant. An area of potential expansion and future research may be the set of such disorders for which treatment is possible and for which *in utero* detection may be desirable.

One possible candidate for this type of prenatal screening is the metabolic testing of newborns for a small panel of metabolic disorders, such as PKU, for which interventions exist. In states where this testing is performed, blood is typically collected from newborns within several days of birth, allowing for appropriate interventions – where applicable – reasonably quickly. However, in principle such testing could be performed prenatally by combining parental haplotypes with targeted cfDNA sequencing of a panel of approximately 50 genes[1], allowing for dietary supplementation or other treatments to begin immediately upon delivery. Prenatal detection of one of these disorders would potentially allow appropriate specialists to be available at delivery, and enable counseling of parents at an early stage, potentially improving outcomes and reducing anxiety immediately following delivery.

Undoubtedly, as the scope of prenatal screening expands, the rates and causes of false-positive tests will change. Continuing optimization of test performance will remain critical to reducing patient anxiety and minimizing the burden of invasive, diagnostic followup. Remarkably, the work presented in Chapter 3 has already borne fruit. In response to the publication and public dissemination of these results, Illumina – the highest volume testing provider as of this writing – performed a retrospective analysis of 11 false positive cases in a cohort of 1,914 tested pregnancies. They identified large maternal CNVs in at least three of these false positive cases and several others in which the presence of CNVs was uncertain (Chudova et al., 2015) . They further described a change to their data analysis pipeline to account for the presence of maternal CNVs. A particularly encouraging aspect of these changes is the quick timeframe in which they occurred, suggesting that the barriers

[1]In the United States, the exact number of genes examined would vary from state to state, depending on the composition of the newborn screening panel.

to clinical translation of at least some academic findings may be diminished.

5.1.1 Discovery of *de novo* mutations

Despite the improvements in resolution and accuracy in noninvasive prenatal screening for disease, the robust detection of *de novo* mutations, important contributors to the burden of genetic disease in newborns (Veltman and Brunner, 2012), remains an elusive goal. The detection of inherited variation, which benefits greatly from the availability of parental haplotypes (as discussed in Chapter 2), has proved a tractable problem, with several groups reporting high accuracy exome- or genome-wide (Chen et al., 2013; Fan et al., 2012; Kitzman et al., 2012) in a small number of examined pregnancies. However, the discovery of the small number of *de novo* variants expected in each fetal genome – mutations newly arising in the maternal or paternal germline – is a much tougher nut to crack, and remains one of the major open challenges for noninvasive prenatal testing.

In principle, *de novo* mutations are easily identified as variants in the sequenced maternal cfDNA that are not found in either parent. In practice, despite ongoing improvement, DNA sequencing technology remains imperfect, and errors introduced during PCR or sequencing far outnumber the approximately 50 to 100 true *de novo* mutations expected in any given fetus (Kong et al., 2012). At a sequencing depth of 100X and fetal fraction of 10%, the two types of events yield signatures that are, on the whole, nearly indistinguishable: at a given site, a small handful of reads suggests the spontaneous emergence of a fetal genotype incompatible with Mendelian inheritance. Separating the true mutations from the spurious errors introduced during the sequencing process remains a challenge and a major area for improvement in both technology and analysis.

One way to address the large number of candidate *de novo* mutations is to apply an increasingly aggressive set of filters designed to improve the signal-to-noise ratio in the candidate set, as described in Chapter 2. For example, we might exclude any candidate with only one or two supporting reads. We might remove sites that are inside or adjacent to specific sequence motifs known to generate elevated error rates. We might discard any

site also identified as a candidate in other samples within the same cohort. At each step, we may trade a small decrease in sensitivity for a suitably large gain in specificity. Even after extensive filtering, we are likely to be left with several thousand candidates – still too many for follow-up. However, only a very small percentage of these candidates are likely to fall within protein coding or known regulatory regions, suggesting that manual review and/or validation of known high-impact candidates may be plausible, if undesirable, in a clinical setting.

The emergence in recent years of genome-wide variant prioritization or scoring frameworks such as CADD (Kircher et al., 2014), as well as the increasing number of gene-level variant effect predictions from deep mutational scans, population surveys, and evolutionary models suggests that some level of automation may provide a useful, albeit far from comprehensive, analytical layer for ranking candidates for followup. Nevertheless, the molecular causes of thousands of additional Mendelian disorders, contributing to the approximately 20% of infant mortality caused by single-gene disorders (Bell et al., 2011; Saunders et al., 2012), remain undiscovered, and the extent to which automated or manual variant prioritization schemes highly rank the unknown, causal alleles involved in these disorders remains to be determined. Despite the increasing catalogs of known molecular causes of disease (Amberger et al., 2009), and a growing number of saturation mutagenesis experiments densely investigating the impact of single nucleotide changes in gene bodies and regulatory regions, the functional impact of most potential and extant human variation remains unknown, representing a major challenge to the interpretation of candidate or validated *de novo* mutations. Furthermore, the degree to which current experimental systems for high-throughput functional profiling of many mutations reflect actual phenotypic consequences at the level of the developing fetus or newborn is not yet clear, further clouding interpretation. Even in the context of neonatal sequencing, when genetic material from the newborn can be directly obtained and sequenced without confounding by maternal background, whole-genome sequencing of cases of severe disease has had mixed success (Kingsmore et al., 2015; Saunders et al., 2012; Willig et al., 2015), albeit with undeniable

benefits for some individuals.

Clinically, diagnostic or validation followup work is expected to remain substantial even as sequencing error rates continue to improve, arguing that a different paradigm for detection of *de novo* variation is desirable. What might this look like in practice? One possibility lies in the isolation of intact fetal cells in the maternal circulation. While such cells are extremely rare, technologies for the enrichment of such cell populations have improved dramatically in recent years (de Bourcy et al., 2014). After isolation of fetal cells, the genetic material of each cell could be amplified and sequenced, providing a direct readout of the fetal genome with no maternal background. Assuming sufficient numbers of fetal cells are available, variant predictions could be made by statistically or heuristically combining evidence across cells to reduce errors inevitably arising from the amplification of picograms of DNA (Gawad et al., 2016). This strategy has the additional advantage of working around the potential confounding effect of confined placental mosaicism by directly interrogating fetal, rather than placental, genetic material. However, even this type of technology is likely to struggle with specific classes of variation, in particular those variant types that bear mutational signatures similar to technical artifacts arising from single-cell amplification methodologies[2], or variants that are truly mosaic in the developing fetus itself.

Ideally, in order to systematically map *de novo* mutations, a sample must be collected from the father. Without knowledge of the paternal genotypes, any paternally transmitted alleles not shared with the mother are indistinguishable from *de novo* mutations in the maternal germline. However, even without a paternal sample, it may still be possible to identify likely *de novo* mutations by searching a predefined panel of genes known to be inherited in a dominant fashion with high penetrance; mutations in these genes could be ruled as unlikely given the father's health status. Nevertheless, for all but the most stereotyped disorders, definitively separating deleterious mutations from benign ones remains

[2]For current single-cell amplification methods including MDA or MALBAC, such artifacts tend to look like copy-number gains or losses, owing to uneven amplification of scant input material. Allelic dropout is a more common error modality than is the spurious introduction of new alleles, suggesting that at the single-nucleotide level, such methods are likely to be more challenged by imperfect sensitivity than by low specificity.

an elusive goal, even for single-gene disorders.

5.1.2 Translational challenges

Although high-resolution, prenatal cfDNA sequencing can yield an accurate picture of the fetal genome, substantial technical and logistical several challenges must be addressed and avenues for improvement explored before this technology can reach the bedside. One major hurdle facing care providers is establishing an informatics infrastructure to process and securely store large volumes of genomic data. Interpreting these data poses an even greater challenge: WGS provides measurements across over 20,000 protein-coding genes that are not readily summarized as a single result. The measurements themselves are complex: WGS reports an entire set of genotypes, far from providing a numerical read-out as analyte testing might, or a "normal/abnormal" status as trisomy screening would provide. While the report is comprehensive in breadth, most of the reported variants have little to no impact on patient health, placing the burden on the physician and genetic counselor to isolate the relevant information (if any) from that volume of data. As discussed above in the context of *de novo* mutations, automated analyses might been applied in the context of neonatal sequencing to select or prioritize only genetic variants in genes deemed relevant in order to streamline the process and to exclude incidental findings (Bell et al., 2011). Additionally, the analytic method described in Chapter 2 focuses primarily on single-nucleotide variants (which account for the majority of human genetic variation by quantity but not necessarily by proportion of the genome impacted). In order to be truly complete, it is necessary to also consider other variants including short insertions and deletions, structural variation such as inversions or translocations, and copy number gains or losses. While necessary, these analyses will further complicate interpretation.

As discussed before, current approaches to prenatal diagnosis incorporate increasingly refined noninvasive screening techniques to identify pregnancies at high risk of fetal abnormalities, thus facilitating direction of invasive diagnostic approaches to a small number of pregnancies (ACOG Committee on Practice Bulletins, 2007). Ultimately, diagnostic ap-

proaches that are both noninvasive and comprehensive would replace screening altogether. Currently, though noninvasive approaches for detection of specific aneuploidies are commercially available, the test performance characteristics for these approaches drive consideration of these tests as sensitive screening tests with persistent reliance on invasive testing for diagnostic confirmation (American College of Obstetricians and Gynecologists Committee on Genetics, 2012). NIFWGS, with its extremely high sensitivity, has the potential to achieve the goal of noninvasive, broad diagnostic capability. However, in the context of prenatal diagnosis, in order to achieve this potential, we must keep in mind the need for absolute minimization of false positive results, matching or surpassing the accuracy of invasive testing (>99% in the case of amniocentesis (American College of Obstetricians and Gynecologists, 2007)). Once technical aspects of the procedure are refined, scalability to larger validation studies carefully evaluating such test performance characteristics will be the essential next step.

