TOOLS AND STATISTICAL APPROACHES FOR INTEGRATING DNA SEQUENCING INTO CLINICAL CARE

Dayne Lewis Filer

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the in the School of Medicine.

Chapel Hill
2020

Approved by:

Bradford C Powell

Kirk C Wilhelmsen

Stan C Ahalt

Yun Li

Neeta L Vora

Page intentionally left blank.

Page intentionally left blank.

ABSTRACT

Dayne Lewis Filer: Tools and statistical approaches for integrating DNA sequencing into clinical care
(Under the direction of Kirk C Wilhelmsen)

Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here.

Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here.

Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here. Some abstract goes here.

Page intentionally left blank.

To my loving and endlessly supportive wife and parents.

Page intentionally left blank.

ACKNOWLEDGEMENTS

Some abstract goes here.

Page intentionally left blank.

PREFACE

This will be the preface describing the work, who was involved, and the pending publication of the two chapters.

TABLE OF CONTENTS

Page intentionally left blank.

LIST OF FIGURES

Page intentionally left blank.

LIST OF TABLES

Page intentionally left blank.

LIST OF ABBREVIATIONS

CNV copy number variant

ES exome sequencing

cfDNA cell-free DNA

DNA deoxyribonucleic acid

MC multiplexed capture

IC independent capture

PMAR proportion of minor allele reads

cfES cell-free DNA exome sequencing

cfGS cell-free DNA genome sequencing

NIPT noninvasive prenatal testing

RNA ribonucleic acid

mRNA messanger RNA

PCR polymerase chain reaction

SBS sequencing by synthesis

dNTP deoxynucleotide triphosphate

ddNTP dideoxynucleotide triphosphate

ATP adenosine triphosphate

BCL binary base call (file format)

BLAST basic local alignment search tool

BWT Burrows-Wheeler Transform

UMI unique molecular identifer

SNV single nucleotide variant

LOH loss of heterozygosity

CHAPTER 1

Introduction

1.1 Outline

1.2 Human genetics primer

1.2.1 Discovery of DNA and the central dogma

The discovery of deoxyribonucleic acid (DNA) took roughly 100 years of work, and has fundamentally changed how we view ourselves, society, and life. The study of genetics begins with the study of peas and the discovery of inheritance by Gregor Mendel in the middle of the 19th century.[1] Shortly after Mendel's work, Friedrich Miescher isolated "nuclein" from lymphocytes noting the uniquenely high proportion of phosphorus in the form of phosphoric acid.[2–4] Albrecht Kossel and Albert Neumann furthered Miescher's work by identifying the four bases and renaming "nuclein" deoxyribonucleic acid.[5] Walther Flemming first described mitosis (the division of cells) showing the doubling and separation of chromosomes.[6] Theodor Boveri and Walter Sutton independently discovered meiosis, establishing chromosomes as the vehicle for inheritance (i.e. the "chromosome theory of inheritance").[7–9]

Despite early suggestions of chromatin containing DNA by Kossel and Neumann, many believed proteins and not DNA coded the fundamental information for inheritance. Oswald Avery, Collin MacLeod, and Maclyn McCarty published the first experiments to establish DNA carries the hedidary code using *Diplococcus pneumoniae*.[10] Erwin Chargaff rightly believed the work by Avery *et al.* and went on to discover equal proportions of adenine/thymine guanine/cytosine ("Chargaff's rule") which disproved the tetranucleotide hypothesis and laid the groundwork for the double helical model.[11] In the early 1950's Roslind Franklin started using X-ray crystallography to study the structure of DNA, producing the first images showing the double helical form.[12] Watson and Crick were given Rosalind's images without her knowledge or permission, allowing them to perform the final work to establish the structure of DNA.[13] Crick went on to establish the Central Dogma of Molecular Biology.[14,15]

The Central Dogma of Molecular Biology describes the process by which DNA codes for the proteins that build and sustain eukaryotic life. To produce proteins, ribonucleic acid (RNA) polymerase first transcribes the DNA message into single-stranded RNA molecules (messanger RNA, mRNA). The mRNA, after post-transcriptional modifications including possible splicing (reorganization), is then translated into a polymer of amino acids by ribosomal RNA complexes. Amino acid polymers, also known as polypeptides or peptide chains, form the primary structure of proteins. Therefore, modifications to DNA have profound impacts on cellular and organismal function.

1.2.2 Types of genetic variation

Humans are diploid organisms, meaning we have two copies of each chromosome. Under normal circumstances, we receive one set of chromosomes each from our biological mother and father. Humans have 46 chromosomes (23 from each parent), including 21 autosomes (chromosomes 1-22) and two sex chromosomes (X and Y).

1.3 DNA Sequencing

1.3.1 First-generation sequencing

Using lessons learned from previous RNA sequencing efforts, the first DNA sequencing techniques arose in the 1970s with Sanger's original plus-minus approach,[16] the Maxam-Gilbert chemical cleavage approach,[17] and Sanger's chain termination approach.[18]

Maxam-Gilbert sequencing works by cleaving DNA sequences at specific base pairs using specific chemical reactions. Before cleaving, radioactive phosphorus is incorporated into the 5 prime terminus of the DNA fragment to be sequenced. The fragment is then cleaved randomly in four separate reactions: at either G, G and A, C, or C and T. The cleaved radio-labeled fragments from each of the four reactions are then size-separated and visualized on a polyacrylamide gel.

Sanger sequencing (chain termination) was the first sequencing by synthesis (SBS) approach. Similar to Maxam-Gilbert sequencing, the target DNA fragment is replicated by the polymerase chain reaction (PCR) in four separate conditions. Each condition contains an equimolar mix of the four deoxynucleotide triphosphate (dNTP, DNA bases) molecules and a small amount of a single radio- or fluorescently-labeled dideoxynucleotide (ddNTP). The PCR reaction cannot proceed after the incorporation of a ddNTP, so each of the four reactions will contain synthesized fragments that stop at the same base. Again, the four reactions are size-separated and visualized on a polyacrylamide gel.

1.3.2 Second-generation sequencing

In the following decade Nyrèn and Lundin discovered an enzymatic method for detecting the incorporation of a new base during sequencing.[19] Pyrophosphate is released when dNTPs are incorporated into a DNA polymer; Nyrèn added two enzymes to the synethesis reaction: (1) ATP sulfurylase, which converts pyrophosphate into ATP; (2) luciferase, which converts ATP molecules into light. After fixing the DNA template to a solid phase, sequencing is performed by watching for light reactions after adding a single base at a time. Pyrosequencing struggles with sequencing over homopolymers (contiguous runs of the same base), with poor performance after 4-5 identical bases.[20]

The next significant breakthrough came in the early 2000s when Li *et al.* developed the first photocleavable fluorescent nucleotide.[21] The novel nucleotides use a fluorescent tag to block the 3 prime hydroxyl group, which can be cleaved using a specific wavelength of light. This allows for SBS with a “reversible termination” of synthesis after each base incorporation. The reversible terminators, in conjunction with the development of glass-bound colony expansion,[22] laid the groundwork for the Solexa system (acquired by Illumina) which currently dominates the sequencing field.[23]

Illumina sequencing works by creating clusters of identical DNA fragments bound to a glass plate (“flow cell”), then performing SBS using fluorescent reversible terminators. To perform Illumina sequencing, specific sequencing adapters are ligated onto short DNA fragments to: (1) bind

DNA fragments to the flow cell; (2) initiate amplification; (3) optionally identify the fragment source. The flow cell contains a “lawn” of two short oligos bound to the glass surface; the fragments have homology to either the forward or reverse adapter. The sequencing library containing the ligated forward and reverse adapters are added to the flow cell, where they hybridize to the lawn. Once bound, polymerase is added and the bound oligo is extended using the hybridized DNA fragment as a template. The original template is then washed away, leaving complementary sequences bound to the flow cell. The free adapter then folds over to hybridize to its complement oligo, forming a bridge, and polymerase fills in the oligo to form a double-stranded fragment (bridge amplification). The double-stranded fragment is denatured, leaving two single stranded fragments bound to the flow cell. Bridge amplification is repeated until each cluster contains hundreds of the same the fragment. The reverse fragments are then cleaved from the flow cell, and the clusters are sequenced by detecting the incorporation of fluorescent reversible terminators. Each cluster is tracked as basepairs are incorporated, giving the final DNA sequence.

1.3.3 Processing short-read sequencing data

With the advancements in sequencing chemistry, we now have the ability to sequence great amounts of DNA cheaply. However, the massively parallel sequencing modalities only sequence small fragments of DNA (typically 50 to 500 basepairs in length) often using a “shotgun” approach – “shotgun” referring to sequencing a randomly fragmented sample rather than a known locus. Therefore, the nature of short-read shotgun sequencing requires robust computational approaches to process and contextualize sequence data for millions of DNA fragments.

Here, I will give an overview of processing sequencing data for a species with an established reference genome. Processing short-read sequencing data follows the following general steps:

1. pre-processing to remove artificially added sequence (sequencing adapters, sample barcodes, etc.) and create FASTA/FASTQ[24,25] output;
2. map individual reads to their original location in the reference genome and create Sequence Alignment Map (SAM/BAM)[26] output;
3. optional post-mapping quality control;

4. variant identification;

5. variant filtering and interpretation.

The pre-processing step depends entirely on the sequencing chemistry and machinery uesd. Illumina sequencers produce binary base call (BCL) files containing all of the raw base call and quality information from the sequencing run. BCL files contain the adapter sequence (including sample barcode sequence and molecular index sequence when used in the library generation), which must be removed prior to mapping. Due to the capacity of modern sequencing machines, most often each lane of the flow cell will contain multiple samples. By convention, reads from each sample are separated into individual FASTQ files. Separating reads by sample must occur prior to discarding the adapater sequence information. Illumina currently provides the `bcl2fastq` command line tool for performing all of the requisite tasks to produce sample-specific FASTQ files with molecular index information when applicable.

The process of mapping individual reads (query sequences) to a reference sequence requires (1) finding the correct starting point in the reference sequence, and (2) accounting for substitutions, insertions, and deletions in the query sequence. Smith and Waterman published the first algorithm meeting both requirements, using dynamic programming on a substitution matrix[27] based on the inital work of Needleman and Wunsch.[28] The Smith-Waterman algorithm requires user-defined scores for matches, mismatches, and gaps (insertions/deletions); the algorithm will find the best possible match with the given scoring system, but requires $O(mn)$ compute time where m and n represent the length of the reference and query sequences.

To reduce the complexity of the problem, Altschul *et al.* developed the basic local alignment search tool (BLAST).[29] BLAST works by breaking the query sequences into a hash table of all possible k-mer sub-sequences and searching the reference sequence for non-gap matches above some threshold. For pairs of matches, BLAST extends the sequence to refine the candidate pool, and then finalizes the best candidates using the Smith-Waterman algorithm. Many other tools take similar hash table approaches, including hashing the reference sequence rather than the query sequence.[30]

Modern alignment algorithms have futher improved efficiency by exploiting the Burrows-Wheeler Transform (BWT).[30,31] The BWT creates a quickly search-able compressed representation of the reference sequence (roughly 1 gigabyte for the complete human genome), which search algorithms can hold in memory for even greater search efficiency.[32] The various BWT-based algorithms differ primarilly on how they handle mismatches[30] For DNA sequencing, the Burrows Wheeler alignment tool (BWA) developed and subsequently refined by Li and Durbin in 2009 remains the *de facto* industry standard.[26,33]

Post-alignment processing prepares the mapped reads for variant calling. Artificial duplicate reads can create bias in downstream variant calling, and deserve careful consideration. Tw types of artificial duplicates can occur with Illumina sequencing: (1) PCR duplicates, (2) technical (optical and cluster) duplicates. In large randomly-fragmented libraries sequenced to moderate depth, duplicate reads are much more likely to represent artificial than true duplicates. Virtually all sequencing library preparation protocols include PCR amplification, producing artificial duplicate reads. Using non-paterned flow cells, the image processing software my incorrectly identify large/odly shaped clusters as two separate clusters. With patterned flow cells, occasionally the same template can "jump" into an adjascent cluster.

In deeply-sequenced libraries with low complexity, we are more likely to observe true read duplicates. Without including unique molecular identifiers (UMIs) in the adapter sequence, we have no way of distinguishing true versus artificial duplicate reads. A UMI is a short (generally 6-12 basepairs) sequence of random bases; all PCR duplicates will contain the same UMI sequence. The exceedingly low probability of two true read duplicates having the same UMI allows properly controlling for artificial duplicates without removing true duplicates.

In addition to removing duplicate reads, the GATK best practices pipeline suggests adjusting the base quality scores prior to variant calling.[34,35] GATK provides the BaseQualityScoreRecalibration tool, which uses machine learning models to correct for known systematic errors in sequencing.

With the final set of aligned reads, we move to identifying deviations (variants) from the reference sequence. Numerous tools exist to perform variant calling; I will discuss the general approaches to calling the different types of variants, highlighting commonly-used algorithms.

Calling single base substitutions – single nucleotide variants (SNVs) – relies fundamentally on counting alleles at each locus. At minimum, the statistical models incorporate the quality of each base call and assumptions about sequencing error rates, e.g. the samtools mpileup/bcftools call programs.[36] GATK previously provided a similar tool, implementing a simple Bayesian genotype likelihood model,[34,35,37] but has moved currently to a haplotype-based calling algorithm (HaplotypeCaller).[38] HaplotypeCaller works by (1) identifying "active" regions containing plausible variants, (2) building possible haplotypes in the active regions using de Bruijn-like graphs, (3) assigning haplotype likelihoods to reads, and (4) calculating genotype likelihoods incorporating the estimated haplotype information. The idea for using haplotype estimates in genotype calling originated with the freebayes algorithm.[39] The above tools all use very similar approaches to call small insertions and deletions (indels). Development continues actively in SNV/indel variant identification, and performance between algorithms predictably differs with condition.[40,41]

Calling larger structural variation from short-read sequencing poses greater difficulty. SNVs and indels exist within single reads; therefore, we can view and count them directly. We cannot directly view variation which spans lengths greater than our read (or read pair) length. To identify larger variation, calling algorithms attempt to identify some combination of the following tw signals: (1) relative changes in sequencing depth (read depth); (2) paired read insert size and orientation (paired end mapping).