One factor that must be considered in test performance evaluation and translation of high-resolution prenatal screening to clinical practice is the placental origin of fetal cfDNA. As with chorionic villus sampling (CVS), which also samples placental material, confined placental mosaicism (CPM) must be considered in interpretation of genetic results derived from fetal cfDNA (Kalousek and Dill, 1983). Empiric evidence supporting the relevance of CPM to fetal cfDNA was recently described in a case report (Choi et al., 2012). In practice, depending on the overall clinical picture, abnormal CVS results may require direct confirmatory testing of fetal cells through amniocentesis. Estimates of the incidence of CPM vary depending on preparation technique – whether performed directly or after culture – but generally range between 1-2% (Hahnemann and Vejerslev, 1997). These estimates derive primarily from first trimester samples evaluated for aneuploidy. Though CVS is typically performed in early pregnancy, some studies of CVS in the second and third trimesters have found an increased incidence of CPM with increasing gestational age (Carroll et al., 1999), a factor to consider in estimation of the effect in noninvasive screens. CPM can result from a postyzygotic event generating genetic error in an initially normal pregnancy, or placental

genetic rescue (e.g. trisomic rescue) in an initially abnormal pregnancy, which can result in fetal uniparental disomy (UPD) or segmental UPD (Engel, 2006). It is important to consider that evaluation of CPM or UPD has typically focused on karyotypic analyses. Studies utilizing genome-wide or array-based approaches suggest greater detection of abnormalities through these techniques (Filges et al., 2011; Inbar-Feigenberg et al., 2012), supporting the possibility that CPM for subchromosomal changes across the genome would be expected to occur more frequently than CPM for aneuploidy.

Sorting out the effect of CPM on diagnostic performance of high-resolution, noninvasive fetal testing is complicated further by our increasing understanding of genetic diversity and even genetic flexibility within an individual. Throughout the field of genetics, technological advances are providing glimpses into the nonabsoluteness of genetic categorization (Engel, 2006). In fact, "CPM" may not be actually confined to the placenta in a substantial proportion of cases, with true fetal mosaicism occurring more commonly than previously understood and with varying phenotypic manifestations (Stetten et al., 2004). While genetic flexibility in disease – for example, loss of heterozygosity at HLA loci to evade immune detection in cancer (Vago et al., 2009) -– has been known for some time, it is becoming increasingly clear that it is also present in health (e.g., somatic revertant mosaicism (Choate et al., 2010)) and development (Engel, 2006). Genetic diversity within an individual, through mosaicism or microchimerism (Nelson, 2012) may in fact be the rule rather than the exception. Understanding the impacts of CPM and true fetal mosaicism at the level of whole genome sequencing is an entirely new area and an essential component of bringing such testing to clinical practice.

5.1.3 Implications for patients and genetic counselors

The complexity and ambiguity of high-resolution results must be considered as the field of noninvasive prenatal diagnostics moves forward, and communication of the resultant ambiguity to patients must be a priority. From a practical perspective, it is clear that as the field advances, there will be an increasing need for subspecialized genetic counseling, as it

is unlikely that these discussions could reasonably be incorporated into busy obstetric or perinatal practices. Though a comprehensive discussion of the ethical implications of such testing is outside the scope of this dissertation, the intersection of these issues with prenatal decision making for families mandates careful consideration of how best to incorporate this technology into practice (Tabor et al., 2012).

Initial targeting of high-resolution testing to specific patient populations will likely include those with current or prior otherwise unexplainable fetal abnormalities or losses, providing a framework for appropriate counseling. Ultimately, after larger studies of test performance, couples at risk for genetic disorders on the basis of race/ethnicity or family history may benefit in the intermediate term. With increasing public awareness and accessibility of commercial genetic screening opportunities, genetic information obtained in other contexts may drive particular patient populations to seek prenatal whole-genome or whole-exome testing, again providing a natural starting point for genetic counseling. In the long term, after significant study in larger cohorts, utilization of this approach for widespread screening – akin to or replacing neonatal screening, as discussed above – may enable prenatal provision of information and also facilitating immediate neonatal intervention for specific conditions. However, the future uptake of high-resolution screening by average- or low-risk pregnancies promises to present a particularly difficult challenge for appropriate genetic counseling and informed decision-making.

Overall, clinical translation of techniques to comprehensively assess fetal genetic health from a maternal blood sample has the potential to reshape the future of prenatal diagnosis. Scientific progress in this area has vastly evolved in the last 30 years and continues to accelerate rapidly. Thoughtful shepherding of this technology to the prenatal bedside must include technical refinement, careful evaluation of test performance, appropriate targeting of patient populations, and effective communication in the face of our increasing appreciation of genetic ambiguity.

5.2 Noninvasive monitoring of cancer and other diseases

The analytical and experimental framework developed in Chapter 4 relies on evidence left behind by native enzymatic cleavage of chromatin to infer the types of cells contributing to the circulating DNA. Application of this methodology to cohorts of healthy individuals and individuals with advanced cancer revealed clear differences in contributing cell populations, recovering the tissue-of-origin for a subset of the cancer samples, including a ductal carcinoma and a hepatocellular carcinoma, and confirming the dominant myeloid and lymphoid sources of cfDNA in healthy individuals.

Moving forward, one key avenue for improvement to this framework involves the estimation of the proportions or "burden" of contributing tissues. The question is twofold: first, can we develop a point estimate and confidence interval for the percentage of cfDNA derived each contributing tissue; and second, can we determine whether the estimated proportions in a clinical sample differ in a biologically significant way from expectation? As discussed in Chapter 1, at least two methods exist for tissue deconvolution, the first based on DNA methylation signatures (Sun et al., 2015) and the second on circulating RNA profiles (Koh et al., 2014). In both of these approaches, a quadratic programming framework is employed to allow for quantification of contributing tissues, although the number of tissues that could be quantified simultaneously and the ability to stably quantify cell types with highly correlated profiles are not directly addressed. To increase the interpretability and clinical utility of the method presented in Chapter 4, a similar approach is needed, and work towards that goal is ongoing. However, even once such estimation is achieved, determining the biological relevance of these tissue contributions remains a key question. To fully address this topic, a thorough characterization of the spectrum of contributions in phenotypically healthy individuals will be required, and possible strategies toward this goal are discussed below in section 5.2.4.

Assuming these key questions have been addressed, what conditions might be monitored with these or related methods? A likely candidate is autoimmune disease such as

systemic lupus erythmatosus. Systemic or localized tissue damage or inflammation may give rise to detectable signals in cfDNA, and the relative proportions of contributing tissues might then be quantified to identify response to treatment or to assist in classifying the severity of flares (Tug et al., 2014) where simple measures of total cfDNA concentration have failed (Atamaniuk et al., 2011). Cardiovascular health also shows promise for non-invasive monitoring with cfDNA (Breitbach et al., 2014), both in the context of acute events such as myocardial infarction, where subsequent necrosis of smooth muscle tissue might be measured for prognostic reasons, or in chronic conditions such as coronary artery disease, where individuals might be placed on a spectrum of disease severity. More speculatively, potential applications of this analytical framework to diseases of the brain may prove fruitful. Because of the blood-brain barrier, cfDNA from brain tissues has been difficult to identify, even in late-stage brain malignancies, with few exceptions (Bettegowda et al., 2014; Koh et al., 2014). However, most attempts at identifying brain-derived cfDNA rely on identifying a small number of somatic variants, typically those stereotyped mutations observed repeatedly across panels of brain tumors (Bettegowda et al., 2014). An orthogonal approach for aggregating a genome-wide signal of gene expression, such as the framework described in Chapter 4, may improve sensitivity to identify rare contributions in brain malignancies, although to date no data is available in support of this claim.

Below, I discuss some avenues for technical improvement in determining cfDNA tissues of origin, and in extending this framework beyond cancer to additional diseases and conditions. In particular, I examine ways in which multiple, orthogonal testing methodologies may in fact be complementary, such that precision of inference would benefit from integration of various data types from a single sample. I describe the possible implications of the findings in Chapter 4 for the expected performance of targeted cfDNA assays. Finally, I discuss how improvements to the analytical framework and standardization of sample collection may permit higher accuracy and more reproducible results.

5.2.1 Integrating multiple data types

As discussed in Chapter 1, several methodologies exist for determining the contribution or burden of specific tissues in a biological sample of cfDNA. For example, the relative proportion of paternal-specific alleles to maternal background in the maternal circulation can be quantified with qPCR or sequencing to yield an estimate of the "fetal fraction" during pregnancy. Similarly, tumor-specific aneuploidy and/or loss of heterozygosity, when known, can be tracked over time with assays designed to measure allelic ratios at specific loci.

In general, such methods currently fall into three categories: profiling DNA sequencing variants, quantifying site-specific DNA methylation, and analyzing fragment endpoints. In at least some embodiments, these approaches are complementary, such that clinical tests may benefit from combining multiple, orthogonal lines of evidence. In particular, these various lines of evidence might be gathered in the same test.

For example, a hypothetical testing approach might first involve the bisulfite conversion of cfDNA to globally identify methylated cytosine residues for inference of tissue composition. Bisulfite conversion preserves fragment endpoints, such that the additional signal provided by nucleosome spacing could be used to improve precision, potentially by discriminating between tissues with similar methylation profiles but different gene expression patterns. Copy-number amplification or deletion could be identified with relatively modest sequencing depth from this bisulfite-converted library. Finally, provided deep sequencing is performed, specific SNVs could be quantified. A target enrichment step – for example, as implemented in CAPP-seq (Newman et al., 2014) – would also be compatible, potentially reducing overall sequencing costs.

5.2.2 Implications for targeted analysis

Unbiased, genome-wide investigations of cfDNA – for example, by shotgun sequencing following conventional library preparation – may be appropriate for screening of individuals whose health status is unknown. However, in the context of monitoring a patient for re-

sponse to treatment for a specific disease, targeted investigation of specific loci may represent a more cost-effective and tractable experimental strategy with minimal risk of incidental findings (Newman et al., 2014). This methodology might involve ultra-deep sequencing or quantitative amplification of a small number of known tumor-associated mutations to evaluate tumor burden, or tiling across a small panel of frequently mutated genes to surveil disease recurrence following treatment.

One method for target enrichment using a tiling approach is the capture of relevant regions with single-molecule molecular inversion probes, or smMIPs (Hiatt et al., 2013). In this experimental framework, a single-stranded DNA oligo is designed to have two targeting arms complementary to DNA sequence flanking a target of interest, along with a molecular tag to quantify the number of unique copies of the target that were interrogated. After hybridization of these targeting arms to a template, the targeted region is copied, or "captured," for downstream readout by sequencing. Typically, genomic DNA is used as input to the capture reaction, and its high molecular weight allows over 100 nucleotides of targeted sequence to be captured by a single probe, with typical efficiencies of 1-2% per target. Owing to the high degree of fragmentation observed in cfDNA samples, shorter targets may be indicated to maintain acceptable capture efficiency. Preliminary results suggest that a target length of up to 40 bases yields capture efficiency between 1-2% for a small number of probes. However, the majority of probes fail to achieve a 1% threshold, suggesting that further optimization is needed if this strategy is to gain traction (Figure 5.1).

One potential confounder of these results is that, due to the limited material available from a typical plasma sample and the biased fragmentation patterns observed in plasma-borne cfDNA, the optimal placement of PCR primers or capture baits may play a critical role in determining capture success. Design of such oligos is frequently optimized for some set of experimentally relevant parameters, including GC content and repetitive sequence composition within some target locus. One implication of the findings presented in Chapter 4 is that the specific locations of primers or baits relative to expected nucleosome positions may govern the likelihood of successful capture or amplification of those loci. For example,

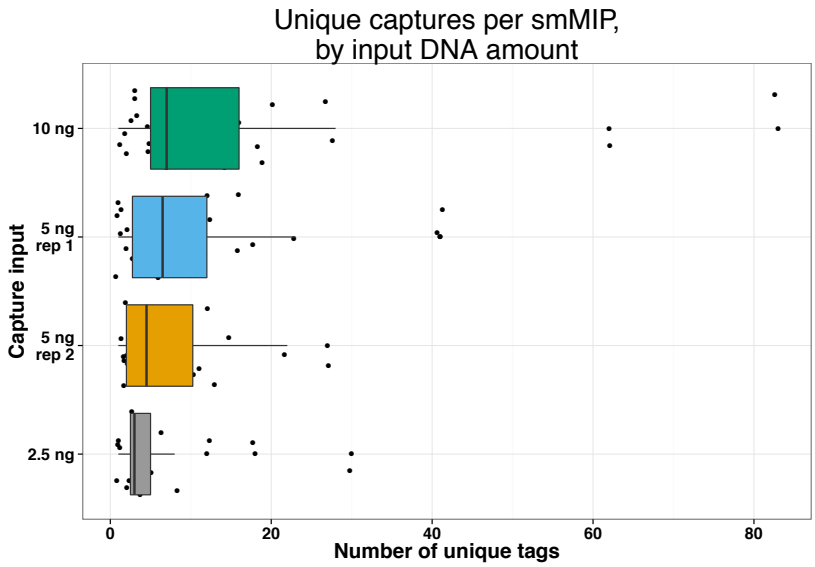


Figure 5.1: **Capture of cfDNA targets with smMIPs, by input**. Cell-free from an anonymous female donor was subjected to capture with single-molecule molecular inversion probes (smMIPs). The four capture reactions contained an identical panel of 32 smMIPs, each designed to capture 40 nucleotides of sequence with targeting arms between 16 and 24 bases in length. The capture reaction was performed as described (O'Roak et al., 2012) with a probe-to-target ratio of 800:1. At least two unique molecular tags were observed for 28 targets (87.5%) in the 10 ng input capture, 26 and 23 targets (81.3% and 71.9%) in the two replicates of the 5 ng input capture, and 26 targets (81.3%) in the 2.5 ng input capture.

primers closely spanning a WPS peak would be expected to result in more efficient amplification than would primers spanning a WPS trough.

Although the extent to which this consideration impacts successful targeting remains to be seen, anecdotal investigation of the success of cfDNA capture with smMIPs suggests that, at least in limited data, successful probe design involves targeting arms placed within WPS peaks (*data not shown*). The correlation between probe placement and capture success rate is not perfect, though, perhaps owing in part to the overall low efficiency of MIP capture obscuring site-to-site variability. Future studies involving multiplex PCR or hybrid capture target enrichment should continue to shed light on this experimental consideration.

5.2.3 Optimizing sample collection and processing

Currently, no universal standards exist for collection of peripheral blood for downstream plasma isolation and cfDNA purification. While some best practices are frequently observed – for example, blood is typically collected in tubes containing EDTA to minimize subsequent degradation of cfDNA fragments – many of the parameters plausibly affecting downstream analysis have yet to be studied. Future work is needed to address the impact of these potential confounders and, where possible, to control for them during sample collection, sample processing, or data analysis.

One key parameter that may impact test results is the time of day at which samples are collected. White blood cell lineages represent the major sources of the "background" in most cfDNA samples, due to frequent apoptosis and renewal of these cell populations. However, evidence suggests that the turnover of these cells may be biased for particular times of the circadian clock. Because of the short half-life of cfDNA fragments in the circulation (Lo et al., 1999), optimal samples may be those collected when white blood cell turnover is at its lowest point, in order to maximize the signal contributed by additional or abnormal cell populations. Even when the time of collection cannot be standardized, understanding the potential impact of this variable on downstream inference and developing strategies to to control for this possible confounder analytically may bear fruit.

Other parameters that may confound or complicate inference include biological considerations such as patient BMI or fitness status, and technical considerations including cfDNA purification methods or sequencing library construction. Patient characteristics, including frequency of strenuous exercise, can influence the concentration and likely the composition of cfDNA in a plasma sample, which may confound inference. Bead-based kits for cfDNA purification, while attractive due to ease of workflow, may preferentially purify longer fragments compared to organic chemistry-based approaches, discarding a potentially informative class of fragments. Conventional library preparation techniques may compound this problem, altering the fragment length distribution and shifting the cleavage periodicity by several base-pairs (Figures 4.2 and 4.4), potentially affecting downstream analysis. Standardizing or controlling for variability in sample processing will remain an important future step as testing expands to additional providers and patient groups.

5.2.4 Comparison to reference cohorts

The method for tissue-of-origin determination described in Chapter 4 relies on comparison of clinical samples to one or more reference maps – specifically, by comparing observed signals of nucleosome spacing across gene bodies to tissue-specific gene expression measurements in a linear correlation framework. The goal of this analytical framework is to identify or rank the closest match or matches between a clinical sample and reference maps in order to identify contributing tissues or cell types.

While this approach represents a reasonable analytical strategy when well-matched reference maps are available, such maps do not exist for every cell type. Even when purportedly matched cell lines do exist, the degree to which they represent a snapshot of a given tumor or other diseased tissue is unclear. For example, evidence from DNase-I hypersensitivity studies suggests that the regulatory landscapes of tumors evolve as cancer progresses, specifically by the reactivation of certain embryonic regulatory regions (Stergachis et al., 2013). If the epigenetic changes are of sufficient magnitude, the best matching reference maps may no longer give clear clues to the anatomical origin of the tumor. Additionally,

collapsing gene expression measurements in a cell line to a single number per gene may obscure true expression heterogeneity *in vivo*, a caveat that may be amplified in the presence of stresses or inflammatory processes plausibly co-occurring with development or progression of disease. If primary tissue samples are instead used to generate reference maps, the maps may again be imperfect owing to the different extents to which each tissue sample is vascularized – or, more generally, to the different proportions of various cell populations present in each sample.

Thus, an alternative strategy based on clustering of patients along phenotypic lines may be preferable. Rather than comparing a clinical sample to reference maps made from primary tissue or cell lines, the comparison could be made to reference cohorts of previously studied samples whose clinical phenotypes include diverse cancer types and other diseases, as well as healthy controls. Familiar distance metrics between samples can be calculated – for example, the Euclidean distance between intensity vectors from Fourier transformation as described in Chapter 4 – to cluster samples on the basis of similarity to one another rather than to reference maps. In essence, the question shifts from *Which reference maps most closely match this sample?* to *Which previous clinical samples most closely match this new sample, and what condition(s) did those samples have in common?*

This strategy, though, is challenged on several fronts. First, it presupposes a broad survey of cancers and other diseases, requiring significant uptake of this screening methodology by diverse patient groups. Second, it requires sufficient numbers of samples within each disease category to evaluate the genetic heterogeneity observed between individuals with similar phenotypes, which for some conditions may plausibly be too great to allow for confident inference in this manner. Third, careful phenotyping of individuals in the reference cohorts is critical. While this requirement is particularly true in the context of disease classification, it is perhaps equally true for healthy or "control" individuals. For example, individuals may present as clinically unaffected, but may in fact harbor low-grade and undiagnosed disease, including diminished cardiovascular health or early malignancies. Careful selection of reference individuals may help reduce the number of incorrect

predictions made on the basis of sample clustering. Finally, the use of reference cohorts argues for some degree of data sharing between testing providers, currently an elusive goal for many genomic analyses. While reference maps – for example, those produced by the Human Protein Atlas, GTEx, or ENCODE – are typically publicly available resources and easily standardized across testing providers, reference patient cohorts are likely to be more restricted. This balkanization could lead to the undesirable outcome of a particular sample receiving two or more different classifications based on testing performed by different providers. A mitigating consideration is that data sharing at the level of FFT intensity vectors may represent less of a privacy concern than at the genotypic level.

5.3 Closing thoughts

There is no doubt that improved noninvasive screening techniques, whether based on cfDNA or other biomarkers, represent an exciting and powerful approach to precision medicine. As the range of conditions that can be monitored this way continues to grow, it is likely – although far from guaranteed – that diagnoses will become more precise, treatments will become more tailored, and health outcomes will improve. However, this technological and medical boon may not be shared equally by all. Will the growth of cfDNA-based screening widen or narrow health disparities? Who benefits?

The expanding reach of precision medicine carries an enormous price tag. These costs are both individual, borne by those wishing to pay for new tests, and collective, in the form of NIH-funded research such as the work presented in this dissertation. A key question touching on political, ethical, and medical concerns is whether the cost of such research is justified in the face of the many more basal and substantial public health challenges that remain, both here and abroad. Might hunger not be alleviated, might the homeless not be housed with the money spent on medical genomic research?

For some such as Daniel Koshland, former editor of *Science* and cheerleader for genomic research (Koshland, 1989), this question is easily answered. In his remarks at the First Human Genome Conference in 1989, he responded to the argument that less indirect

interventions were still needed: “What these people don’t realize is that the homeless are impaired. [...] Indeed, no group will benefit more from the application of human genetics” (Kevles and Hood, 1993).

Now more than 25 years later, it is hard to find evidence to support that prediction, even if we restrict our view to diagnostics and ignore the slow pace of growth of therapeutic interventions. The underlying hypothesis that genetic variation is the major driver of behaviors such as addiction and mental illness has thus far found limited support in studies of large cohorts, with many variants of small effect explaining some of the traits’ heritability without opening up clear avenues for potential intervention. When we consider the entire spectrum of complex human traits, the picture is not much more encouraging. Some of the basic biology of lipid traits and heart health has been explained using genomic approaches, but high cholesterol and cardiovascular disease remain important health concerns in the United States. It may be that Dick Lewontin was, for the most part, correct: the research program of linking genetic variation to deviations from a phenotypic mean reinforces a deterministic view and minimizes the role of environmental interventions in effecting change at the individual level. In 1974, he suggested that we “stop the endless search for better methods of estimating useless quantities. There are plenty of real problems” (Lewontin, 1974). Indeed, many of those real problems persist, 40 years later.