Read-depth methods, e.g. CNVnator,[42] work by building statistical models utilizing the relative sequencing depth across the genome. The depth bias introduced by the capture step in targeted sequencing necessitates comparing to a set of control samples, e.g. ExomeDepth,[43] rather than calculating the relative depth across the genome. Paired-end mapping methods identify sets of reads with insert sizes outside a specified range, indicating insertions or deletions, and reads with the incorrect orientation suggesting genomic rearrangements.[44] The Lumpy algorithm[45] utilizes both the read depth and paired end mapping approaches for greater detection sensitivity. The ERDS algorithm[46] combines read depth information with allele ratios when possible.

1.4 Medical genetics primer

1902 – Mendel's theories were finally associated with a human disease by Sir Archibald Edward Garrod, who published the first findings from a study on recessive inheritance in human beings in 1902. Garrod opened the door for our understanding of genetic disorders resulting from errors in chemical pathways in the body.

Late 1940s – Barbara McClintock discovered the mobility of genes, ultimately challenging virtually everything that was once thought to be. Her discovery of the "jumping gene," or the idea that genes can move on a chromosome, earned her the Nobel Prize in Physiology.

CHAPTER 2

A novel copy number variant algorithm

2.1 Introduction

In human genetics, individuals normally have two copies of each locus in the genome (one inherited from each parent). Deviations from the normal diploid state, known broadly as copy number variation, can cause phenotypic changes and Mendelian disorders. Technologies, e.g. microarray, exist for reliably detecting large (greater than 100 kilobases) copy number variants (CNVs). Over the last decade, the availability short-read DNA sequencing compelled numerous efforts to identify and characterize smaller variants. Sequencing cost, data burden, and the problem of classifying intronic and non-coding variants have led to exome sequencing (ES) as the preferred clinical sequencing modality. ES analysis most often focuses on identifying pathogenic single-nucleotide variants and insertion/deletions. CNV analysis has demonstrated limited improvement in diagnostic yield,[47] but existing data/analysis lacks power to detect exon-level variation.[48,49]

Current analytic methodologies adequately detect large CNVs, but require large amounts of data and lack resolution for intragenic exon-level variation.[43,50–52] The prevalence and clinical importance of exon-level CNVs remains largely unknown due to inadequate power in ES studies and limited access to clinical genome sequencing data. Recent work on a subset of 1507 genes suggests intragenic CNVs account for 1.9% of total variants, but 9.8% of pathogenic variants.[53] Additionally, the authors demonstrated 627/2844 (22%) of identified CNVs spanned a single (598) or partial (29) exon.[53]

Targeted sequencing requires capturing the desired loci (e.g. exons) using sequence-specific oligonucleotide baits. The differential efficiency of baits, even when carefully designed and balanced, leads to variable read-depth across the exome. The GC content and length of targeted fragments both contribute to the observed variable read-depth;[54] most ES analysis platforms incorporate correction for GC content and exon length.[55] The variable read-depth in ES precludes the single-sample window-smoothing approaches successfully applied in GS data,[56] therefore we must rely on comparative analysis for interrogating copy number.

Comparing multiple samples, each captured independently, compounds the variable read-depth problem. The capture probability for each exon correlates between samples but with high variability.[43] In other words, we can gain information from similarly captured samples, but independent captures introduce significant noise. ExomeDepth attempts to circumvent the capture-to-capture variation by identifying a subset of samples from a large pool with low inter-sample variability.[43] Alternatively, CoNIFER,[52] XHMM,[50] and CODEX[51] use a latent factor model with spectral value decomposition to remove systematic noise, presumably introduced by capture-to-capture variation. These methods generally require very large sample sizes, and often still lack power for exon-level resolution (e.g. CODEX defines a “short” CNV as spanning 5 contiguous exons).

Here, we explore how multiplexing the capture across samples reduces inter-sample variance, increasing the power to detect CNVs. We also introduce our own algorithm, mcCNV (“multiplexed capture CNV”), specifically designed to utilize multiplexed capture exome data for estimating exon-level variation without prior information.

2.2 Methods

2.2.1 Exome sequencing

We performed sequencing on human samples of purified DNA obtained from the Wilhelmsen laboratory collection, the NCGENES cohort,[57] and the Coriell Institute in compliance with the UNC Institutional Review Board. We also utilized existing read-level data from the NCGENES[57] project. We compared the performance of two capture platforms: (1) Agilent SureSelect XT2 (multiplexed capture)/Agilent SureSelect XT (independent capture); (2) Integrated DNA Technologies (IDT) xGen Lockdown Probes. We utilized Human All Exome v4 baits (Agilent) and Exome Research Panel v1 baits (IDT). All captures performed according to manufacturer protocol, with the following exceptions: (1) we multiplexed 16 samples versus the recommended 8 for the XT2 protocol for some pools; (2) for Pool2, we performed the fragmentation step 5 times, to test whether a more uniform fragment length distribution would improve capture.

All sequencing performed with Illumina paired-end chemistry. We aligned paired reads to hg19v0 (GATK resource bundle) using BWA-MEM[58] and removed duplicate reads using Picard tools. We then used our novel R package, mcCNV, to count the number of overlapping molecules (read-pairs) per exon. For inclusion, we required properly-paired molecules with unambiguous mapping for one read and mapping quality greater than or equal to 20 for both reads. Full Snakemake[59] pipeline provided in supplemental materials. Table 2.1 provides an overview of the exome sequencing included.

2.2.2 Genome sequencing

For the 16 samples in the "WGS" pool, we performed genome sequencing to an average 50x coverage. We followed Trost et al. recommendations for making read-depth based CNV calls.[60] Briefly, we mapped paired-reads identical to our targeted sequencing data. We then interrogated the read depth interquartile range using samtools depth,[26] recalibrated base-quality scores and called sequence variants using GATK,[37] and called copy number variants using the ERDS[46] and cnvpytor (updated implementation of CNVnator)[42] algorithms. Full Snakemake[59] pipeline provided in supplemental materials.

2.2.3 Simulating targeted sequencing

To simulate targeted capture, we represent the capture process as a large multinomial distribution defining the probability of capture at each target. We use an alternate definition of copy state, such that 1 represents the normal diploid state. Let N represent the total number of molecules (read pairs) and $e_j \in \mathbb{E}$ represent the probability of capturing target j, then for each subject, i:

1. Randomly select $s_{ij} \in \mathbb{S}_i$ from $S = \{0.0, 0.5, 1, 1.5, 2\}$ as the copy number at target j

2. Adjust the subject-specific capture probabilities by the copy number, $\mathbb{E}_i = \frac{\mathbb{E} \odot \mathbb{S}_i}{\sum_j \mathbb{E} \odot \mathbb{S}_i}$

3. Draw N times from Multinomial($\mathbb{E}_i$), giving the molecule counts at each target j for sample i, $c_{ij} \in \mathbb{C}_i$

We provide functionality within the mcCNV R package for producing reproducible simulations.

2.2.4 mcCNV algorithm

The mcCNV algorithm was adapted from the sSEQ method for quantifying differential expression in RNA-seq experiments with small sample sizes.[61] Yu et al. provide detailed theoretical background of the negative binomial model and using shrinkage to improve dispersion estimates. The mcCNV algorithm adjusts the sSEQ probability model by adding a multiplier for the copy state:

$$C_{ij} \sim \mathcal{NB}(f_i s_{ij} \hat{\mu}_j, \tilde{\phi}_j / f_i)$$

where the random variable C_{ij} represents observed molecule counts for subject i at target j, f_i is the size factor for subject i, s_{ij} is the copy state, μ_j is the expected mean under the diploid state at target j, and $\tilde{\phi}_j$ is the shrunken phi at target j. We observe c_{ij} and wish to estimate s_{ij}, $\hat{s}_{ij}$. Initialize by setting $\hat{s}_{ij} = 1$ for all i, j. Then,

1. Adjust the observed values for the estimated copy-state,

$$c'_{ij} = \frac{c_{ij}}{\hat{s}_{ij}}.$$

2. Subset c'_{ij} such that $c'_{ij} > 10,\ \hat{s}_{ij} > 0$

3. Calculate the size-factor for each subject

$$f_i = \text{median}\left(\frac{c'_{ij}}{g_j}\right),$$

   where g_j is the geometric mean at each exon.

4. Use method of moments to calculate the expected dispersion

$$\hat{\phi}_j = \max\left(0, \frac{\hat{\sigma}_j^2 - \hat{\mu}_j}{\hat{\mu}_j^2}\right)$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the sample mean and variance of c'_{ij}/f_i.

5. Let J represent the number of targets. Shrink the phi values to

$$\tilde{\phi}_j = (1-\delta)\hat{\phi}_j + \delta\hat{\xi}$$

such that

$$\delta = \frac{\sum\limits_j \left(\hat{\phi}_j - \frac{1}{n_j}\sum\limits_j \hat{\phi}_j\right)^2 / (J-1)}{\sum\limits_j \left(\hat{\phi}_j - \hat{\xi}\right)^2 / (n_j - 2)}$$

and

$$\hat{\xi} = \underset{\xi}{\operatorname{argmin}} \left\{ \frac{d}{d\xi} \frac{1}{\sum\limits_j \left(\hat{\phi}_j - \xi\right)^2} \right\}.$$

6. Update $\hat{s}_{ij}$,

$$\underset{s \in S}{\operatorname{argmax}} \left\{ \mathcal{L}(s|c_{ij}, f_i, \hat{\mu}_j, \tilde{\phi}_j) \right\}$$

where $S = \{0.001, 0.5, 1, 1.5, 2\}$.

7. Repeat until the number of changed states falls below a threshold or a maximum number of iterations is reached.

8. After convergence, calculate p-values for the diploid state, $\pi_{ij} = \Pr(s_{ij} = 1)$.

9. Adjust p-values using the Benjamini–Hochberg procedure[62] and filter to a final call-set such that adjusted p-values fall below some threshold, α.

2.3 Results

2.3.1 Multiplexed capture reduces inter-sample variance

ES requires using molecular baits to "capture" the exonic DNA fragments during the library preparation (prior to sequencing). Most laboratories capture each sample individually. The capture efficiency varies with timing, temperature, and substrate concentrations, making identical capture reproduction impossible. Alternatively, one could multiplex (pool) samples prior to capture, cap-

Table 2.1: Summary of whole-exome sequencing. "pool" indicates the name of the pool of samples; "capture" indicates the capture platform for the pool; "N" gives the number of samples in the pool; "medExon" gives the pool median of the subject median mapped molecule count per exon; "medTotal" gives the median by pool of total mapped molecule counts per subject; "minTotal" and "maxTotal" give the minimum and maximum total mapped molecules; "rsdTotal" gives the relative standard deviation (SD/mean*100) of total mapped molecules. † indicates captures were performed independently on each sample within the pool, otherwise captures were multiplexed across all samples within the pool.

pool	capture	N	medExon	medTotal	minTotal	maxTotal	rsdTotal
IDT-IC†	IDT	16	143	55,149,058	37,453,015	85,138,915	22.4
IDT-MC	IDT	16	93	29,772,684	16,674,468	118,147,912	64.2
IDT-RR	IDT	16	272	79,079,629	61,289,322	120,147,888	22.9
NCGENES†	Agilent	112	93	24,451,245	12,749,793	68,565,471	27.6
Pool1	Agilent	16	56	13,265,614	8,911,132	17,324,903	18.5
Pool2	Agilent	16	86	21,076,056	4,585,195	27,846,146	27.6
SMA1	Agilent	8	56	12,256,002	11,051,840	13,600,697	6.2
SMA2	Agilent	8	25	5,622,040	4,904,000	6,545,360	10.4
WGS	Agilent	16	196	46,406,224	36,496,097	65,200,410	16.4

turing the pool of samples simultaneously. Here we profile the inter-sample variance of individual capture versus multiplexed capture.

A multinomial process provides a logical framework for modeling targeted capture, each target represented by an individual outcome. We can estimate the multinomial probability simplex for an exome capture by dividing the observed counts at each exon by the total mapped reads for the exome. The dirichlet distribution, conjugate prior to the multinomial, defines distributions of probability simplexes. The dirichlet distribution is parameterized by $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, where the expected probability for outcome i is given by $\alpha_i/\alpha_0,\ \alpha_0 = \sum \boldsymbol{\alpha}$. If $\boldsymbol{\pi}$ is a probability simplex drawn from a dirichlet with $\boldsymbol{\alpha}$, then the variance of $\boldsymbol{\pi}$ is inversely proportional to α_0. Therefore, we can approximate the inter-sample variance by fitting the dirichlet distribution to each pool and interrogating the mean α.

Using multiplexed capture, we sequenced 3 16-sample pools and 2 8-sample pools with Agilent baits and 2 16-sample pools with IDT baits (Table 2.1). To compare to individually-captured Agilent data, we randomly-selected 5 16-sample pools from the NCGENES cohort. For numeric stability, we subset to exons with at least 5 and no greater than 2000 counts across all samples within a pool. We then used a Newton-Raphson algorithm[63] to fit the dirichlet distribution to each

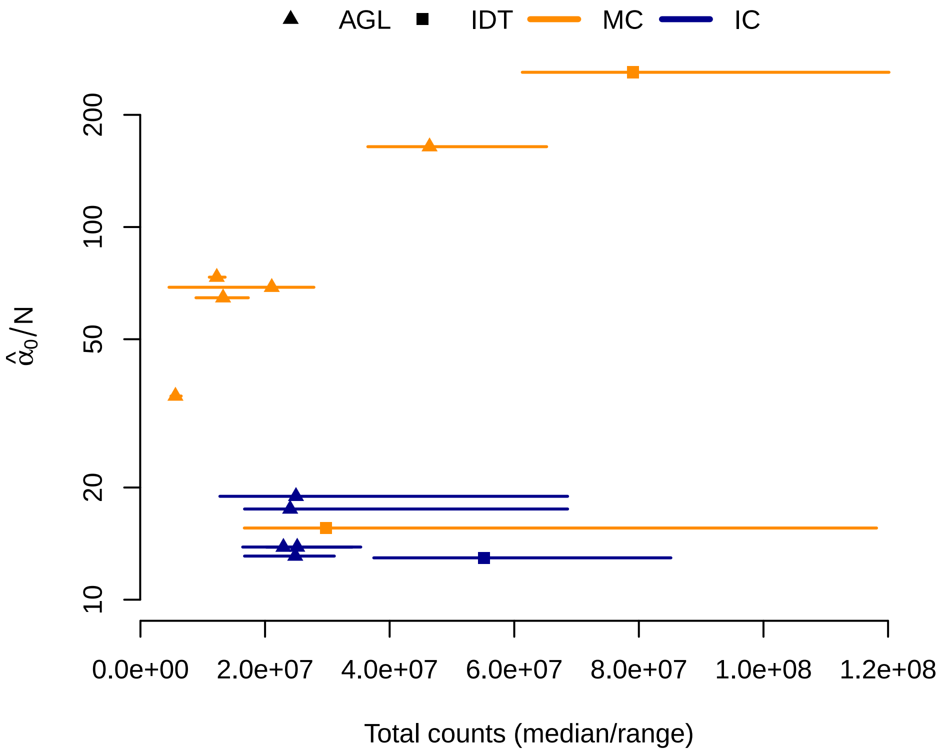

Figure 2.1: Multiplexed capture (MC) decreases variance with respect to independent captures (IC), as estimated by fitting the dirichlet distribution. Total counts/sample given on the horizontal axis; mean α given on the vertical axis. α_0 is inversely proportional to inter-sample variance. Each line/point represents a single pool. The point indicates the median total counts across the pool, with the range given by the line. Orange indicates a multiplexed capture; blue indicates independent captures. Triangles indicate pools using Agilent (AGL) capture; squares indicate Integrated DNA Technologies (IDT).

pool; all pools converged to stable estimates. We found, with one exception, multiplexed capture pools had greater α_0 of their independently-captured counterparts (Figure 2.1).