Nevertheless, noninvasive screening and testing in pregnancy represents one of the real success stories for public health in the genomic era, reducing risk to patients and enabling appropriate support for children with genetic disease to begin at birth. While the challenge of predicting the entire complex trait makeup of a child from noninvasive sequencing may be hopeless, the comprehensive assessment of risk for thousands of severe diseases and disorders is now within reach. The impact of this category of testing in oncology is still preliminary, but there is no doubt that at least some patients will be positively affected by therapeutic choices guided by cfDNA-based profiling. Similar promise exists for screening and testing for an variety of other conditions and disease states, with much work yet to be done. Critically, it is up to us to ensure that the benefits of these public health advances are

shared by all.

BIBLIOGRAPHY

ACOG Committee on Practice Bulletins, . ACOG Practice Bulletin No. 77: screening for fetal chromosomal abnormalities., January 2007.

Adey, Andrew; Morrison, Hilary G; Asan, ; Xun, Xu; Kitzman, Jacob O; Turner, Emily H; Stackhouse, Bethany; Mackenzie, Alexandra P; Caruccio, Nicholas C; Zhang, Xiuqing, and Shendure, Jay. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, 11(12):R119, 2010.

Ajay, Subramanian S; Parker, Stephen C J; Abaan, Hatice Ozel; Fajardo, Karin V Fuentes, and Margulies, Elliott H. Accurate and comprehensive sequencing of personal genomes. *Genome Research*, 21(9):1498–1505, September 2011.

Alberry, M; Maddocks, D; Jones, M; Abdel Hadi, M; Abdel-Fattah, S; Avent, N, and Soothill, P W. Free fetal DNA in maternal plasma in anembryonic pregnancies: confirmation that the origin is the trophoblast. *Prenatal Diagnosis*, 27(5): 415–418, May 2007.

Amberger, Joanna; Bocchini, Carol A; Scott, Alan F, and Hamosh, Ada. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research*, 37 (Database issue):D793–6, January 2009.

American College of Obstetricians and Gynecologists, . ACOG Practice Bulletin No. 88, December 2007. Invasive prenatal testing for aneuploidy., December 2007.

American College of Obstetricians and Gynecologists Committee on Genetics, . Committee Opinion No. 545: Noninvasive prenatal testing for fetal aneuploidy. *Obstetrics and gynecology*, 120(6):1532–1534, December 2012.

Andersson, Robin; Enroth, Stefan; Rada-Iglesias, Alvaro; Wadelius, Claes, and Komorowski, Jan. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, 19(10):1732–1741, October 2009.

Angert, Robert M; Leshane, Erik S; Yarnell, Ralph W; Johnson, Kirby L, and Bianchi,

Diana W. Cell-free fetal DNA in the cerebrospinal fluid of women during the peripartum period. *American journal of obstetrics and gynecology*, 190(4):1087–1090, April 2004.

Ashoor, G; Syngelaki, A; Wagner, M; Birdir, C, and Nicolaides, K H. Reports of Major Impact. *American journal of obstetrics and gynecology*, 206(4):322.e1–322.e5, April 2012.

Atamaniuk, Johanna; Hsiao, Yu-Yang; Mustak, Monika; Bernhard, Duhm; Erlacher, Ludwig; Fodinger, Manuela; Tiran, Beate, and Stuhlmeier, Karl M. Analysing cell-free plasma DNA and SLE disease activity. *European journal of clinical investigation*, 41(6):579–583, June 2011.

Bansal, V and Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159, August 2008.

Bell, C J; Dinwiddie, D L; Miller, N A; Hateley, S L; Ganusova, E E; Mudge, J; Langley, R J; Zhang, L; Lee, C C; Schilkey, F D; Sheth, V; Woodward, J E; Peckham, H E; Schroth, G P; Kim, R W, and Kingsmore, S F. Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. *Science translational medicine*, 3(65):65ra4–65ra4, January 2011.

Bettegowda, C; Sausen, M; Leary, R J; Kinde, I; Wang, Y; Agrawal, N; Bartlett, B R; Wang, H; Luber, B; Alani, R M; Antonarakis, E S; Azad, N S; Bardelli, A; Brem, H; Cameron, J L; Lee, C C; Fecher, L A; Gallia, G L; Gibbs, P; Le, D; Giuntoli, R L; Goggins, M; Hogarty, M D; Holdhoff, M; Hong, S M; Jiao, Y; Juhl, H H; Kim, J J; Siravegna, G; Laheru, D A; Lauricella, C; Lim, M; Lipson, E J; Marie, S K N; Netto, G J; Oliner, K S; Olivi, A; Olsson, L; Riggins, G J; Sartore-Bianchi, A; Schmidt, K; Shih, l M; Oba-Shinjo, S M; Siena, S; Theodorescu, D; Tie, J; Harkins, T T; Veronese, S; Wang, T L; Weingart, J D; Wolfgang, C L; Wood, L D; Xing, D; Hruban, R H; Wu, J; Allen, P J; Schmidt, C M; Choti, M A; Velculescu, V E; Kinzler, K W; Vogelstein, B; Papadopoulos, N, and Diaz, L A. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Science translational medicine*, 6(224):224ra24–224ra24, February 2014.

Bianchi, Diana W; Parker, R Lamar; Wentworth, Jeffrey; Madankumar, Rajeevi; Saffer, Craig; Das, Anita F; Craig, Joseph A; Chudova, Darya I; Devers, Patricia L; Jones, Keith W; Oliver, Kelly; Rava, Richard P, and Sehnert, Amy J. DNA Sequencing versus Standard Prenatal Aneuploidy Screening. *The New England journal of medicine*, 370(9): 799–808, February 2014.

Botezatu, I; Serdyuk, O; Potapova, G; Shelepov, V; Alechina, R; Molyaka, Y; Ananév, V; Bazin, I; Garin, A; Narimanov, M; Knysh, V; Melkonyan, H; Umansky, S, and Lichtenstein, A. Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. *Clinical Chemistry*, 46(8 Pt 1):1078–1084, August 2000.

Breitbach, Sarah; Tug, Suzan; Helmig, Susanne; Zahn, Daniela; Kubiak, Thomas; Michal, Matthias; Gori, Tommaso; Ehlert, Tobias; Beiter, Thomas, and Simon, Perikles. Direct quantification of cell-free, circulating DNA from unpurified plasma. *PLoS ONE*, 9(3):e87838, 2014.

Browning, Sharon R and Browning, Brian L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5): 1084–1097, November 2007.

Carroll, S G; Davies, T; Kyle, P M; Abdel-Fattah, S, and Soothill, P W. Fetal karyotyping by chorionic villus sampling after the first trimester. *British journal of obstetrics and gynaecology*, 106(10):1035–1040, October 1999.

Chan, Allen K C; Chiu, Rossa W K; Lo, Y M Dennis, and Clinical Sciences Reviews Committee of the Association of Clinical Biochemists, . Cell-free nucleic acids in plasma, serum and urine: a new tool in molecular diagnosis. *Annals of clinical biochemistry*, 40(Pt 2):122–130, March 2003.

Chang, Christine P-Y; Chia, Rhu-Hsin; Wu, Tsu-Lan; Tsao, Kuo-Chien; Sun, Chien-Feng, and Wu, James T. Elevated cell-free serum DNA detected in patients with myocardial infarction. *Clinica Chimica Acta*, 327(1-2):95–101, January 2003.

Chen, Shengpei; Ge, Huijuan; Wang, Xuebin; Pan, Xiaoyu; Yao, Xiaotian; Li, Xuchao; Zhang, Chunlei; Chen, Fang; Jiang, Fuman; Li, Peipei; Jiang, Hui; Zheng, Hancheng; Zhang, Lei; Zhao, Lijian; Wang, Wei; Li, Songgang; Wang, Jun; Wang, Jian; Yang, Huanming; Li, Yingrui, and Zhang, Xiuqing. Haplotype-assisted accurate non-invasive fetal whole genome recovery through maternal plasma sequencing. *Genome medicine*, 5 (2):18, 2013.

Chitty, Lyn S and Bianchi, Diana W. Next generation sequencing and the next generation: how genomics is revolutionizing reproduction. *Prenatal Diagnosis*, 35 (10):929–930, October 2015.

Chiu, Rossa W K; Chan, K C Allen; Gao, Yuan; Lau, Virginia Y M; Zheng, Wenli; Leung, Tak Y; Foo, Chris H F; Xie, Bin; Tsui, Nancy B Y; Lun, Fiona M F; Zee, Benny C Y; Lau, Tze K; Cantor, Charles R, and Lo, Y M Dennis. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20458–20463, December 2008.

Choate, K A; Lu, Y; Zhou, J; Choi, M; Elias, P M; Farhi, A; Nelson-Williams, C; Crumrine, D; Williams, M L; Nopper, A J; Bree, A; Milstone, L M, and Lifton, R P. Mitotic Recombination in Patients with Ichthyosis Causes Reversion of Dominant Mutations in KRT10. *Science*, 330(6000): 94–97, September 2010.

Chodavarapu, Ramakrishna K; Feng, Suhua; Bernatavichute, Yana V; Chen, Pao-Yang; Stroud, Hume; Yu, Yanchun; Hetzel, Jonathan A; Kuo, Frank; Kim, Jin; Cokus, Shawn J; Casero, David; Bernal, Maria; Huijser, Peter; Clark, Amander T; Krämer, Ute; Merchant, Sabeeha S; Zhang, Xiaoyu; Jacobsen, Steven E, and Pellegrini, Matteo. Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304):388–392, July 2010.

Choi, H; Lau, T K; Jiang, F M; Chan, M K; Zhang, H Y; Lo, P S S; Chen, F; Zhang, L, and Wang, W. Fetal aneuploidy screening by maternal plasma DNA sequencing: 'False positive' due to confined placental

mosaicism. *Prenatal Diagnosis*, pages n/a–n/a, November 2012.

Chudova, D I; Curnow, K J; Bhatt, S; Sehnert, A J, and Bianchi, D W. Maternal copy number variants are a significant reason for false positive noninvasive prenatal test results (Abstract 2103). Presented at the 65th Annual Meeting of the American Society of Human Genetics, 2015.

Committee on Genetics, Society for Maternal-Fetal Medicine, . Committee Opinion No. 640: Cell-Free DNA Screening For Fetal Aneuploidy. *Obstetrics and gynecology*, 126(3):e31–7, September 2015.

Conrad, Donald F; Keebler, Jonathan E M; DePristo, Mark A; Lindsay, Sarah J; Zhang, Yujun; Casals, Ferran; Idaghdour, Youssef; Hartl, Chris L; Torroja, Carlos; Garimella, Kiran V; Zilversmit, Martine; Cartwright, Reed; Rouleau, Guy A; Daly, Mark; Stone, Eric A; Hurles, Matthew E; Awadalla, Philip, and 1000 Genomes Project, . Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, 43(7):712–714, July 2011.

Consortium, 1000 Genomes Project. A map of human genome variation from population-scale sequencing. *Nature*, 2010.

Cooper, Gregory M and Shendure, Jay. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9): 628–640, September 2011.

Dar, Pe'er; Curnow, Kirsten J; Gross, Susan J; Hall, Megan P; Stosic, Melissa; Demko, Zachary; Zimmermann, Bernhard; Hill, Matthew; Sigurjonsson, Styrmir; Ryan, Allison; Banjevic, Milena; Kolacki, Paula L; Koch, Susan W; Strom, Charles M; Rabinowitz, Matthew, and Benn, Peter. Clinical experience and follow-up with large scale single-nucleotide polymorphism-based noninvasive prenatal aneuploidy testing. *American journal of obstetrics and gynecology*, 211(5):527.e1–527.e17, November 2014.

de Bourcy, Charles F A; De Vlaminck, Iwijn; Kanbar, Jad N; Wang, Jianbin; Gawad, Charles, and Quake, Stephen R. A Quantitative Comparison of Single-Cell Whole

Genome Amplification Methods. *PLoS ONE*, 9(8):e105585, August 2014.

De Vlaminck, Iwijn; Valantine, Hannah A; Snyder, Thomas M; Strehl, Calvin; Cohen, Garrett; Luikart, Helen; Neff, Norma F; Okamoto, Jennifer; Bernstein, Daniel; Weisshaar, Dana; Quake, Stephen R, and Khush, Kiran K. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine*, 6(241):241ra77, June 2014.

De Vlaminck, Iwijn; Martin, Lance; Kertesz, Michael; Patel, Kapil; Kowarsky, Mark; Strehl, Calvin; Cohen, Garrett; Luikart, Helen; Neff, Norma F; Okamoto, Jennifer; Nicolls, Mark R; Cornfield, David; Weill, David; Valantine, Hannah; Khush, Kiran K, and Quake, Stephen R. Noninvasive monitoring of infection and rejection after lung transplantation. *Proceedings of the National Academy of Sciences*, 112(43):13336–13341, October 2015.

DePristo, Mark A; Banks, Eric; Poplin, Ryan; Garimella, Kiran V; Maguire, Jared R; Hartl, Christopher; Philippakis, Anthony A; del Angel, Guillermo; Rivas, Manuel A; Hanna, Matt; McKenna, Aaron; Fennell, Tim J; Kernytsky, Andrew M; Sivachenko, Andrey Y; Cibulskis, Kristian; Gabriel, Stacey B; Altshuler, David, and Daly, Mark J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, May 2011.

Diaz, Luis A and Bardelli, Alberto. Liquid biopsies: genotyping circulating tumor DNA. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 32(6):579–586, February 2014.

Diaz, Luis A; Sausen, Mark; Fisher, George A, and Velculescu, Victor E. Insights into therapeutic resistance from whole-genome analyses of circulating tumor DNA. *Oncotarget*, 4(10):1856–1857, October 2013.

Diehl, Frank; Schmidt, Kerstin; Choti, Michael A; Romans, Katharine; Goodman, Steven; Li, Meng; Thornton, Katherine; Agrawal, Nishant; Sokoll, Lori; Szabo, Steve A; Kinzler, Kenneth W; Vogelstein, Bert, and JrDiaz, Luis A. Circulating mu-

tant DNA to assess tumor dynamics. *Nature Medicine*, 14(9):985–990, July 2007.

Eckmann-Scholz, Christel; Tönnies, Holger; Liehr, Thomas; Gesk, Stefan; Jonat, Walter, and Caliebe, Almuth. Normal prenatal ultrasound findings in a case with de novo mosaic small supernumerary marker chromosome 18–how to counsel? *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, 25(2):200–202, February 2012.

Engel, Eric. A fascination with chromosome rescue in uniparental disomy: Mendelian recessive outlaws and imprinting copyrights infringements. *European Journal of Human Genetics*, 14(11):1158–1169, May 2006.

Fan, H Christina; Blumenfeld, Yair J; Chitkara, Usha; Hudgins, Louanne, and Quake, Stephen R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proceedings of the National Academy of Sciences of the United States of America*, 105(42): 16266–16271, October 2008.

Fan, H Christina; Wang, Jianbin; Potanina, Anastasia, and Quake, Stephen R. Whole-genome molecular haplotyping of single cells. *Nature Biotechnology*, 29(1):51–57, December 2010.

Fan, H Christina; Gu, Wei; Wang, Jianbin; Blumenfeld, Yair J; El-Sayed, Yasser Y, and Quake, Stephen R. Non-invasive prenatal measurement of the fetal genome. *Nature*, 487(7407):320–324, July 2012.

Filges, Isabel; Kang, Anjeung; Klug, Vanessa; Wenzel, Friedel; Heinimann, Karl; Tercanli, Sevgi, and Miny, Peter. aCGH on chorionic villi mirrors the complexity of fetoplacental mosaicism in prenatal diagnosis. *Prenatal Diagnosis*, 31(5):473–478, February 2011.

Fleischhacker, M and Schmidt, B. Circulating nucleic acids (CNAs) and cancer—A survey. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1775(1):181–232, January 2007.

Forshew, Tim; Murtaza, Muhammed; Parkinson, Christine; Gale, Davina; Tsui,

Dana W Y; Kaper, Fiona; Dawson, Sarah-Jane; Piskorz, Anna M; Jimenez-Linan, Mercedes; Bentley, David; Hadfield, James; May, Andrew P; Caldas, Carlos; Brenton, James D, and Rosenfeld, Nitzan. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Science translational medicine*, 4(136):136ra68, May 2012.

Fu, Yutao; Sinha, Manisha; Peterson, Craig L, and Weng, Zhiping. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genetics*, 4(7):e1000138, 2008.

Gaffney, Daniel J; McVicker, Graham; Pai, Athma A; Fondufe-Mittendorf, Yvonne N; Lewellen, Noah; Michelini, Katelyn; Widom, Jonathan; Gilad, Yoav, and Pritchard, Jonathan K. Controls of Nucleosome Positioning in the Human Genome. *PLoS Genetics*, 8(11):e1003036, November 2012.

Galeazzi, M; Morozzi, G; Piccini, M; Chen, J; Bellisai, F; Fineschi, S, and Marcolongo, R. Dosage and characterization of circulating DNA: present usage and possible applications in systemic autoimmune disorders. *Autoimmunity reviews*, 2(1):50–55, January 2003.

Gansauge, Marie-Theres and Meyer, Matthias. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature protocols*, 8 (4):737–748, April 2013.

García-Olmo, Dolores C; Picazo, María G; Toboso, Inmaculada; Asensio, Ana I, and García-Olmo, Damián. Quantitation of cell-free DNA and RNA in plasma during tumor progression in rats. *Molecular cancer*, 12:8, 2013.

Gawad, Charles; Koh, Winston, and Quake, Stephen R. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, March 2016.

Grant, Charles E; Bailey, Timothy L, and Noble, William Stafford. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

Grati, Francesca R; Malvestiti, Francesca; Ferreira, Jose C P B; Bajaj, Komal;

Gaetani, Elisa; Agrati, Cristina; Grimi, Beatrice; Dulcetti, Francesca; Ruggeri, Anna M; De Toffol, Simona; Maggi, Federico; Wapner, Ronald; Gross, Susan, and Simoni, Giuseppe. Fetoplacental mosaicism: potential implications for false-positive and false-negative noninvasive prenatal screening results. *Genetics in Medicine*, February 2014.

Greco, F Anthony and Hainsworth, John D. Introduction: unknown primary cancer. *Seminars in oncology*, 36(1):6–7, February 2009.

Hahnemann, J M and Vejerslev, L O. Accuracy of cytogenetic findings on chorionic villus sampling (CVS)–diagnostic consequences of CVS mosaicism and non-mosaic discrepancy in centres contributing to EUCROMIC 1986-1992. *Prenatal Diagnosis*, 17(9):801–820, September 1997.

Harshman, Sean W; Young, Nicolas L; Parthun, Mark R, and Freitas, Michael A. H1 histones: current perspectives and challenges. *Nucleic acids research*, 41 (21):9593–9609, November 2013.

Haymon, Lori; Simi, Eve; Moyer, Kelly; Aufox, Sharon, and Ouyang, David W. Clinical implementation of noninvasive prenatal testing among maternal fetal medicine specialists. *Prenatal Diagnosis*, 34(5): 416–423, May 2014.

Hiatt, J B; Pritchard, C C; Salipante, S J; O'Roak, B J, and Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Research*, 23(5):843–854, May 2013.

Hill, Melissa; Twiss, Philip; Verhoef, Talitha I; Drury, Suzanne; McKay, Fiona; Mason, Sarah; Jenkins, Lucy; Morris, Stephen, and Chitty, Lyn S. Non-invasive prenatal diagnosis for cystic fibrosis: detection of paternal mutations, exploration of patient preferences and cost analysis. *Prenatal Diagnosis*, 35(10):950–958, October 2015.

Holdenrieder, Stefan; Stieber, Petra; Chan, Lisa Y S; Geiger, Sandra; Kremer, Andreas; Nagel, Dorothea, and Lo, Y M Dennis. Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. *Clinical Chemistry*, 51(8):1544–1546, August 2005.

Inbar-Feigenberg, Michal; Choufani, Sanaa; Cytrynbaum, Cheryl; Chen, Yi-An; Steele, Leslie; Shuman, Cheryl; Ray, Peter N, and Weksberg, Rosanna. Mosaicism for genome-wide paternal uniparental disomy with features of multiple imprinting disorders: Diagnostic and management issues. *American Journal of Medical Genetics Part A*, 161(1):13–20, December 2012.