The multiplexed pool without decreased inter-sample variance, IDT-MC, had a much larger spread in sequencing depth across the pool (Table 2.1, Figure 2.1). Looking at the total mapped molecules, the IDT-MC pool had a relative standard deviation of 64.2%, over double the next highest pool. We hypothesized the absent reduction in variation stemmed from poor library balance during the multiplexing step. We subsequently captured a new pool using the same DNA input, IDT-RR, and found comparable reductions in inter-sample variance (the pool with the highest α_0 in Figure 2.1).

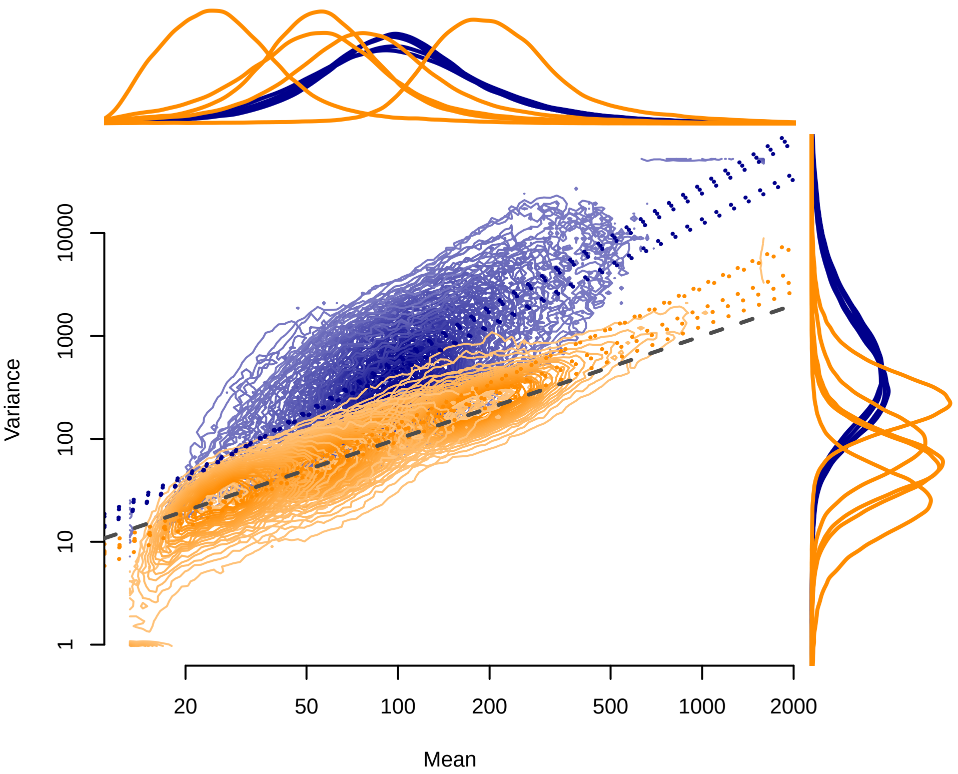

Figure 2.2: Mean-variance relationship for Agilent (AGL) pools. Mean counts per exon given on the horizontal axis; mean variance per exon given on the vertical axis. Contours show the distribution of points by pool. Dotted lines show the ordinary least squares regression fit. Orange indicates multiplexed capture pools; blue indicates independently captured pools. The dashed gray line represents the 1:1 relationship expected under a Poisson process. Lines above the plot show the density of mean values by pool; lines to the right of the plot show the density of variance values by pool.

Examining the mean-variance relationship demonstrated the same inter-sample variance reduction suggested by the dirichlet parameter estimates (Figures 2.2 and 2.3). The Agilent pools (Figure 2.2) segregated cleanly, with less dispersion in the multiplexed capture pools. Again, we found no variance reduction for the IDT-MC pool, overlapping with independently-captured IDT-IC pool (Figure 2.2). We did, however, observe near-complete reduction in dispersion for the better-balanced IDT-RR pool.

2.3.2 Multiplexed capture provides controls for ExomeDepth

ExomeDepth requires a set of control subjects, summed into a reference vector of counts at each exon. ExomeDepth provides functionality to select appropriate controls from a set of subjects, often requiring large numbers of subjects to identify appropriate controls. Smaller research groups

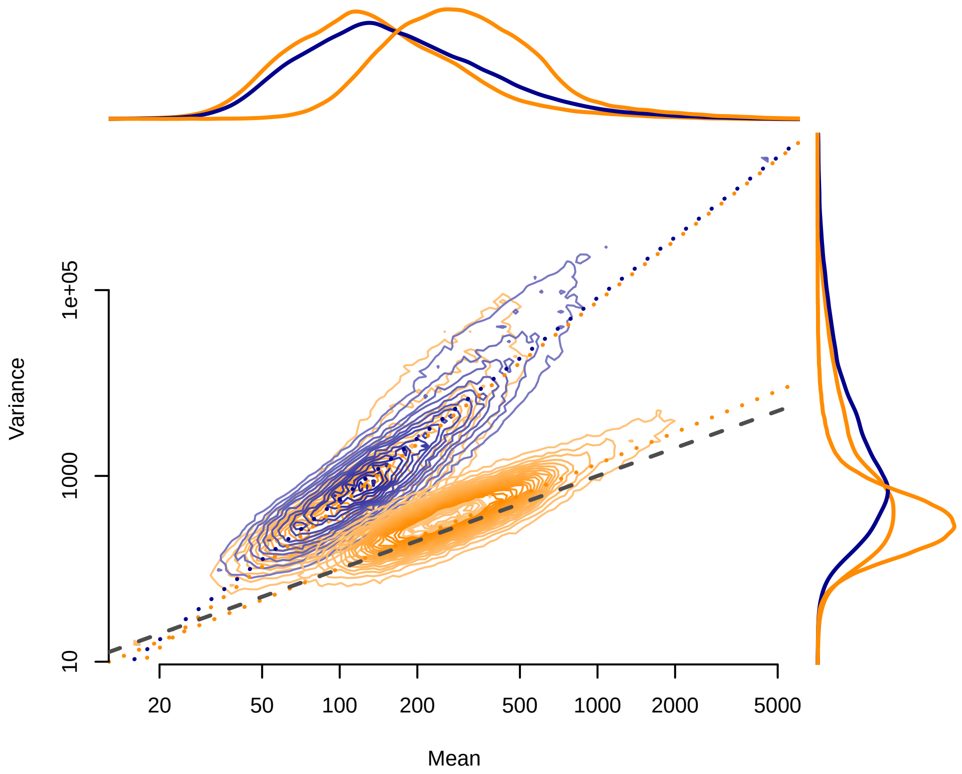

Figure 2.3: Mean-variance relationship for Integrated DNA Technologies (IDT) pools. Mean counts per exon given on the horizontal axis; mean variance per exon given on the vertical axis. Contours show the distribution of points by pool. Dotted lines show the ordinary least squares regression fit. Orange indicates multiplexed capture pools; blue indicates independently captured pools. The dashed gray line represents the 1:1 relationship expected under a Poisson process. Lines above the plot show the density of mean values by pool; lines to the right of the plot show the density of variance values by pool.

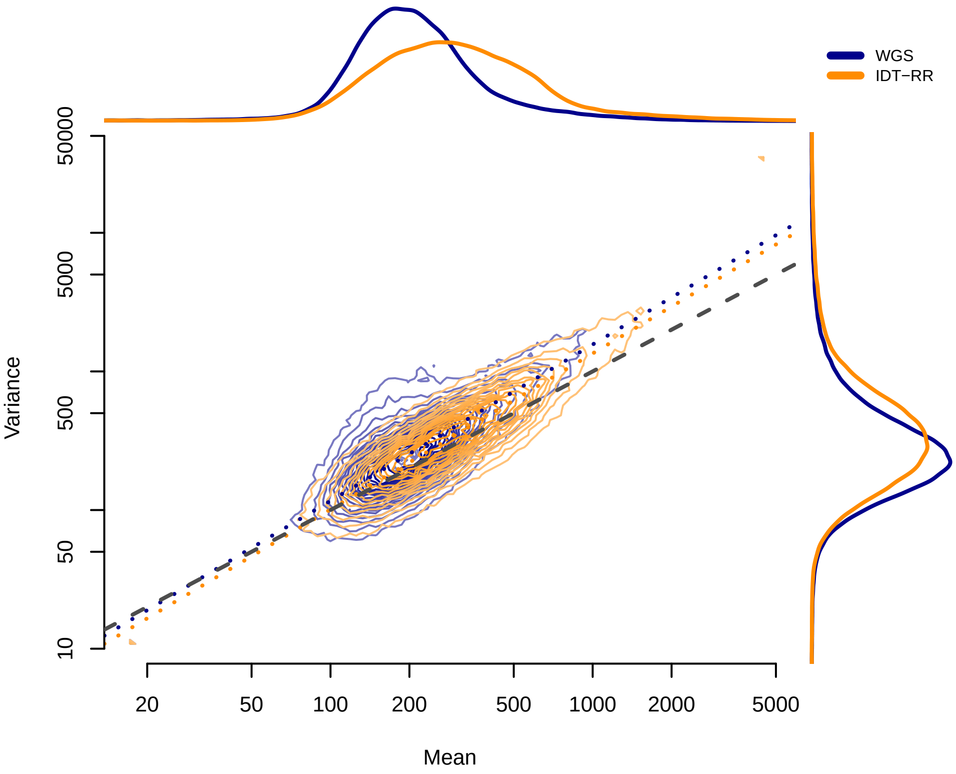

Figure 2.4: Comparison of mean-variance relationship between WGS pool (blue) and IDT-RR pool (orange). Mean count by exon given on horizontal axis; variance of exon counts given on horizontal axis. Dotted lines show the ordinary least-squares fit. Lines above plot show the distribution of mean values; lines to the right of the plot show the distribution of variance values.

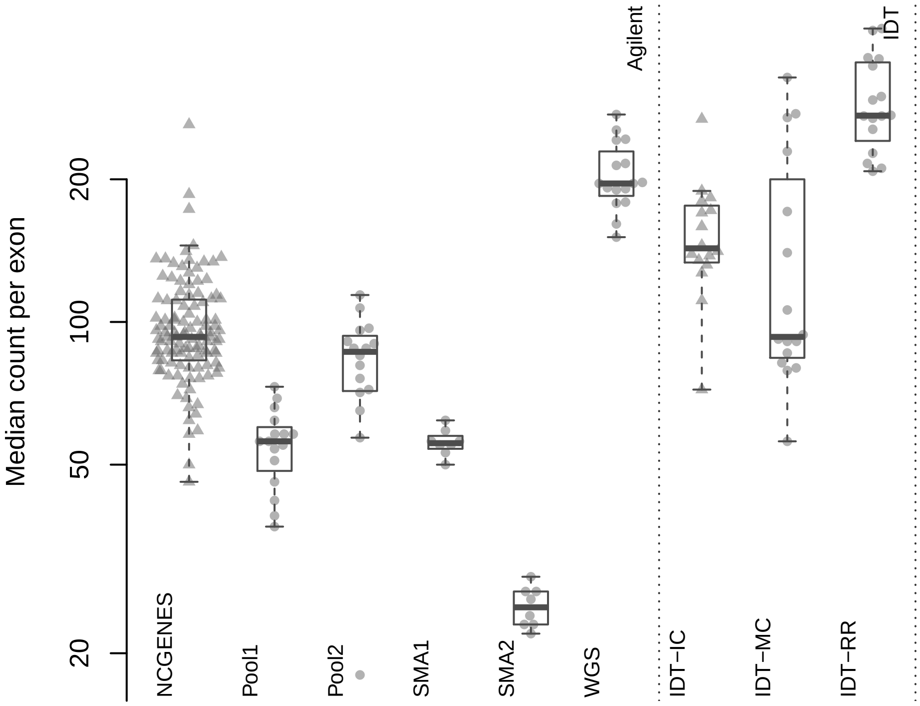

Figure 2.5: Median count per exon. Each point represents a single sample, with samples grouped by pool. Triangles indicate independently-captured samples; circles indiciate a single multiplexed capture within the pool. Dotted vertical line separates the two capture platforms.

and clinical laboratories may struggle building large databases of exomes, with the difficulty compounded by lot-to-lot variation and regular improvements to capture and sequencing chemistries. We wanted to know if the reduced inter-sample variance with multiplexed capture could provide an appropriate control set for ExomeDepth, eliminating the need for large databases of similarly-captured exomes. We found the reduced inter-sample variance with multiplexed capture leads to appropriate control selection for ExomeDepth (Figures 2.5). Pool2, where we repeated the initial fragmentation 5 times, did not perform as well as the other multiplexed pools. We also found two samples within the WGS pool did not correlate well with the rest of the pool.