International Parkinson Disease Genomics Consortium, ; Nalls, Michael A; Plagnol, Vincent; Hernandez, Dena G; Sharma, Manu; Sheerin, Una-Marie; Saad, Mohamad; Simón-Sánchez, J; Schulte, Claudia; Lesage, Suzanne; Sveinbjörnsdóttir, Sigurlaug; Stefansson, Kari; Martinez, Maria; Hardy, John; Heutink, Peter; Brice, Alexis; Gasser, Thomas; Singleton, Andrew B, and Wood, Nicholas W. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, 377(9766): 641–649, February 2011.

Kalousek, D K and Dill, F J. Chromosomal mosaicism confined to the placenta in human conceptions. *Science*, 221(4611): 665–667, August 1983.

Kevles, Daniel J and Hood, Leroy E. *The Code of Codes*. Scientific and Social Issues in the Human Genome Project. Harvard University Press, January 1993.

Kinde, Isaac; Wu, Jian; Papadopoulos, Nick; Kinzler, Kenneth W, and Vogelstein, Bert. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(23):9530–9535, June 2011.

Kingsmore, Stephen F; Petrikin, Josh; Willig, Laurel K, and Guest, Erin. Emergency medical genomes: a breakthrough application of precision medicine. *Genome medicine*, 7(1):82, 2015.

Kircher, Martin; Witten, Daniela M; Jain, Preti; O'Roak, Brian J; Cooper, Gregory M, and Shendure, Jay. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, March 2014.

Kitzman, Jacob O; Mackenzie, Alexandra P; Adey, Andrew; Hiatt, Joseph B; Patward-

han, Rupali P; Sudmant, Peter H; Ng, Sarah B; Alkan, Can; Qiu, Ruolan; Eichler, Evan E, and Shendure, Jay. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnology*, 29(1):59–63, January 2011.

Kitzman, Jacob O; Snyder, Matthew W; Ventura, Mario; Lewis, Alexandra P; Qiu, Ruolan; Simmons, Lavone E; Gammill, Hilary S; Rubens, Craig E; Santillan, Donna A; Murray, Jeffrey C; Tabor, Holly K; Bamshad, Michael J; Eichler, Evan E, and Shendure, Jay. Noninvasive whole-genome sequencing of a human fetus. *Science translational medicine*, 4 (137):137ra76, June 2012.

Koh, Winston; Pan, Wenying; Gawad, Charles; Fan, H Christina; Kerchner, Geoffrey A; Wyss-Coray, Tony; Blumenfeld, Yair J; El-Sayed, Yasser Y, and Quake, Stephen R. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences*, 111(20): 7361–7366, May 2014.

Kong, Augustine; Frigge, Michael L; Masson, Gisli; Besenbacher, Soren; Sulem, Patrick; Magnusson, Gisli; Gudjonsson, Sigurjon A; Sigurdsson, Asgeir; Jonasdottir, Aslaug; Jonasdottir, Adalbjorg; Wong, Wendy S W; Sigurdsson, Gunnar; Walters, G Bragi; Steinberg, Stacy; Helgason, Hannes; Thorleifsson, Gudmar; Gudbjartsson, Daniel F; Helgason, Agnar; Magnusson, Olafur Th; Thorsteinsdottir, Unnur, and Stefansson, Kari. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412): 471–475, August 2012.

Koshland, D E. Sequences and consequences of the human genome. *Science*, 246(4927):189, October 1989.

Kumar, Akash; Ryan, Allison; Kitzman, Jacob O; Wemmer, Nina; Snyder, Matthew W; Sigurjonsson, Styrmir; Lee, Choli; Banjevic, Milena; Zarutskie, Paul W; Lewis, Alexandra P; Shendure, Jay, and Rabinowitz, Matthew. Whole genome prediction for preimplantation genetic diagnosis. *Genome medicine*, 7 (1):35, 2015.

Lau, T K; Cheung, S W; Lo, P S S; Pursley, A N; Chan, M K; Jiang, F; Zhang, H; Wang, W; Jong, L F J; Yuen, O K C;

Chan, H Y C; Chan, W S K, and Choy, K W. Non-invasive prenatal testing for fetal chromosomal abnormalities by low-coverage whole-genome sequencing of maternal plasma DNA: review of 1982 consecutive cases in a single center. *Ultrasound in Obstetrics & Gynecology*, 43 (3):254–264, February 2014.

Lau, Tze Kin; Jiang, Fu Man; Stevenson, Robert J; Lo, Tsz Kin; Chan, Lin Wai; Chan, Mei Ki; Lo, Pui Shan Salome; Wang, Wei; Zhang, Hong-Yun; Chen, Fang, and Choy, Kwong Wai. Secondary findings from non-invasive prenatal testing for common fetal aneuploidies by whole genome sequencing as a clinical service. *Prenatal Diagnosis*, 33(6):602–608, June 2013.

Leary, R J; Sausen, M; Kinde, I; Papadopoulos, N; Carpten, J D; Craig, D; O'Shaughnessy, J; Kinzler, K W; Parmigiani, G; Vogelstein, B; Diaz, L A, and Velculescu, V E. Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. *Science translational medicine*, 4(162): 162ra154–162ra154, November 2012.

Lewontin, R C. Annotation: the analysis of variance and the analysis of causes. *American journal of human genetics*, 26(3): 400–411, May 1974.

Li, Heng and Durbin, Richard. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

Li, Heng and Durbin, Richard. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26 (5):589–595, March 2010.

Li, Heng; Handsaker, Bob; Wysoker, Alec; Fennell, Tim; Ruan, Jue; Homer, Nils; Marth, Gabor; Abecasis, Goncalo; Durbin, Richard, and 1000 Genome Project Data Processing Subgroup, . The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

Li, Meng; Diehl, Frank; Dressman, Devin; Vogelstein, Bert, and Kinzler, Kenneth W. BEAMing up for detection and quantification of rare sequence variants. *Nature Methods*, 3(2):95–97, February 2006.

Lo, Y M; Corbetta, N; Chamberlain, P F; Rai,

V; Sargent, I L; Redman, C W, and Wainscoat, J S. Presence of fetal DNA in maternal plasma and serum. *Lancet*, 350 (9076):485–487, August 1997.

Lo, Y M; Tein, M S; Lau, T K; Haines, C J; Leung, T N; Poon, P M; Wainscoat, J S; Johnson, P J; Chang, A M, and Hjelm, N M. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *American journal of human genetics*, 62 (4):768–775, April 1998a.

Lo, Y M; Tein, M S; Pang, C C; Yeung, C K; Tong, K L, and Hjelm, N M. Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. *Lancet*, 351(9112):1329–1330, May 1998b.

Lo, Y M; Zhang, J; Leung, T N; Lau, T K; Chang, A M, and Hjelm, N M. Rapid clearance of fetal DNA from maternal plasma. *American journal of human genetics*, 64 (1):218–224, January 1999.

Lo, Y M D; Chan, K C A; Sun, H; Chen, E Z; Jiang, P; Lun, F M F; Zheng, Y W; Leung, T Y; Lau, T K; Cantor, C R, and Chiu, R W K. Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Science translational medicine*, 2 (61):61ra91–61ra91, December 2010.

Lo, Y M Dennis and Chiu, Rossa W K. Prenatal diagnosis: progress through plasma nucleic acids. *Nature Reviews Genetics*, 8 (1):71–77, January 2007.

Lui, Yanni Y N; Chik, Ki-Wai; Chiu, Rossa W K; Ho, Cheong-Yip; Lam, Christopher W K, and Lo, Y M Dennis. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clinical Chemistry*, 48(3):421–427, March 2002.

Lui, Yanni Y N; Woo, Kam-Sang; Wang, Angela Y M; Yeung, Chung-Kwong; Li, Philip K T; Chau, Elaine; Ruygrok, Peter, and Lo, Y M Dennis. Origin of plasma cell-free DNA after solid organ transplantation. *Clinical Chemistry*, 49(3):495–496, March 2003.

Ma, Li; Xiao, Yan; Huang, Hui; Wang, Qingwei; Rao, Weinian; Feng, Yue; Zhang, Kui, and Song, Qing. Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods*, 7(4): 299–301, March 2010.

MacArthur, Daniel G; Balasubramanian, Suganthi; Frankish, Adam; Huang, Ni; Morris, James; Walter, Klaudia; Jostins, Luke; Habegger, Lukas; Pickrell, Joseph K; Montgomery, Stephen B; Albers, Cornelis A; Zhang, Zhengdong D; Conrad, Donald F; Lunter, Gerton; Zheng, Hancheng; Ayub, Qasim; DePristo, Mark A; Banks, Eric; Hu, Min; Handsaker, Robert E; Rosenfeld, Jeffrey A; Fromer, Menachem; Jin, Mike; Mu, Xinmeng Jasmine; Khurana, Ekta; Ye, Kai; Kay, Mike; Saunders, Gary Ian; Suner, Marie-Marthe; Hunt, Toby; Barnes, If H A; Amid, Clara; Carvalho-Silva, Denise R; Bignell, Alexandra H; Snow, Catherine; Yngvadottir, Bryndis; Bumpstead, Suzannah; Cooper, David N; Xue, Yali; Romero, Irene Gallego; 1000 Genomes Project Consortium, ; Wang, Jun; Li, Yingrui; Gibbs, Richard A; McCarroll, Steven A; Dermitzakis, Emmanouil T; Pritchard, Jonathan K; Barrett, Jeffrey C; Harrow, Jennifer; Hurles, Matthew E; Gerstein, Mark B, and Tyler-Smith, Chris. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, February 2012.

Mandel, P and Métais, P. Les acides nucléiques du plasma sanguin chez l'Homme. *Comptes rendus des séances de la Société de biologie et de ses filiales*, 142(3-4):241–243, February 1948.

Mao, Jun; Wang, Ting; Wang, Ben-Jing; Liu, Ying-Hua; Li, Hong; Zhang, Jianguang; Cram, David, and Chen, Ying. Clinica Chimica Acta. *Clinica Chimica Acta*, 433:190–193, 2014.

Marle, N; Martinet, D; Aboura, A; Joly-Helas, G; Andrieux, J; Flori, E; Puechberty, J; Vialard, F; Sanlaville, D; Fert Ferrer, S; Bourrouillou, G; Tabet, A C; Quilichini, B; Simon-Bouy, B; Bazin, A; Becker, M; Stora, H; Amblard, S; Doco-Fenzy, M; Molina Gomes, D; Girard-Lemaire, F; Cordier, M P; Satre, V; Schneider, A; Lemeur, N; Chambon, P; Jacquemont, S; Fellmann, F; Vigouroux-Castera, A; Molignier, R; Delaye, A; Pipiras, E; Liquier, A; Rousseau, T; Mosca, A L; Kremer, V; Payet, M; Rangon, C; Mugneret, F; Aho, S; Faivre, L, and Callier, P. Molecular characterization of 39 de novo sSMC: contribution to prognosis and genetic coun-

selling, a prospective study. *Clinical genetics*, 85(3):233–244, March 2014.

Masuzaki, H; Miura, K; Yoshiura, K-i; Yoshimura, S; Niikawa, N, and Ishimaru, T. Detection of cell free placental DNA in maternal plasma: direct evidence from three cases of confined placental mosaicism. *Journal of Medical Genetics*, 41 (4):289–292, April 2004.

Maurano, Matthew T; Humbert, Richard; Rynes, Eric; Thurman, Robert E; Haugen, Eric; Wang, Hao; Reynolds, Alex P; Sandstrom, Richard; Qu, Hongzhu; Brody, Jennifer; Shafer, Anthony; Neri, Fidencio; Lee, Kristen; Kutyavin, Tanya; Stehling-Sun, Sandra; Johnson, Audra K; Canfield, Theresa K; Giste, Erika; Diegel, Morgan; Bates, Daniel; Hansen, R Scott; Neph, Shane; Sabo, Peter J; Heimfeld, Shelly; Raubitschek, Antony; Ziegler, Steven; Cotsapas, Chris; Sotoodehnia, Nona; Glass, Ian; Sunyaev, Shamil R; Kaul, Rajinder, and Stamatoyannopoulos, John A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

Mennuti, Michael T; Cherry, Athena M; Morrissette, Jennifer J D, and Dugoff, Lorraine. Is it time to sound an alarm about false-positive cell-free DNA testing for fetal aneuploidy? *American journal of obstetrics and gynecology*, 209(5):415–419, November 2013.

Mouliere, Florent; El Messaoudi, Safia; Pang, Dalong; Dritschilo, Anatoly, and Thierry, Alain R. Multi-marker analysis of circulating cell-free DNA toward personalized medicine for colorectal cancer. *Molecular oncology*, 8(5):927–941, July 2014.

Murtaza, Muhammed; Dawson, Sarah-Jane; Tsui, Dana W Y; Gale, Davina; Forshew, Tim; Piskorz, Anna M; Parkinson, Christine; Chin, Suet-Feung; Kingsbury, Zoya; Wong, Alvin S C; Marass, Francesco; Humphray, Sean; Hadfield, James; Bentley, David; Chin, Tan Min; Brenton, James D; Caldas, Carlos, and Rosenfeld, Nitzan. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 497(7447):108–112, May 2013.

Nelson, J Lee. The otherness of self:

microchimerism in health and disease. *Trends in immunology*, 33(8):421–427, August 2012.

Newman, Aaron M; Bratman, Scott V; To, Jacqueline; Wynne, Jacob F; Eclov, Neville C W; Modlin, Leslie A; Liu, Chih Long; Neal, Joel W; Wakelee, Heather A; Merritt, Robert E; Shrager, Joseph B; Loo, Billy W; Alizadeh, Ash A, and Diehn, Maximilian. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine*, pages 1–9, April 2014.

Nicolaides, Kypros H; Syngelaki, Argyro; Ashoor, Ghalia; Birdir, Cahit, and Touzet, Gisele. Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. *American journal of obstetrics and gynecology*, 207 (5):374.e1–6, November 2012.

Norton, Mary E; Jacobsson, Bo; Swamy, Geeta K; Laurent, Louise C; Ranzini, Angela C; Brar, Herb; Tomlinson, Mark W; Pereira, Leonardo; Spitz, Jean L; Hollemon, Desiree; Cuckle, Howard; Musci, Thomas J, and Wapner, Ronald J. Cell-free DNA analysis for noninvasive examination of trisomy. *The New England journal of medicine*, 372(17):1589–1597, April 2015.

Nygren, A O H; Dean, J; Jensen, T J; Kruse, S; Kwong, W; van den Boom, D, and Ehrich, M. Quantification of Fetal DNA by Use of Methylation-Based DNA Discrimination. *Clinical Chemistry*, 56(10):1627–1635, September 2010.

Ong, Chin-Tong and Corces, Victor G. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics*, 15(4):234–246, April 2014.

O'Roak, Brian J; Vives, Laura; Fu, Wenqing; Egertson, Jarrett D; Stanaway, Ian B; Phelps, Ian G; Carvill, Gemma; Kumar, Akash; Lee, Choli; Ankenman, Katy; Munson, Jeff; Hiatt, Joseph B; Turner, Emily H; Levy, Roie; O'Day, Diana R; Krumm, Niklas; Coe, Bradley P; Martin, Beth K; Borenstein, Elhanan; Nickerson, Deborah A; Mefford, Heather C; Doherty, Dan; Akey, Joshua M; Bernier, Raphael; Eichler, Evan E, and Shendure, Jay. Multiplex targeted sequencing identifies recurrently mutated genes in autism spec-

trum disorders. *Science*, 338(6114):1619–1622, December 2012.

Osborne, C Michael; Hardisty, Emily; Devers, Patricia; Kaiser-Rogers, Kathleen; Hayden, Melissa A; Goodnight, William, and Vora, Neeta L. Discordant noninvasive prenatal testing results in a patient subsequently diagnosed with metastatic disease. *Prenatal Diagnosis*, 33(6):609–611, April 2013.

Palomaki, Glenn E; Deciu, Cosmin; Kloza, Edward M; Lambert-Messerlian, Geralyn M; Haddow, James E; Neveux, Louis M; Ehrich, Mathias; van den Boom, Dirk; Bombard, Allan T; Grody, Wayne W; Nelson, Stanley F, and Canick, Jacob A. DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. *Genetics in Medicine*, 14(3):296–305, March 2012.

Pedersen, J S; Valen, E; Velazquez, A M V; Parker, B J; Rasmussen, M; Lindgreen, S; Lilje, B; Tobin, D J; Kelly, T K; Vang, S; Andersson, R; Jones, P A; Hoover, C A; Tikhonov, A; Prokhortchouk, E; Rubin, E M; Sandelin, A; Gilbert, M T P; Krogh, A; Willerslev, E, and Orlando, L. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research*, 24(3): 454–466, March 2014.

Quake, Stephen. Sizing up cell-free DNA. *Clinical Chemistry*, 58(3):489–490, March 2012.

Rainer, Timothy H; Wong, Lawrence K S; Lam, Wynnie; Yuen, Eddie; Lam, Nicole Y L; Metreweli, Constantine, and Lo, Y M Dennis. Prognostic use of circulating plasma nucleic acid concentrations in patients with acute stroke. *Clinical Chemistry*, 49(4):562–569, April 2003.

Rao, Suhas S P; Huntley, Miriam H; Durand, Neva C; Stamenova, Elena K; Bochkov, Ivan D; Robinson, James T; Sanborn, Adrian L; Machol, Ido; Omer, Arina D; Lander, Eric S, and Aiden, Erez Lieberman. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

Rava, R P; Srinivasan, A; Sehnert, A J, and Bianchi, D W. Circulating Fetal Cell-Free

DNA Fractions Differ in Autosomal Aneuploidies and Monosomy X. *Clinical Chemistry*, 60(1):243–250, December 2013.

Rhodes, C H; Honsinger, C, and Sorenson, G D. Detection of tumor-derived DNA in cerebrospinal fluid. *Journal of neuropathology and experimental neurology*, 53(4):364–368, July 1994.

Roach, Jared C; Glusman, Gustavo; Smit, Arian F A; Huff, Chad D; Hubley, Robert; Shannon, Paul T; Rowen, Lee; Pant, Krishna P; Goodman, Nathan; Bamshad, Michael; Shendure, Jay; Drmanac, Radoje; Jorde, Lynn B; Hood, Leroy, and Galas, David J. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978): 636–639, April 2010.

Sambuy, Y; De Angelis, I; Ranaldi, G; Scarino, M L; Stammati, A, and Zucco, F. The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell biology and toxicology*, 21(1):1–26, January 2005.

Saunders, C J; Miller, N A; Soden, S E; Dinwiddie, D L; Noll, A; Alnadi, N A; Andraws, N; Patterson, M L; Krivohlavek, L A; Fellis, J; Humphray, S; Saffrey, P; Kingsbury, Z; Weir, J C; Betley, J; Grocock, R J; Margulies, E H; Farrow, E G; Artman, M; Safina, N P; Petrikin, J E; Hall, K P, and Kingsmore, S F. Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. *Science translational medicine*, 4(154):154ra135–154ra135, October 2012.

Savitzky, Abraham and Golay, M J E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964.

Schep, Alicia N; Buenrostro, Jason D; Denny, Sarah K; Schwartz, Katja; Sherlock, Gavin, and Greenleaf, William J. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, August 2015.

Schones, Dustin E; Cui, Kairong; Cuddapah, Suresh; Roh, Tae-Young; Barski, Artem; Wang, Zhibin; Wei, Gang, and Zhao, Keji.

Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132 (5):887–898, March 2008.

Snyder, Thomas M; Khush, Kiran K; Valantine, Hannah A, and Quake, Stephen R. Universal noninvasive detection of solid organ transplant rejection. *Proceedings of the National Academy of Sciences*, 108 (15):6229–6234, April 2011.

Sparks, Andrew B; Wang, Eric T; Struble, Craig A; Barrett, Wade; Stokowski, Renee; McBride, Celeste; Zahn, Jacob; Lee, Kevin; Shen, Naiping; Doshi, Jigna; Sun, Michel; Garrison, Jill; Sandler, Jay; Hollemon, Desiree; Pattee, Patrick; Tomita-Mitchell, Aoy; Mitchell, Michael; Stuelpnagel, John; Song, Ken, and Oliphant, Arnold. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. *Prenatal Diagnosis*, 32(1):3–9, January 2012.

Srinivasan, Anupama; Bianchi, Diana W; Huang, Hui; Sehnert, Amy J, and Rava, Richard P. Noninvasive Detection of Fetal Subchromosome Abnormalities via Deep Sequencing of Maternal Plasma. *American journal of human genetics*, pages 1–10, January 2013.