When we looked at independently-captured subjects, we found appropriate control sets for most of the 112 NCGENES subjects (Figure 2.6). However, ExomeDepth only selected 12.2% of available samples as controls, on average (Figure 2.7). Similarly, with the independently-captured IDT-IC pool we find low control numbers for most samples. While possible to select the same number of controls but exhibit differing dispersion, we observed little difference in the dispersion

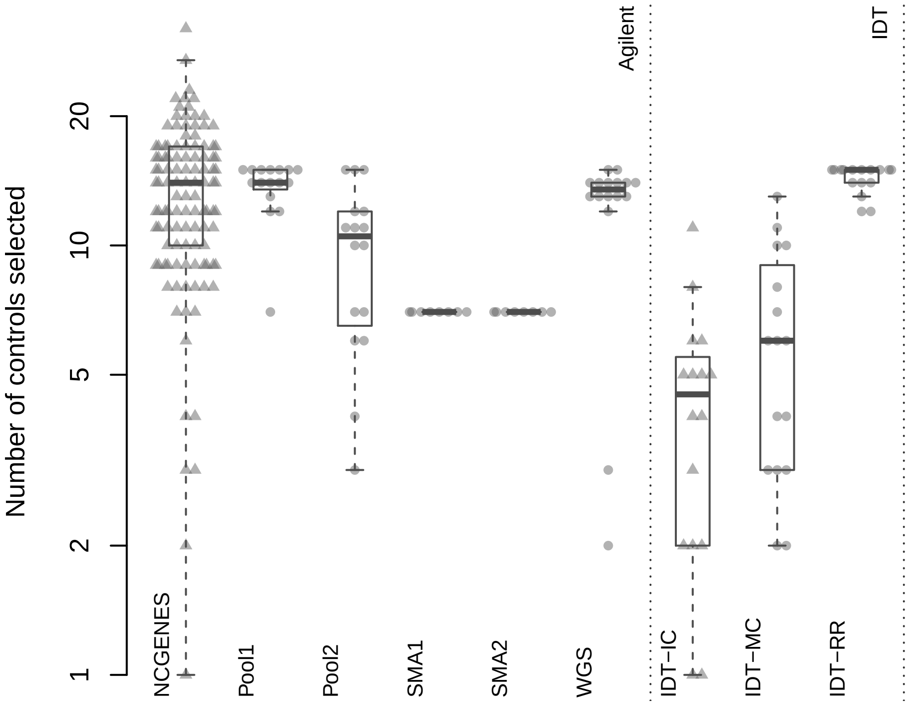

Figure 2.6: Total number of controls selected by ExomeDepth. Each point represents a single sample, with samples grouped by pool. Triangles indicate independently-captured samples; circles indiciate a single multiplexed capture within the pool. Dotted vertical line separates the two capture platforms.

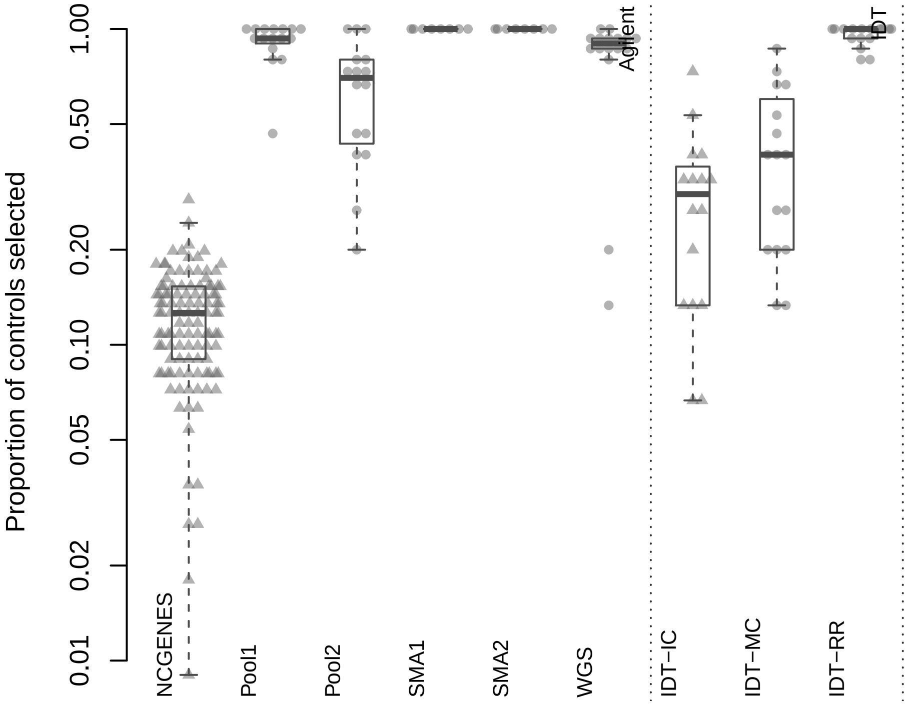

Figure 2.7: Proportion of available samples selected by ExomeDepth as a control. Each point represents a single sample, with samples grouped by pool. Triangles indicate independently-captured samples; circles indiciate a single multiplexed capture within the pool. Dotted vertical line separates the two capture platforms.

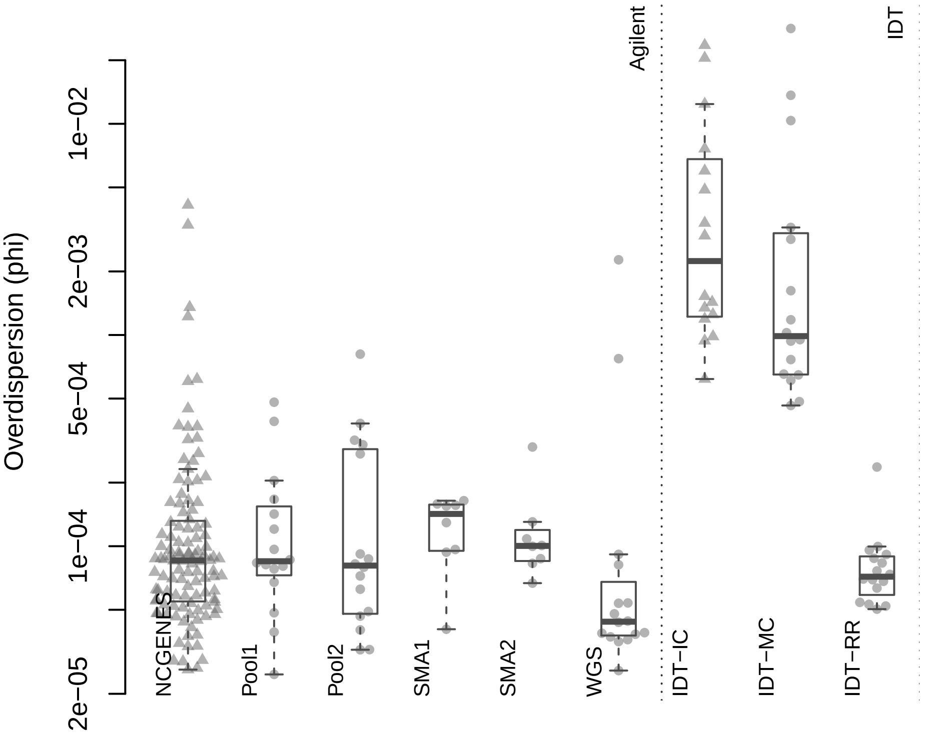

Figure 2.8: Estimated phi parameter from ExomeDepth. Each point represents a single sample, with samples grouped by pool. Triangles indicate independently-captured samples; circles indiciate a single multiplexed capture within the pool. Dotted vertical line separates the two capture platforms.

between independent and multiplexed capture (Figure 2.8). Overall, multiplexed capture provided appropriate controls for most samples tested, however an adequately-large set of available controls delivered comparable performance.

2.3.3 mcCNV & ExomeDepth perform comparably in simulation study

To compare our mcCNV algorithm and ExomeDepth, we created synthetic pools of data across different sequencing depths. Based on our observations with the real data, we selected the total number of molecules for each sample from a uniform distribution defined as a 30% window on either side of the specified depth; for example, for a specified depth of 10 million molecules, we drew the molecules per sample from 7 to 13 million molecules. For each depth ranging from 5 to 100 million molecules, we simulated 200 16-sample pools with single-exon variants. We allowed for homozygous and heterozygous deletions and duplications (0 to 4 copies), such that all variants were equally likely and the total variant probability was $1/1,000$. We used, as the starting capture probabilities ($\mathbb{E}$), the empiric capture probabilities observed by summing across the Pool1 pool.

We analyzed each of the 4,000 pools (200 replicates by 20 depths) using our algorithm and two iterations of ExomeDepth. For the first iteration of ExomeDepth, we used the default values for transition probability ($1/10,000$) and expected variant length (50 kb). For the second iteration, we used the true simulated variant prior for the transition probability ($1/1,000$) and an expected variant length of 1 kb. As expected, the sensitivity increased and false discovery rate decreased as the sequencing depth increased (Figures 2.9). In both comparisons, mcCNV demonstrated a superior false-discovery rate. When interrogating Matthew's correlation coefficient[64] and the sensitivity, we found mcCNV had marginal performance over ExomeDepth with default parameters and marginal performance under ExomeDepth with simulation-matched parameters. Table 2.2 provides the actual values.

2.3.4 mcCNV & ExomeDepth perform comparably on WGS pool

To establish a truth set on real data, we performed matched genome sequencing on the subjects included in the WGS pool. Following the best practices suggested by Trost et al. [60], we performed read-depth based CNV calling using the genome data. In line with recommendations by Trost et al.,

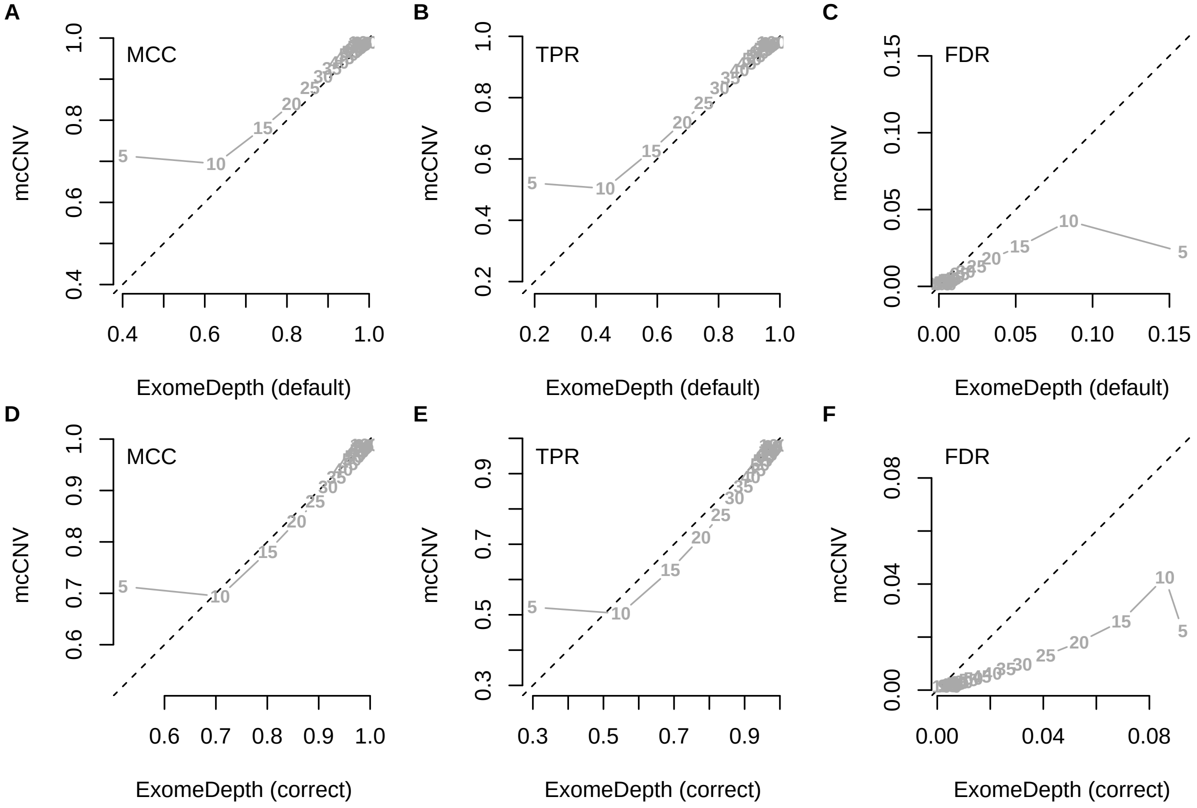

Figure 2.9: Algorithm performance comparing mcCNV and ExomeDepth on simulated exomes. (A-C) mcCNV versus ExomeDepth with default parameters, 1/10, 000 transition probability and 50 kb expected variant length. (D-F) mcCNV versus ExomeDepth with simulation-matched parameters, 1/1, 000 transition probability and 1 kb expected variant length. Numbered points indicate the simulated depth in millions of molecules. 'MCC' indicates Matthew's correlation coefficient; 'TPR' indicates true positive rate/sensitivity; 'FDR' indicates false discovery rate. Dashed black line shows the 1:1 relationship.

Table 2.2: Simulation results by algorithm. ED-def: ExomeDepth with default parameters; ED-sim: ExomeDepth with simulation-matched parameters. Values represent the mean over 200 simulations.

	MCC			TPR			FDR		
dep	mcCNV	ED-def	ED-sim	mcCNV	ED-def	ED-sim	mcCNV	ED-def	ED-sim
5	0.713	0.401	0.519	0.522	0.192	0.298	0.02230	0.15900	0.09260
10	0.694	0.628	0.708	0.503	0.431	0.549	0.04250	0.08450	0.08590
15	0.781	0.742	0.801	0.627	0.581	0.690	0.02600	0.05270	0.06940
20	0.840	0.811	0.857	0.719	0.682	0.777	0.01810	0.03420	0.05360
25	0.879	0.856	0.893	0.783	0.752	0.832	0.01310	0.02460	0.04090
30	0.907	0.889	0.918	0.831	0.804	0.872	0.00967	0.01750	0.03210
35	0.926	0.909	0.935	0.864	0.839	0.897	0.00807	0.01370	0.02600
40	0.941	0.927	0.948	0.891	0.869	0.917	0.00638	0.01060	0.02080
45	0.952	0.940	0.957	0.911	0.892	0.932	0.00527	0.00846	0.01680
50	0.961	0.950	0.965	0.927	0.910	0.944	0.00437	0.00701	0.01370
55	0.966	0.957	0.969	0.937	0.921	0.951	0.00377	0.00569	0.01180
60	0.972	0.963	0.974	0.947	0.933	0.959	0.00318	0.00517	0.00986
65	0.976	0.969	0.978	0.955	0.943	0.964	0.00290	0.00433	0.00837
70	0.978	0.972	0.980	0.960	0.949	0.968	0.00252	0.00381	0.00735
75	0.981	0.976	0.983	0.965	0.955	0.972	0.00212	0.00321	0.00625
80	0.983	0.978	0.985	0.969	0.960	0.975	0.00200	0.00294	0.00560
85	0.985	0.980	0.986	0.972	0.963	0.977	0.00181	0.00263	0.00491
90	0.987	0.982	0.987	0.975	0.967	0.979	0.00169	0.00243	0.00451
95	0.988	0.984	0.988	0.978	0.970	0.981	0.00156	0.00223	0.00393
100	0.989	0.985	0.989	0.980	0.973	0.982	0.00150	0.00195	0.00359

Table 2.3: Number of CNV calls by subject and algorithm for the 'WGS' pool. 'MC' indicates the mcCNV algorithm; 'ED' indicates the ExomeDepth algorithm; 'WG' indicates the overlap of ERDS/cnvpytor calls from matched whole-genome sequencing. Exons with any overlap of the repetitive and low-complexity regions, as defined in the Trost et al. manuscript,[60] omitted from analysis.

	Total			Duplications			Deletions		
subject	MC	ED	WG	MC	ED	WG	MC	ED	WG
NCG_00012	90	106	143	61	73	121	29	33	22
NCG_00237	82	101	165	50	64	129	32	37	36
NCG_00525	68	74	151	30	33	110	38	41	41
NCG_00593	45	58	142	22	28	81	23	30	61
NCG_00676	66	78	112	38	46	92	28	32	20
NCG_00790	5,156	2,204	121	19	37	92	5,137	2,167	29
NCG_00819	68	76	134	30	41	100	38	35	34
NCG_00840	78	92	157	44	52	115	34	40	42
NCG_00851	1,151	859	141	28	51	102	1,123	808	39
NCG_00857	59	75	119	10	15	81	49	60	38
NCG_00976	46	58	114	25	37	93	21	21	21
NCG_01023	59	95	143	32	60	113	27	35	30
NCG_01043	73	94	128	40	64	105	33	30	23
NCG_01076	36	57	105	7	22	78	29	35	27
NCG_01077	135	157	230	103	121	184	32	36	46
NCG_01117	95	101	154	72	78	129	23	23	25

we excluded from comparative analysis any exons overlapping repetitive or low-complexity regions (34,856 out of 179,250). We then compared the exome calls using mcCNV and ExomeDepth to the genome calls using the overlap of ERDS and cnvpytor. Table 2.3 lists the total calls by subject. Overall, mcCNV predicted the largest number of variants; however, 85.7% of predicted variants were deletions from two samples (NCG_00790 and NCG_00851). ExomeDepth also predicted a disproportionate number of deletions for NCG_00790 and NCG_00851, totaling 69.4% of calls.

Looking at the control selection, for NCG_00790 and NCG_00851 ExomeDepth only selected 2 and 3 controls, respectively. Furthermore, NCG_00790 and NCG_00851 had substantially higher dispersion than the rest of the pool (two outliers in Figure 2.8).

Recognizing the genome calls do not represent an accurate truth set, we looked at the ability of mcCNV and ExomeDepth to predict the genome calls. Due to the large number of deletions called for NCG_00790 and NCG_00851, both algorithms performed poorly in predicting the genome

Table 2.4: mcCNV (MC)/ExomeDepth (ED) calls for 'WGS' pool (used as prediction) versus the ERDS/cnvpytor calls from matched genome sequencing (used as truth). Calls are subdivided by duplications (DUP) and deletions (DEL). 'Full' gives performance across the full pool; 'Sub' gives the performance excluding the poorly correlated samples NCG_00790 and NCG_00851 (gray rows). 'MCC' is Matthew's correlation coefficient, 'TPR' is true positive rate/sensitivity, 'FDR' is false discovery rate, 'PPV' is positive predictive value, 'BalAcc' is balanced accuracy. Exons with any overlap of the repetitive and low-complexity regions, as defined in the Trost et al. manuscript,[60] omitted from analysis.

			MCC	TPR	FDR	PPV
ALL	Total	MC	0.185	0.335	0.897	0.1030
		ED	0.263	0.363	0.809	0.1910
	Sub	MC	0.487	0.345	0.311	0.6890
		ED	0.482	0.378	0.383	0.6170
DUP	Total	MC	0.396	0.236	0.334	0.6660
		ED	0.347	0.240	0.496	0.5040
	Sub	MC	0.404	0.246	0.333	0.6670
		ED	0.384	0.266	0.446	0.5540
DEL	Total	MC	0.180	0.639	0.949	0.0509
		ED	0.219	0.558	0.914	0.0861
	Sub	MC	0.683	0.661	0.294	0.7060
		ED	0.541	0.554	0.471	0.5290

calls (Table 2.4). When we excluded NCG_00790 and NCG_00851 from the analysis, mcCNV had comparable, but uniformly better performance. Both algorithms demonstrated greater power to detect deletions. Figures 2.10 and 2.11 show the call overlap between the three approaches, including and excluding NCG_00790 and NCG_00851, respectively.

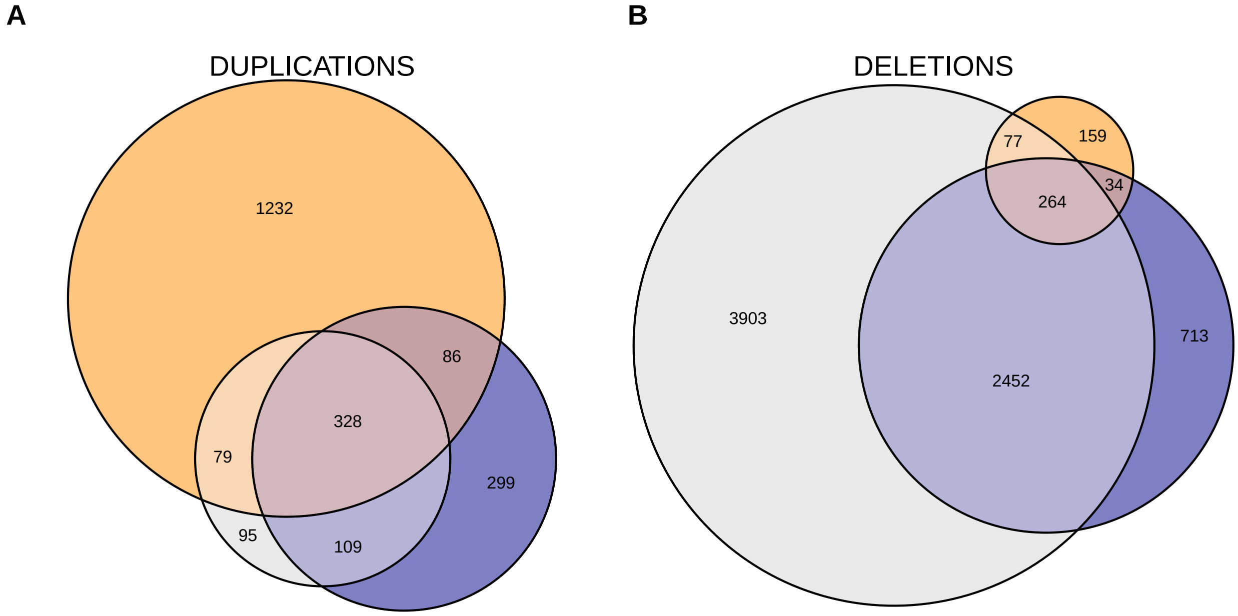

Figure 2.10: Copy number variant call concordance for the WGS pool. (A) predicted duplications; (B) predicted deletions. mcCNV in grey; ExomeDepth in blue; ERDS/cnvpytor in orange. Values within overlaps give the number of variants.

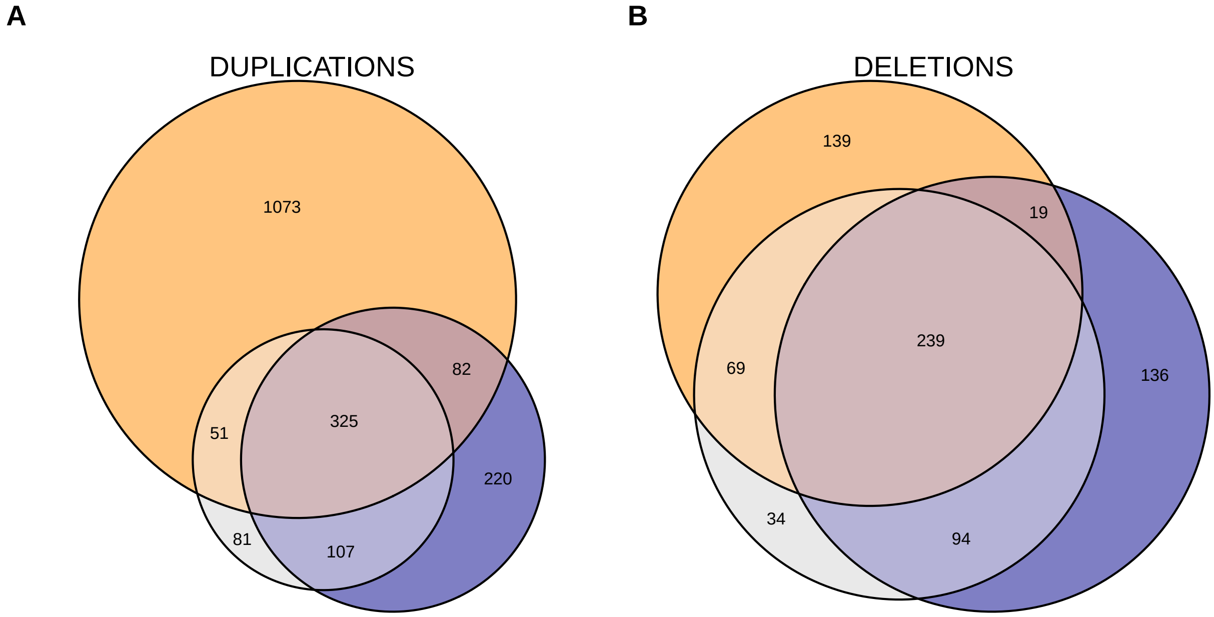

Figure 2.11: Copy number variant call concordance for the WGS pool, excluding subjects NCG_00790 and NCG_00851 due to poor correlation to the rest of the pool. (A) predicted duplications; (B) predicted deletions. mcCNV in grey; ExomeDepth in blue; ERDS/cnvpytor in orange. Values within overlaps give the number of variants.

2.4 Discussion

The medical genetics community still lacks robust exome-wide information about small (exon-level) variant prevalence. Others have established the reliability and cost-efficiency of pre-capture multiplexing,[65–69] and most commercial exome capture platforms have protocols for pre-capture multiplexing. Here, we demonstrate the reduction in inter-sample variance when pre-capture multiplexing, leading to increased power to detect exon-level copy number variation. Despite the benefits, many clinical laboratories do not employ a multiplexed capture protocol because multiplexing requires waiting to fill a pool and may delay results. While we understand the increased complexity, multiplexed capture may uncover otherwise missed copy number variation and increase the diagnostic yield for patients.

Multiplexed capture is not without limitations. We presented an example (pool IDT-MC) where multiplexed capture provided little to no improvement over independently-captured samples. We concluded the absent improvement in inter-sample variance stemmed from the poor library balance prior to capture. Rebuilding a more-balanced pool with the same samples (pool IDT-RR) demonstrated a large reduction in inter-sample variance.

In assessing the inter-sample variance, we compared two capture platforms: (1) Agilent SureSelectXT2 and (2) Integrated DNA Technologies xGen Lockdown Probes. We do not have enough data to suggest definitively one over the other. Comparing the mean-variance relationship, the IDT-RR pool appeared to have less dispersion overall (Figure 2.4); however, the sample-specific dispersion estimates from ExomeDepth suggest better performance by the WGS pool (Figure 2.8) and the higher pool-wide dispersion comes entirely from the two poorly correlated samples.

Our results suggest having a sufficiently large database of samples most-often provides appropriate control samples to estimate copy number variation (Figure 2.5). However, we show laboratories can circumvent the need for large samples by multiplexing the capture step. Defining the capture pool as the set of controls both limits the need for regular reanalysis as the database grows and eliminates potential over-selecting of samples with the same variants.

At the depth of the WGS pool, our simplistic simulation study would suggest both mcCNV and ExomeDepth have the power to detect single-exon variants with >85% sensitivity while maintaining a low false-discovery rate (Figure 2.9, Table 2.2). However, comparing the exome calls to the genome calls for the WGS pool revealed lackluster concordance. As Trost et al. point out, the genome CNV callers still struggle with variants less than 1 kb.[60] We do not dismiss the possibility of exome calls providing greater reliability than the genome calls, given multiplexed capture and adequate sequencing depth. However, given the distribution of calls throughout the exome, we doubt the thousands of excess deletions called for NCG_00790 and NCG_00851. Confirmation of the individual calls is beyond the scope of this work.

Unsurprisingly, both mcCNV and ExomeDepth failed to call many of the duplications called from the genome data. The variance for the negative binomial increases as the mean increases; we expect greater variation in read depth from duplicated loci, making duplications more difficult to distinguish. Similarly, the variance of the binomial proportion increases monotonically over [0, 0.5). More sensitive detection of duplications will likely require greater sequencing depth.

The simulation study emphasizes the importance of sequencing depth (in terms of absolute molecules). We can collect increased basepair coverage for less money by sequencing longer reads (e.g. 2x150 versus 2x50), but doing so decreases power for depth-based CNV calling. Typically, exome sequencing targets 30-50x coverage to ensure most targets have sufficient coverage for accurate basepair calling. We demonstrate the need for much deeper sequencing if we wish to establish exon-level variants.