Stenson, Peter D; Ball, Edward V; Howells, Katy; Phillips, Andrew D; Mort, Matthew, and Cooper, David N. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Human genomics*, 4(2):69–72, December 2009.

Stergachis, Andrew B; Neph, Shane; Reynolds, Alex; Humbert, Richard; Miller, Brady; Paige, Sharon L; Vernot, Benjamin; Cheng, Jeffrey B; Thurman, Robert E; Sandstrom, Richard; Haugen, Eric; Heimfeld, Shelly; Murry, Charles E; Akey, Joshua M, and Stamatoyannopoulos, John A. Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell*, 154 (4):888–903, August 2013.

Stetten, Gail; Escallon, Cathleen S; South, Sarah T; McMichael, Joseph L; Saul, Daniel O, and Blakemore, Karin J. Reevaluating confined placental mosaicism. *American Journal of Medical Genetics Part A*, 131A(3):232–239, 2004.

Stroun, M; Anker, P; Lyautey, J; Lederrey, C, and Maurice, P A. Isolation and characterization of DNA from the plasma of cancer patients. *European journal of cancer & clinical oncology*, 23(6):707–712, June 1987.

Sudmant, P H; Kitzman, J O; Antonacci, F; Alkan, C; Malig, M; Tsalenko, A; Sampas, N; Bruhn, L; Shendure, J; 1000 Genomes Project, , and Eichler, E E. Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, 330(6004):641–646, October 2010.

Sun, Kun; Jiang, Peiyong; Chan, K C Allen; Wong, John; Cheng, Yvonne K Y; Liang, Raymond H S; Chan, Wai-kong; Ma, Edmond S K; Chan, Stephen L; Cheng, Suk Hang; Chan, Rebecca W Y; Tong, Yu K; Ng, Simon S M; Wong, Raymond S M; Hui, David S C; Leung, Tse Ngong; Leung, Tak Y; Lai, Paul B S; Chiu, Rossa W K, and Lo, Yuk Ming Dennis. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences*, 112(40):E5503–12, October 2015.

Tabor, Holly K; Murray, Jeffrey C; Gammill, Hilary S; Kitzman, Jacob O; Snyder, Matthew W; Ventura, Mario; Lewis, Alexandra P; Qiu, Ruolan; Simmons, Lavone E; Rubens, Craig E; Santillan, Mark K; Eichler, Evan E; Cheng, Edith Y; Bamshad, Michael J, and Shendure, Jay. Non-invasive fetal genome sequencing: Opportunities and challenges. *American Journal of Medical Genetics Part A*, pages n/a–n/a, August 2012.

Teif, Vladimir B; Vainshtein, Yevhen; Caudron-Herger, Maïwen; Mallm, Jan-Philipp; Marth, Caroline; Höfer, Thomas, and Rippe, Karsten. Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology*, 19(11):1185–1192, November 2012.

Thierry, Alain R; Mouliere, Florent; Gongora, Celine; Ollier, Jeremy; Robert, Bruno; Ychou, Marc; Del Rio, Maguy, and Molina, Franck. Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nu-*

cleic acids research, 38(18):6159–6175, October 2010.

Thierry, Alain R; Mouliere, Florent; El Messaoudi, Safia; Mollevi, Caroline; Lopez-Crapez, Evelyne; Rolet, Fanny; Gillet, Brigitte; Gongora, Celine; Dechelotte, Pierre; Robert, Bruno; Del Rio, Maguy; Lamy, Pierre-Jean; Bibeau, Frederic; Nouaille, Michelle; Loriot, Virginie; Jarrousse, Anne-Sophie; Molina, Franck; Mathonnet, Muriel; Pezet, Denis, and Ychou, Marc. Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. *Nature Medicine*, 20(4):430–435, April 2014.

Tug, Suzan; Helmig, Susanne; Menke, Julia; Zahn, Daniela; Kubiak, Thomas; Schwarting, Andreas, and Simon, Perikles. Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cellular immunology*, 292(1-2): 32–39, November 2014.

Uhlén, Mathias; Fagerberg, Linn; Hallström, Björn M; Lindskog, Cecilia; Oksvold, Per; Mardinoglu, Adil; Sivertsson, Åsa; Kampf, Caroline; Sjöstedt, Evelina; Asplund, Anna; Olsson, IngMarie; Edlund, Karolina; Lundberg, Emma; Navani, Sanjay; Szigyarto, Cristina Al-Khalili; Odeberg, Jacob; Djureinovic, Dijana; Takanen, Jenny Ottosson; Hober, Sophia; Alm, Tove; Edqvist, Per-Henrik; Berling, Holger; Tegel, Hanna; Mulder, Jan; Rockberg, Johan; Nilsson, Peter; Schwenk, Jochen M; Hamsten, Marica; von Feilitzen, Kalle; Forsberg, Mattias; Persson, Lukas; Johansson, Fredric; Zwahlen, Martin; von Heijne, Gunnar; Nielsen, Jens, and Pontén, Fredrik. Proteomics. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, January 2015.

Vago, Luca; Perna, Serena Kimi; Zanussi, Monica; Mazzi, Benedetta; Barlassina, Cristina; Stanghellini, Maria Teresa Lupo; Perrelli, Nicola Flavio; Cosentino, Cristian; Torri, Federica; Angius, Andrea; Forno, Barbara; Casucci, Monica; Bernardi, Massimo; Peccatori, Jacopo; Corti, Consuelo; Bondanza, Attilio; Ferrari, Maurizio; Rossini, Silvano; Roncarolo, Maria Grazia; Bordignon, Claudio; Bonini, Chiara; Ciceri, Fabio, and Fleischhauer, Katharina. Loss of mismatched

HLA in leukemia after stem-cell transplantation. *The New England journal of medicine*, 361(5):478–488, July 2009.

Valouev, Anton; Johnson, Steven M; Boyd, Scott D; Smith, Cheryl L; Fire, Andrew Z, and Sidow, Arend. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, April 2012.

van den Oever, Jessica M E; Bijlsma, Emilia K; Feenstra, Ilse; Muntjewerff, Nienke; Mathijssen, Inge B; Bakker, Egbert; van Belzen, Martine J, and Boon, Elles M J. Noninvasive prenatal diagnosis of Huntington disease: detection of the paternally inherited expanded CAG repeat in maternal plasma. *Prenatal Diagnosis*, 35(10):945–949, October 2015.

Veltman, Joris A and Brunner, Han G. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575, August 2012.

Vierstra, Jeff; Wang, Hao; John, Sam; Sandstrom, Richard, and Stamatoyannopoulos, John A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature Methods*, 11 (1):66–72, November 2013.

Wang, Hao; Maurano, Matthew T; Qu, Hongzhu; Varley, Katherine E; Gertz, Jason; Pauli, Florencia; Lee, Kristen; Canfield, Theresa; Weaver, Molly; Sandstrom, Richard; Thurman, Robert E; Kaul, Rajinder; Myers, Richard M, and Stamatoyannopoulos, John A. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, 22(9): 1680–1688, September 2012.

Wang, Y; Chen, Y; Tian, F; Zhang, J; Song, Z; Wu, Y; Han, X; Hu, W; Ma, D; Cram, D, and Cheng, W. Maternal Mosaicism Is a Significant Contributor to Discordant Sex Chromosomal Aneuploidies Associated with Noninvasive Prenatal Testing. *Clinical Chemistry*, 60(1):251–259, December 2013.

Wapner, Ronald J; Babiarz, Joshua E; Levy, Brynn; Stosic, Melissa; Zimmermann, Bernhard; Sigurjonsson, Styrmir; Wayham, Nicholas; Ryan, Allison; Banjevic, Milena; Lacroute, Phil; Hu, Jing; Hall, Megan P; Demko, Zachary; Siddiqui, Asim; Rabinowitz, Matthew; Gross, Su-

san J; Hill, Matthew, and Benn, Peter. Expanding the scope of noninvasive prenatal testing: detection of fetal microdeletion syndromes. *American journal of obstetrics and gynecology*, 212(3):332.e1–9, March 2015.

Willig, Laurel K; Petrikin, Josh E; Smith, Laurie D; Saunders, Carol J; Thiffault, Isabelle; Miller, Neil A; Soden, Sarah E; Cakici, Julie A; Herd, Suzanne M; Twist, Greyson; Noll, Aaron; Creed, Mitchell; Alba, Patria M; Carpenter, Shannon L; Clements, Mark A; Fischer, Ryan T; Hays, J Allyson; Kilbride, Howard; McDonough, Ryan J; Rosterman, Jamie L; Tsai, Sarah L; Zellmer, Lee; Farrow, Emily G, and Kingsmore, Stephen F. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *The Lancet. Respiratory medicine*, 3(5):377–387, May 2015.

Wimberger, Pauline; Roth, Carina; Pantel, Klaus; Kasimir-Bauer, Sabine; Kimmig, Rainer, and Schwarzenbach, Heidi. Impact of platinum-based chemotherapy on circulating nucleic acid levels, protease activities in blood and disseminated tumor cells in bone marrow of ovarian cancer patients. *International Journal of Cancer*, 128(11):2572–2580, October 2010.

Yang, Hong; Chen, Xi, and Wong, Wing Hung. Completely phased genome sequencing through chromosome sorting. *Proceedings of the National Academy of Sciences*, 108(1):12–17, January 2011.

Zimmermann, Bernhard; Hill, Matthew; Gemelos, George; Demko, Zachary; Banjevic, Milena; Baner, Johan; Ryan, Allison; Sigurjonsson, Styrmir; Chopra, Nikhil; Dodd, Michael; Levy, Brynn, and Rabinowitz, Matthew. Noninvasive prenatal aneuploidy testing of chromosomes 13, 18, 21, X, and Y, using targeted sequencing of polymorphic loci. *Prenatal Diagnosis*, 32(13):1233–1241, October 2012.

VITA

Matthew Snyder graduated in 2006 from the University of Pennsylvania, where he studied economics for far too long. He then spent two years working at a rural clinic in Kenya providing free medical care to underserved populations, during which time he developed a keen interest in cheap and robust diagnostics for point-of-care applications. In 2011, he earned his M.S. in Biostatistics from the University of Michigan School of Public Health, where he worked on statistical methods in genomics with Dr. Gonçalo Abecasis. His current research interests involve the development of noninvasive diagnostic and screening tests for cancer and maternal-fetal medicine, with a particular focus on developing low-cost assays for use in the developing world.

Outside of the lab, he plays checkers, rides his many bicycles, and writes angry letters to city agencies. He rarely receives responses.