Taken together, we recommend the following: (1) research and clinical endeavors consider adjusting protocols to multiplex samples prior to any targeted capture; (2) prior to capture, we suggest checking the library balance and adjusting as necessary (we achieved reasonable performance with relative standard deviation values less than 25%); (3) collecting an average of 225 filtered read-pairs per target. We then provide a simple-to-use and efficient R package to estimate copy number utilizing the negative bionimal distribution.

We believe the uncertainty about the prevalence and clinical significance of exon-level variants warrants a large undertaking. Even if we take the conservative approach and looking only at concordant calls between genome and exome sequencing (Figure 2.11), we have an average of 40 variants per sample to contend with. Two possibilities exist: (1) the algorithms all fail over specific regions, or (2) some genes can tolerate intrageneic copy-number variation better than others. Having eliminated calls from repetitive and low-complexity regions, we believe possibility (2) is more likely. To truly determine the prevalence (and therefore, clinical significance) of exon-level variants we need to interrogate exon-level variants on a large cohort. Confirmation testing for the tens to thousands of predicted variants from the exome and genome calls would allow true determination of algorithm performance and inform the clinical utility.

Page intentionally left blank.

CHAPTER 3

Shortcomings of exome sequencing in noninvasive prenatal genetics

3.1 Introduction

The beneficial health outcomes from newborn screening programs (NBS) are indisputable. We envision future NBS will begin with prenatal genetic testing to enable care in the immediate newborn period, and open up new possibilities for *in utero* and genetic therapies. During pregnancy, placental DNA is released into maternal circulation, enabling noninvasive interrogation of fetal genetics (noninvasive prenatal testing, NIPT). NIPT has a well-established clinical utility in screening for common chromosomal abnormalities such as Down syndrome with high sensitivity and specificity.[70] More recently, efforts have demonstrated sequencing-based testing for *de novo* pathogenic variants in a list of 30 genes associated with dominant Mendelian disorders[71] and PCR-based testing for a small number of recessive Mendelian disorders.[72] Using relative haplotype dosage analysis (RHDO),[73] multiple groups have successfully diagnosed single gene disorders[74–76] including a new offering of noninvasive prenatal diagnosis for cystic fibrosis in the UK Public Health Service.[77] RHDO typically relies on collecting parental, and ideally proband, genetic information to resolve parental haplotypes; Jang et al. demonstrated success in diagnosing Duchenne muscular dystrophy by estimating haplotypes solely from maternal long-read sequencing.[75] Scotchman et al. provide an excellent review summarizing the history of noninvasive testing.[78] To date, no one has reported reliable fetal genotyping purely from maternal cell-free DNA using a sequencing-based approach.

To begin NBS with prenatal genetic testing, we believe we first need a reliable noninvasive test only requiring a maternal sample. Others could reasonably argue the availability of carrier screening, and the immeasurably small risk of invasive testing,[79] removes the need for noninvasive testing. Such an argument, however, dismisses (1) the ethical and practical issues surrounding the necessity of involving the biological father, (2) the fact that many genetic disorders arise due to *de novo* mutations, and (3) the understandable fear and apprehension around invasive testing (especially for rare conditions). Additionally, we believe the prenatal diagnosis community should focus work on sequencing-based (as opposed to PCR-based) approaches. Sequencing generalizes across disorders more easily than PCR techniques, allows multiplexing to a degree not feasible using PCR, and will only continue to decrease in cost.

Snyder et al. provide a review of previous attempts to perform noninvasive fetal genome sequencing, illustrating the cost-infeasibility and suggesting more targeted approaches such as exome sequencing (ES).[73,80–82] As an exploratory exercise, we performed ES on cell-free DNA (cfES) from three pregnant women with singleton fetuses.

3.2 Methods

3.2.1 Participant selection

Genetic counselors identified pregnant women with suspected genetic disorders based either on family history or fetal sonographic findings. We enrolled three women, blinded to their family history and sonographic findings. All participants were consented and enrolled at UNC Hospitals by certified genetic counselors with approval from the UNC Institutional Review Board (IRB Number: 18-2618); we do not include any identifying information in this manuscript.

3.2.2 Exome sequencing and analysis

We collected cell-free DNA from maternal plasma, prepared sequencing libraries for the Illumina platform, and performed exome capture using the IDT xGen Exome Research Panel v1.0 (Cases 1 & 2) or Agilent SureSelect Human All Exon v7 (Case 3). We processed the data using a novel analytic pipeline developed in Snakemake[59] using Anaconda environments for reproducibility (provided in supplemental materials). Briefly, sequencing reads were aligned to hg38 (excluding alternate contigs) using BWA-MEM,[58] then base quality scores were re-calibrated using GATK4.[34,37,38] We only retained non-duplicate, properly-paired reads with unambiguous mapping and mapping quality >30 for each read. We called variants using bcftools,[36] requiring basepair quality scores >20. We suggest the review by Seaby et al. for more information on the specifics of collecting and processing ES data for clinical use.[83] Analyses were restricted to the regions overlapping between the IDT and Agilent capture platforms. For cell-free analyses, we required 5 alternate allele-supporting read-pairs, and at least 80 total read-pairs. Using the identified single-nucleotide variants, we applied a novel empirical Bayesian procedure to estimate the fetal fraction (FF; the proportion of placental/fetal to maternal sequencing reads). We then estimated maternal and fetal genotypes using a maximal likelihood model incorporating the FF estimate and observed proportion of minor allele (alternate) reads (PMAR).

3.2.3 Genotyping algorithm

Represent maternal and fetal genotype pairs, given by the random variable G, with capital and lowercase letters, where 'A' and 'B' represent the major and minor alleles (e.g. 'AAab' represents the fetus uniquely heterozygous for the minor allele).

Let X, Y be random variables for major and minor allele read counts. Define the fetal fraction and PMAR as the random variables F and M. Then, by definition, $\mathrm{E}[M] = \mathrm{E}[Y/(X+Y)]$. It's easily proven:

$$\mathrm{E}[M|G = \mathrm{AAab}, F = f] = \frac{f}{2} \tag{3.1}$$

$$\mathrm{E}[M|G = \mathrm{ABaa}, F = f] = \frac{1-f}{2} \tag{3.2}$$

$$\mathrm{E}[M|G = \mathrm{ABab}, F = f] = \frac{1}{2} \tag{3.3}$$

$$\mathrm{E}[M|G = \mathrm{ABbb}, F = f] = \frac{1+f}{2} \tag{3.4}$$

$$\mathrm{E}[M|G = \mathrm{BBab}, F = f] = 1 - \frac{f}{2} \tag{3.5}$$

We can then rearrange equations (3.1) and (3.5) and solve for the expected fetal fraction in terms of the PMAR:

$$\mathrm{E}[F|G=\mathrm{AAab}, M=m] = 2m \tag{3.6}$$

$$\mathrm{E}[F|G=\mathrm{BBab}, M=m] = 2-2m \tag{3.7}$$

Given the average population allele frequency for sequenced variants, we know the probability distribution of maternal/fetal genotypes under Hardy-Weinberg, $\Pr\{G=g\}$. As shown above, given the fetal fraction, $F=f$, we know the expected PMAR for each genotype, M. We observe the major and minor allele reads, $\mathbb{X}$ and $\mathbb{Y}$ respectively, and wish to estimate $\mathbb{G}, \hat{\mathbb{G}}$.

We employ an empirical Bayesian expectation-maximization algorithm to identify loci with unique fetal heterozygosity, i.e. $g \in \{\mathrm{AAab}, \mathrm{BBab}\}$. We pick reasonable starting values for the fetal fraction, $F=f$, and the average minor allele frequency, then iteratively update the expected allele distribution and expected PMAR values until some convergence:

1. Initialize the genotype probabilities, $p_g^* = \Pr\{G=g\}$, and the expected PMAR, $m_g^* = m_g$, based on reasonable estimates for the average minor allele frequency and fetal fraction

2. Update $\hat{\mathbb{G}}$:

$$\hat{g}_i = \underset{g\in G}{\operatorname{argmax}} \left\{ p_g^* \mathcal{L}(g|m_g^*, x_i, y_i) \right\}, Y_i \sim \mathrm{Bin}(x_i+y_i, m_g^*)$$

3. Update the genotype probabilities:

$$p_g^* = \frac{\sum_i \mathrm{I}(\hat{g}=g) + N\Pr\{G=g\} - 1}{\sum_g \{\sum_i \mathrm{I}(\hat{g}=g) + N\Pr\{G=g\} - 2\}}$$

where N is the weight given to the initial estimate of the genotype probability, $\Pr\{G=g\}$.

4. Update the expected PMAR:

$$m_g^* = \frac{\sum_i y_i \mathrm{I}(\hat{g}=g) + Nm_g - 1}{\sum_i (x_i+y_i)\mathrm{I}(\hat{g}=g) + N - 2}$$

where N is the weight given to the initial estimate of the PMAR, m_g.

5. Continue updating $\hat{\mathbb{G}}$ (2), p_g^* (3), and m_g^* (4) until $\hat{\mathbb{G}}$ converges.

6. For all loci j, such that $\hat{g} \in \{\text{AAab}, \text{BBab}\}$, calculate $\hat{f}_j$:

$$\hat{f}_j = \begin{cases} \dfrac{2y_j}{x_j + y_j}, & \hat{g} = AAab \\ \\ 2 - \dfrac{2y_j}{x_j + y_j}, & \hat{g} = BBab \end{cases}$$

7. Let

$$\hat{f} = \text{median}\left(\hat{f}_j\right)$$

8. Calculate the expected PMAR using the fetal fraction estimate,

$$m_g = \mathrm{E}[M|\hat{f}, g]$$

9. Finally, for all loci, i, estimate $\hat{g}_i \in \hat{\mathbb{G}}$,

$$\hat{g}_i = \underset{g \in G}{\text{argmax}} \left\{\mathcal{L}(g|m_g, x_i, y_i)\right\}, Y_i \sim \text{Bin}(x_i + y_i, m_g)$$

3.2.4 Data availability

The data that support the findings of this study are available on request from the corresponding author. The raw sequencing data are not publicly available due to privacy or ethical restrictions. Allele depths, with the alleles masked and genomic location rounded to 10 kilobases are available in the self-contained R[84] package reproducing the analysis herein (`https://github.com/daynefiler/filer2020B`).

3.3 Results

Using the final set of filtered reads, we analyzed single nucleotide loci with >80x coverage and at least 5 reads supporting the alternate allele. At each analyzed site, we alternate allele sequencing

Table 3.1: Case summaries. GA: gestational age at the time of blood draw for cfES. FF: estimated fetal fraction. Depth: median depth used to estimate genotypes (does not include duplicated/filtered reads). %Dup: percentage of total mapped read pairs discarded as PCR and/or optical duplicates. %Filt: percentage of total mapped read pairs discarded for improper pairing and/or mapping quality.

	GA	Clinical findings	Genetic diagnosis	FF	Dep	%Dup	%Filt
1	32w2d	5 prior pregnancies affected with X-linked recessive Menke's syndrome	Menke's syndrome; del. ATP7A exon 1	0.117	241	42.80	21.96
2	24w5d	Fetal sonogram at 21w5d showed femoral bowing with shortened length (<3% for GA) bilaterally	Osteogenesis imperfecta type VIII; P3H1 c.1120G>T (rs140468248)	0.122	152	33.32	22.09
3	34w0d	Fetal sonogram at 19w0d showed bilateral club foot with bilateral upper limb arthrogryposis	None, to date, despite exome and genome sequencing of newborn	0.169	330	53.67	32.65

depth and total sequencing depth to estimate the fetal fraction and maternal-fetal genotypes using our novel algorithm (Figure 3.1). Table 3.1 lists the known genetic diagnoses for the three cases presented. Genetic counselors recruited the three participants; investigators and cfES analysis was blinded to the eventual genetic diagnoses. In Cases 1 & 2, specific gene sequencing based on family history and sonographic findings, respectively, provided genetic diagnoses. To date, Case 3 does not have a genetic diagnosis. We learned the mother in Case 1 carries a deletion of exon 1 in the gene most-often responsible for Menke's syndrome (ATP7A). Neither exome capture platform targets ATP7A exon 1; therefore, cfES could not have identified the diagnosis for Case 1 with the platform used. In Case 2, we identified the causal variant using cfES. In this case, we correctly genotyped the fetus, but lacked the power to make the genotyping call with any level of confidence acceptable for clinical use (Figure 3.1B, note the widely-overlapping distributions at the causal variant). We did not identify any known pathogenic variants in the sequencing of Case 3, and despite performing genome sequencing on the newborn, we still do not have a genetic diagnosis for the family.

In Case 3, in addition to cfES, we performed exome sequencing (ES) on fetal, maternal, and paternal samples. Based on previous work demonstrating the differential length of maternal and

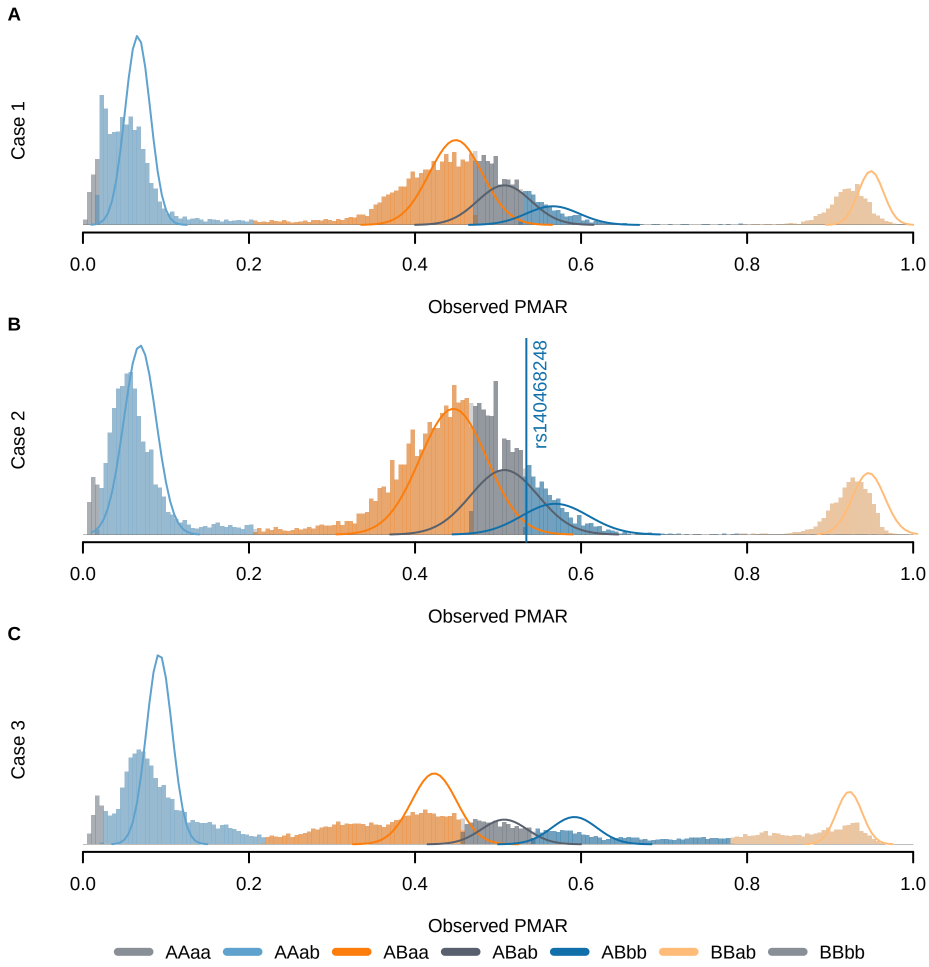

Figure 3.1: Distribution of observed proportion of minor allele reads (PMAR) values for the three cases across the possible maternal-fetal genotype pairs. Uppercase letters give the estimated maternal genotype, lowercase letters give the estimated fetal genotype; 'A/a' indicates the reference allele, 'B/b' indicates the alternate allele. Solid lines show the normal approximation for the theoretical distribution of binomial probabilities, given the frequency of the estimated genotypes. The vertical line in (B) shows the observed PMAR for the known pathogenic variant, rs140468248.

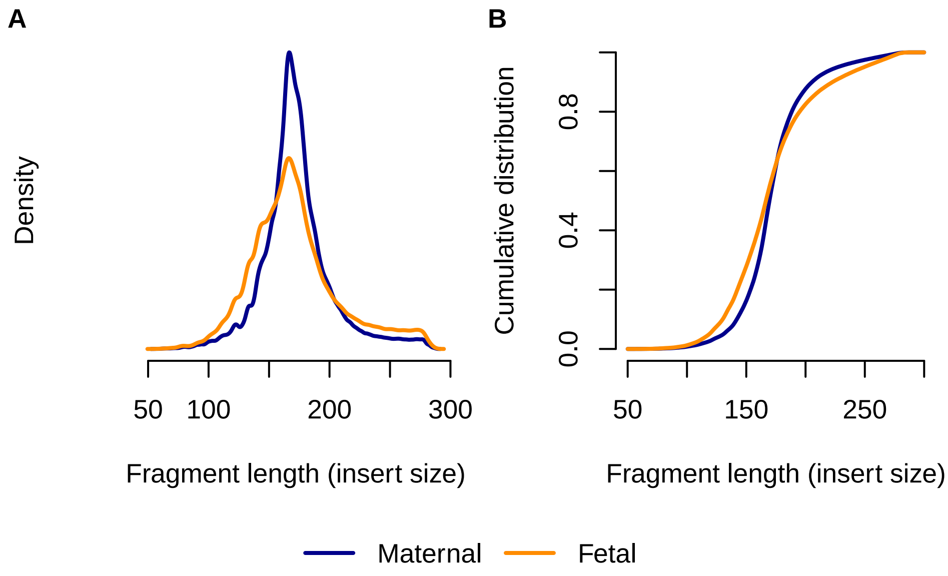

Figure 3.2: Distribution of maternal versus fetal fragment length in Case 3. (A) shows the density; (B) shows the emperic cumulative distribution. The horiztonal axis shows the fragment length (insert size taken from aligned read-pairs). Blue lines show maternal reads, orange lines show fetal reads. We only included cfES reads supporting alleles unique to the mother or fetus, as identified from the direct maternal and fetal ES.

fetal fragments,[85–88] we interrogated the distribution of presumed maternal and fetal reads (Figure 3.2). We identified maternal and fetal reads by identifying sites with unique heterozygosity in the direct maternal and fetal ES results; at the informative sites, we extracted reads supporting the allele unique to the mother or fetus. In total, we identified 654,619 maternal reads and 279,508 fetal reads. We found, as others have, a higher proportion of fetal reads falling below 150 basepairs; however, we also observed a slightly higher proportion of longer reads, as well.

Rabinowitz et al. proposed the Hoobari method which incorporates fragment lengths into fetal genotype estimates,[88] finding the difference in accuracy varied from -0.25% to 1.89% when using versus not using fragment length in their exome analyses. To explore the utility of correcting for fragment length in our analysis, we interrogated the PMAR as a function of the short read proportion (fraction of reads with insert sizes less than 140 basepairs; Figure 3.3). We selected 140 as the cutoff based on the Hoobari algorithm. Overall, we found no meaningful relationship between the short read proportion and the observed PMAR and chose not to incorporate fragment length into our genotype estimates.

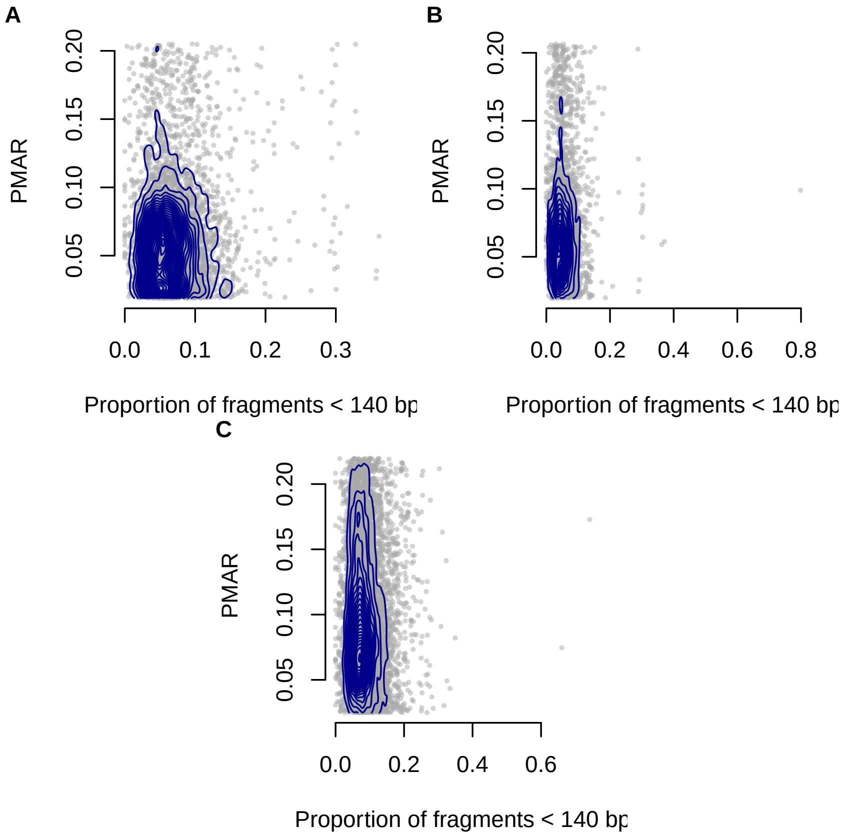

Figure 3.3: Proportion of minor allele reads (PMAR) as a function of the short read proportion for genotypes estimated as ‘AAab.’ Short reads defined as fragments less than 140 basepairs. (A-C) show Cases 1 to 3, respectively. Gray points show the individual sites; blue contour lines show the two-dimensional distribution of values.

Table 3.2: Case 3 fetal versus cell-free genotype calls. ‘0’ represents the major allele; ‘1’ represents the minor allele. Sites with cell-free estimates and reliable direct fetal calls included (reliable defied as passing all quality checks and having a total sequencing depth greater than 30).

		Cell-free		
		0/0	0/1	1/1
Fetal	0/0	1,063	1,857	9
	0/1	3,598	7,079	1,454
	1/1	76	2,197	1,391

Returning to Case 3, we interrogated the fetal genotyping accuracy at all sites with cell-free genotype estimates and reliable calls from the direct fetal sample. Overall, we found a ‘\Sexpr{50.91% accuracy (Table 3.2). Table 3.3 provides the full set of maternal, fetal, and cell-free calls.

Table 3.3: Maternal, fetal, and cell-free genotype calls. '0' represents the major allele; '1' represents the minor allele. Sites with cell-free estimates and reliable direct fetal calls included (reliable defied as passing all quality checks and having a total sequencing depth greater than 30).

Maternal	Fetal	Cell-free	N
0/0	0/0	0/0	468
		0/1	1,159
	0/1	0/0	64
		0/1	352
	1/1	0/0	1
		0/1	2
0/1	0/0	0/0	387
		0/1	107
		1/1	6
	0/1	0/0	3,072
		0/1	1,967
		1/1	713
	1/1	0/0	68
		0/1	458
		1/1	1,291
1/1	0/0	0/1	1
		1/1	2
	0/1	0/0	3
		0/1	1,308
		1/1	648
	1/1	0/1	1,601
		1/1	23

3.4 Discussion

Without the ability to reliably exclude maternal DNA fragments, noninvasive sequencing-based methods to genotype the fetus either require additional sequencing of parental samples or distinguishing genotypes by the proportion of minor allele reads (PMAR). Here, we make no attempt to utilize parental genetic information and demonstrate the difficulty of inferring the genotypes directly from the PMAR. We model the PMAR as a binomial proportion; given the fetal fraction, one can prove the true PMAR defines the maternal and fetal genotypes (supplemental document).

The theoretical bounds of the binomial distribution, therefore, confine our ability to discriminate maternal-fetal genotypes. Using the normal approximation for the binomial variance (valid when the number of observations, i.e. sequencing depth, times the binomial proportion, i.e. PMAR, is greater than 10), we can clearly explain the poor results we observed (Figures 3.4 and 3.5). At sequencing depths up to 500x, the 95% confidence intervals on PMAR distributions still overlap for fetal fractions up to roughly 0.17 (Figure 3.4). When we calculate the degree of distribution overlap (a proxy for classification error rate), we see required sequencing depths in excess of 8,000x for low fetal fraction samples (Figure 3.5). We, therefore, cannot expect cell-free sequencing to reliably differentiate genotypes without substantially higher depth or additional genetic information. No amount of cleverness in the analysis can overcome the fundamental variance bounds when estimating binomial proportions.

The sequencing herein suffers from three problems: (1) inadequate sequencing depth; (2) biased PMAR values from the removal of duplicate reads; (3) errors in sequencing and/or PCR. We have already illustrated the inadequate depth, but emphasize that the theoretical results we present speak to the final depths (not the raw sequencing depth). In our three cases, we excluded over half the reads taken off the sequencer due to sequencing quality thresholds (Table 3.1). We observe the evidence of problems (2) and (3) by observing the high proportion of both duplicate reads and PMAR values outside the theoretic distributions. Additionally, we observed very poor accuracy in the Case 3 genotype estimates.

Typical sequencing workflows start with randomly fragmenting DNA molecules to build sequencing libraries. Standard bioinformatic practices suggest we remove read-pairs with identical

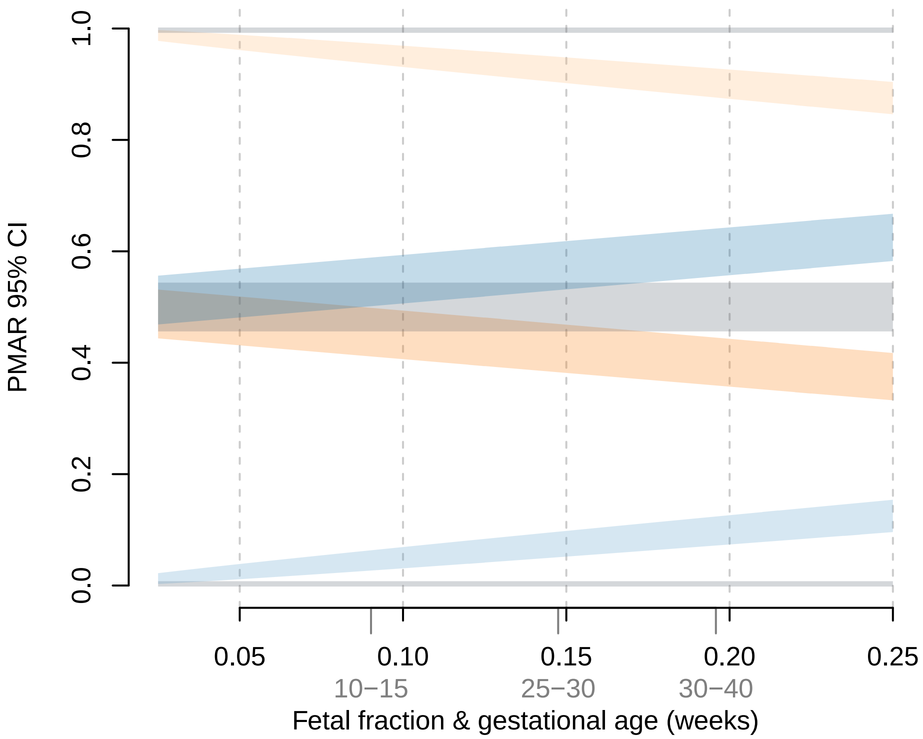

Figure 3.4: 95% confidence intervals on the binomial proportions for possible maternal-fetal genotype pairs across increasing fetal fractions. Confidence intervals represent a sequencing depth of 500x. Average fetal fractions by gestational age (in weeks) given in light gray.[89]

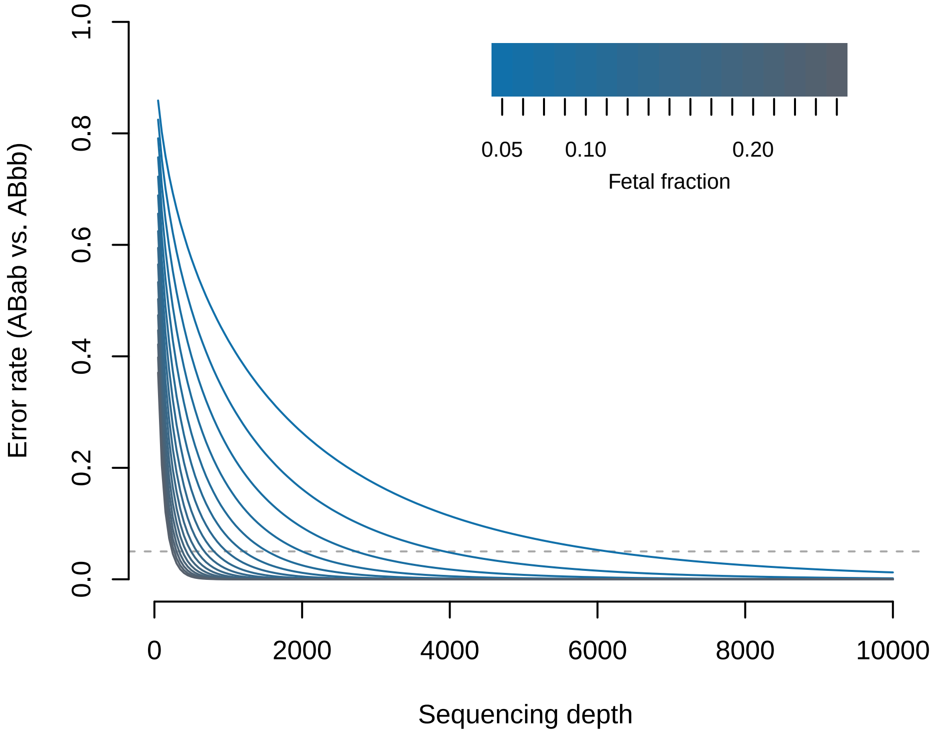

Figure 3.5: Expected misclassification rate (Weitzman overlapping coefficient; i.e. the area of overlapping distributions in Figure 3.4) considering ABab versus ABbb as a function of sequencing depth and fetal fraction. The dashed horizontal line shows 5% error. The theoretical error rates for ABab vs ABaa are symmetric and equal; however, the frequency of errors will depend on the population frequency of the reference versus alternate allele.

endpoints, because the duplicate read-pairs more likely represent PCR amplification of a single molecule than two molecules with the same fragmentation. Cell-free DNA molecules are shorter than nuclear DNA, not requiring manual fragmentation, and have a non-random distribution of endpoints.[86] Therefore, compared to standard sequencing libraries, the likelihood of observing true duplicates in cell-free libraries increases and we cannot necessarily assume duplicates represent PCR amplification. However, for this work we have no way of differentiating reads representing true duplicate molecules versus PCR duplicates and thus excluded duplicate reads from our analysis.

Assuming adequate depth and appropriate handling of duplicate reads and sequencing errors, incorporating the fragment length into the statistical model may prove more beneficial. The high variability of the binomial distribution for small n obfuscates any meaningful relationship between fragment length and PMAR in our data. We reiterate, however, incorporating fragment length may give better estimates of the binomial proportion but cannot decrease variance beyond the distribution bounds.

To solve the above issues, we advocate a more targeted approach with much greater sequencing depth and unique molecular identifiers. Unique molecular identifiers allow identification of sequencing errors and differentiate true versus artifactual duplicate reads. Given the depth requirements for estimating fetal genotypes by the PMAR, and the challenge of variants of uncertain clinical significance, we advocate against broad sequencing modalities on noninvasive samples. Recognizing that all capture methods introduce bias in the relative sequencing efficiency of different targeted regions,[83] the sequencing depths needed for noninvasive fetal genotyping necessitate a targeted approach. Despite the challenges raised by this work, we believe assessing hundreds to thousands of basepairs, rather than the tens of millions targeted in ES, will prove economical and clinically reliable. Doing so, we hope, will foster population-level screening for Mendelian disorders during the prenatal period and, ultimately, unlock new avenues in the treatment of these disorders.

Page intentionally left blank.

REFERENCES

1. Mendel G. Versuche über pflanzenhybriden. *Verh Naturf Ver Brünn.* 1866;4:3—47.

2. Miescher F. Über die chemische zusammensetzung der eiterzellen. *Hoppe-Seyler's Med Chem Unters.* 1871;4:441-460.

3. Miescher F. Das protamin, eine neue organische basis aus den samenfäden des rheinlachses. *Ber Deutsch Chem Ges.* 1874;7:376—379.

4. Miescher F. Die spermatozoen einiger wilbertiere. *Ein Beitrag zur histochemie Verh Naturf Ges.* 1874;6:138—208.

5. Kossel A, Neumann A. Über das thymin, ein spaltungsprodukt der nukleinsäure. *Ber Deutsch Chem Ges.* 1893;26:2753—2756.

6. Flemming W. Zur kenntniss der zelle und ihrer theilungs-erscheinungen. *Schriften des Naturwissenschaftlichen Vereins für Schleswig-Holstein.* 1878;3:23—27.

7. Boveri TH. Über mehrpolige mitosen als mittel zur analyse des zellkerns. *Verh Phys Med Ges Vürzb.* 1902;35:60-90.

8. Boveri TH. Über die konstitution der chromatischen kernsubstanz. *Verh Deutsch Zool Ges Würzb.* 1903;13(10–33).

9. Sutton WS. The chromosomes in heredity. *Biol Bull (Woods Hole).* 1903;4:231-251.

10. Avery OT, Macleod CM, McCarty M. STUDIES on the chemical nature of the substance inducing transformation of pneumococcal types : INDUCTION of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *J Exp Med.* 79(2):137-158.

11. Chargaff E, Vischer E, Doniger R, Green C, Misani F. The composition of the desoxypentose nucleic acids of thymus and spleen. *J Biol Chem.* 177(1):405-416.

12. FRANKLIN RE, GOSLING RG. Molecular configuration in sodium thymonucleate. *Nature.* 171(4356):740-741.

13. WATSON JD, CRICK FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* 171(4356):737-738.

14. CRICK FH. On protein synthesis. *Symp Soc Exp Biol.* 1958;12:138-163.

15. Crick F. Central dogma of molecular biology. *Nature.* 227(5258):561-563.

16. Sanger F, Coulson AR. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *J Mol Biol.* 94(3):441-448.

17. Maxam AM, Gilbert W. A new method for sequencing dna. *Proc Natl Acad Sci U S A.* 74(2):560-564.

18. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 74(12):5463-5467.

19. Nyrén P, Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem.* 151(2):504-509.

20. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. *Science.* 281(5375):363, 365.

21. Li Z, Bai X, Ruparel H, Kim S, Turro NJ, Ju J. A photocleavable fluorescent nucleotide for dna sequencing and analysis. *Proc Natl Acad Sci U S A.* 100(2):414-419.

22. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for dna attachment on glass and efficient generation of solid-phase amplified dna colonies. *Nucleic Acids Res.* 34(3):e22.

23. Turcatti G, Romieu A, Fedurco M, Tairi A-P. A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for dna sequencing by synthesis. *Nucleic Acids Res.* 36(4):e25.

24. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 85(8):2444-2448.

25. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res.* 38(6):1767-1771.

26. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and samtools. *Bioinformatics.* 25(16):2078-2079.

27. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol*. 147(1):195-197.

28. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 48(3):443-453.

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 215(3):403-410.

30. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform*. 11(5):473-483.

31. Burrows M, Wheeler DJ. *A Block-Sorting Lossless Data Compression Algorithm.*; 1994.

32. Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM. Compressed indexing and local alignment of dna. *Bioinformatics*. 24(6):791-797.

33. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*. 26(5):589-595.

34. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*. 20(9):1297-1303.

35. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*. 43(5):491-498.

36. Li H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 27(21):2987-2993.

37. Van der Auwera GA, Carneiro MO, Hartl C, et al. From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43(1110):11.10.1-11.10.33.

38. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. Published online 2018.

39. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. Published online 2012.

40. Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci Rep.* 9(1):9345.

41. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J.* 2018;16:15-24.

42. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome Res.* 21(6):974-984.

43. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics.* 28(21):2747-2754.

44. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science.* 318(5849):420-426.

45. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* 15(6):R84.

46. Zhu M, Need AC, Han Y, et al. Using erds to infer copy-number variants in high-coverage genomes. *Am J Hum Genet.* 91(3):408-421.

47. Marchuk DS, Crooks K, Strande N, et al. Increasing the diagnostic yield of exome sequencing by copy number variant analysis. *PLoS One.* 2018;13(12):e0209185.

48. Retterer K, Scuffins J, Schmidt D, et al. Assessing copy number from exome sequencing and exome array cgh based on cnv spectrum in a large clinical cohort. *Genet Med.* 17(8):623-629.

49. Yao R, Zhang C, Yu T, et al. Evaluation of three read-depth based cnv detection tools using whole-exome sequencing data. *Mol Cytogenet.* 2017;10:30.

50. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet.* 91(4):597-607.

51. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: A normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* 43(6):e39.

52. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22(8):1525-1532.

53. Truty R, Paul J, Kennemer M, et al. Prevalence and properties of intragenic copy-number variation in mendelian disease genes. *Genet Med.* 21(1):114-123.

54. Benjamini Y, Speed TP. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Res.* 40(10):e72.

55. Kadalayil L, Rafiq S, Rose-Zerilli MJJ, et al. Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform.* 16(3):380-392.

56. Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods.* 6(1):99-103.

57. Foreman AKM, Lee K, Evans JP. The NCGENES project: Exploring the new world of genome sequencing. *N C Med J.* 74(6):500-504.

58. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. Published online 2013.

59. Koster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics.* 28(19):2520-2522.

60. Trost B, Walker S, Wang Z, et al. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet.* 102(1):142-155.

61. Yu D, Huber W, Vitek O. Shrinkage estimation of dispersion in negative binomial models for rna-seq experiments with small sample size. *Bioinformatics.* 29(10):1275-1282.

62. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological).* 1995;57(1):289-300.

63. Minka TP. *Estimating a Dirichlet Distribution.*; 2000.

64. Matthews BW. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta.* 405(2):442-451.

65. Neiman M, Sundling S, Grönberg H, et al. Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. *PLOS ONE.* 2012;7(11):1-6.

66. Ramos E, Levinson BT, Chasnoff S, et al. Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. *BMC Genomics.* 13:683.

67. Rohland N, Reich D. Cost-effective, high-throughput dna sequencing libraries for multiplexed target capture. *Genome Res.* 22(5):939-946.

68. Wesolowska A, Dalgaard MD, Borst L, et al. Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant snps in childhood acute lymphoblastic leukemia. *Leukemia.* 25(6):1001-1006.

69. Shearer AE, Hildebrand MS, Ravi H, et al. Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics.* 13:618.

70. Mackie FL, Hemming K, Allen S, Morris RK, Kilby MD. The accuracy of cell-free fetal dna-based non-invasive prenatal testing in singleton pregnancies: A systematic review and bivariate meta-analysis. *BJOG.* 124(1):32-46.

71. Zhang J, Li J, Saucier JB, et al. Non-invasive prenatal sequencing for multiple mendelian monogenic disorders using circulating cell-free fetal dna. *Nat Med.*

72. Tsao DS, Silas S, Landry BP, et al. A novel high-throughput molecular counting method with single base-pair resolution enables accurate single-gene nipt. *Sci Rep.* 9(1):14382.

73. Lo YMD, Chan KCA, Sun H, et al. Maternal plasma dna sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med.* 2(61):61ra91.

74. Hui WWI, Jiang P, Tong YK, et al. Universal haplotype-based noninvasive prenatal testing for single gene diseases. *Clin Chem.* 63(2):513-524.

75. Jang SS, Lim BC, Yoo S-K, et al. Targeted linked-read sequencing for direct haplotype phasing of maternal dmd alleles: A practical and reliable method for noninvasive prenatal diagnosis. *Sci Rep.* 8(1):8678.

76. Vermeulen C, Geeven G, Wit E de, et al. Sensitive monogenic noninvasive prenatal diagnosis by targeted haplotyping. *Am J Hum Genet.* 101(3):326-339.

77. Chandler NJ, Ahlfors H, Drury S, et al. Noninvasive prenatal diagnosis for cystic fibrosis: Implementation, uptake, outcome, and implications. *Clin Chem.* 66(1):207-216.

78. Scotchman E, Chandler NJ, Mellis R, Chitty LS. Noninvasive prenatal diagnosis of single-gene diseases: The next frontier. *Clin Chem.* 66(1):53-60.

79. Salomon LJ, Sotiriadis A, Wulff CB, Odibo A, Akolekar R. Risk of miscarriage following amniocentesis or chorionic villus sampling: Systematic review of literature and updated meta-analysis. *Ultrasound Obstet Gynecol.* 54(4):442-451.

80. Fan HC, Gu W, Wang J, Blumenfeld YJ, El-Sayed YY, Quake SR. Non-invasive prenatal measurement of the fetal genome. *Nature.* 487(7407):320-324.

81. Kitzman JO, Snyder MW, Ventura M, et al. Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med.* 4(137):137ra76.

82. Snyder MW, Simmons LE, Kitzman JO, et al. Noninvasive fetal genome sequencing: A primer. *Prenat Diagn.* 33(6):547-554.

83. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: A practical guide to its clinical application. *Brief Funct Genomics.* 15(5):374-384.

84. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2019.

85. Chan KCA, Zhang J, Hui ABY, et al. Size distributions of maternal and fetal dna in maternal plasma. *Clin Chem.* 50(1):88-92.

86. Chan KCA, Jiang P, Sun K, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred dna ends. *Proc Natl Acad Sci U S A.* 113(50):E8159-E8168.

87. Jiang P, Lo YMD. The long and short of circulating cell-free dna and the ins and outs of molecular diagnostics. *Trends Genet.* 32(6):360-371.

88. Rabinowitz T, Polsky A, Golan D, et al. Bayesian-based noninvasive prenatal diagnosis of single-gene disorders. *Genome Res.* 29(3):428-438.

89. Kinnings SL, Geis JA, Almasri E, et al. Factors affecting levels of circulating cell-free fetal dna in maternal plasma and their implications for noninvasive prenatal testing. *Prenat Diagn.* 35(8):816-822.