Statistical Methods, Analyses and Applications for Next-Generation Sequencing Studies

by

Yan Yancy Lo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Sebastian K. Zöllner, Chair
Professor Gonçalo Abecasis
Professor Betsy Foxman
Professor Timothy D. Johnson
Assistant Professor Hyun Min Kang

© Yan Yancy Lo 2015

All Rights Reserved

To my parents,
who are my source of love.

To my husband, Henry,
who is my source of support.

To my sister, Wai,
who is my source of joy.

And to God,
who is the source of all wisdom.

ACKNOWLEDGEMENTS

The pursuit of my doctorate degree has been a rewarding yet extremely humbling experience. This dissertation would not have been possible without the support and guidance from my committee, colleagues, friends and family.

First, I would like to express my deepest gratitude to my committee chair, Dr. Sebastian Zöllner, who introduced me to statistical and population genetics when I was in transition between graduate programs. Since then, he has provided me invaluable research opportunities and guidance to build a solid foundation in genetics and biostatistics. He has been extremely patient, accommodating and encouraging. His excellent mentorship has helped me mature as an independent scientist and as an individual alike.

I would like to extend the gratitude to Dr. Gonçalo Abecasis, who is always able to offer insightful advice for my research projects. I am also extremely grateful for the financial support Dr. Abecasis has provided for my graduate studies. I would also like to express my heartfelt thanks to Dr. Betsy Foxman, who has provided me valuable advice on the biological motivations of statistical methods, which has expanded my understanding of genetics and epidemiology. I truly look up to Dr. Foxman as my role model of a successful woman scientist. I would also like to thank Dr. Hyun Kang and Dr. Timothy Johnson for their patient instructions both inside the classrooms and specific to my research projects.

In addition, I would like to express my deep appreciation to Kirsten Herold who has offered tremendous help on my scientific writing. I have benefited greatly from her instruction on becoming a better writer. I would also like to thank the past and current members of the Zöllner lab, in particular Matthew Zawistowski, Ziqian Geng, Mark Reppell and Keng-han Lin, for their generous exchange of scientific ideas. The same gratitude goes to my colleagues at the Center for Statistical Genetics and the Department of Biostatistics, whom I have spent a lot of time learning, working and having fun with. I truly treasure our scholastic discussions and friendships.

Next, I would like to thank my friends from the Ann Arbor Chinese Christian Church who are more than friends - they are my family away from home. They have never ceased to express their support and care for me. My graduate studies in Ann Arbor have been genuinely blessed with many fond memories and lifelong companionships with these wonderful people.

Last but not least, I would like to thank my husband, Henry Fan, my parents, Norman Lo and Man-Yee Chan, and my sister, Wai Lo, for their continuous encouragement and support. My family has shown me that love surpasses all physical distances and intellectual understanding; it is their unconditional love that has kept me motivated throughout the ten years of higher education in the United States.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical Methods, Analyses and Applications

for Next-Generation Sequencing Studies

by
Yan Yancy Lo

Chair: Sebastian Zöllner

Current genetics studies rely heavily on next-generation sequencing (NGS) techniques. This dissertation addresses methodological developments and statistical strategies to efficiently and accurately analyze the large amounts of NGS data, thereby to understand the genetic contributions to diseases.

In chapter 2, we evaluated the benefits of different variant calling strategies by performing a comparative analysis of calling methods on large-scale exonic sequencing datasets. We found that individual-based analyses identified the most high quality singletons, but had lower genotype accuracy at common variants than population-based and LD-aware analyses. Therefore, we recommend population-based analyses for high quality variant calls with few missing genotypes, complemented by individual-based analyses to obtain the most singleton variants.

In chapters 3 and 4, we addressed the issue of overlapping read pairs in NGS studies arising from short fragments. In chapter 3, we proposed novel models to separately estimate machine and fragment errors of a NGS experiment from overlapping

read pairs. Using a Markov chain Monte Carlo algorithm, our models suggested that machine and fragment errors were largely predicted by the reported quality scores of the overlapping bases and were uniform across individual samples from the same experiment. In chapter 4, we proposed an algorithm, RESCORE, to resolve the fragment dependence while retaining machine error estimates in overlapping reads. When compared to soft-clipping the overlapping regions, RESCORE increased the recalibrated base quality scores for the majority of overlapping bases, leading to a decrease in estimated false positive rate of novel variant discovery.

In chapter 5, we presented an application of whole-genome sequencing for understanding the evolutionary history of uropathogenic *Escherichia coli* (UPEC). We sequenced 14 UPEC and 5 commensals at >190x, and found a deep split between UPEC and commensal *E. coli*. We observed high between-strain diversity, which suggests multiple origins of pathogenicity. We detected no selective advantage of virulence genes over other genomic regions. These results suggest that UPEC acquired uropathogenicity a long time ago and used it opportunistically to cause extraintestinal infections. In summary, this dissertation presented practical strategies for NGS studies that will contribute to further genetic advances.

CHAPTER I

Introduction

Since the publication of the first finished-grade human genome sequence in 2004 [36], scientific resources and efforts have been heavily allocated to develop genome analysis technologies [106]. In the past decade, particularly the past few years, genome analysis has shifted from using first-generation capillary sequencing to next-generation sequencing [5, 71]. Next-generation sequencing (NGS) is the massively parallel, high-throughput technology that allows the generation of large amounts of genetic data [72, 77]. In contrast to capillary sequencing technologies, which typically produces 500-800 bases per reaction and required over three years and three million dollars to sequence one human genome [36, 92], NGS platforms are able to generate in a much shorter time and lower cost large numbers of short reads that can be reconstructed into the full genome [7, 107]. The recent Illumina HiSeq X Ten is the first platform to deliver full coverage human genomes for close to a thousand dollars, with the yield of up to 6 terabases per day, which is approximately 2000 times of the number of bases in a full human genome [34, 119].

Rapid advancement in sequencing technologies has prompted the success of numerous genetics studies, with a common goal of cataloging variations among human populations or populations of other species [41, 44, 93, 102, 113, 116]. An accu-

rate and detailed database of genetic variations has extensive biological and medical applications. For example, the 1000 Genomes Project sequenced the genomes of $> 2,500$ individuals from populations around the world and created a detailed catalog of human genetic variations, including variants with population allele frequency as low as 1% [116, 117]. This catalog of variation complemented many sequencing and genome-wide association studies in improving data quality and expanding sample sizes, thereby increasing the power to detect associations between genetic loci and complex diseases [51, 110, 121].

As large-scale sequencing studies become more widespread, the analyses of individual genomes have led to discoveries of novel rare polymorphisms (allele frequency <1%), which were difficult to detect from small samples or from genotyping chips which identify alleles from predetermined genomic positions [120]. Rare polymorphisms are of particular interest and importance, because they are more likely to be functional [12, 69, 95]. In the recent study published by Nelson et al. [82] that sequenced over 14,000 individuals at 202 drug targeted genes, they discovered rare variants as frequent as one every 17 base pairs; over half are expected to be deleterious. Moreover, rare polymorphisms have shown to have larger effects on disease phenotype than common polymorphisms [1, 15, 21]. Therefore, identifying such rare variants can explain significant proportions of the missing heritability [48, 139]. From the population genetics perspective, variants are rare because they arise from the recent past. NGS analysis of over 6,500 exomes suggested that most protein-coding variants are rare and have recent origin of 5000-10000 years [24]. The excess of rare variants present in the current population indicate that the human population has expanded faster than exponentially in the recent past [43, 99].

To efficiently and accurately analyze the large amounts of NGS data, advanced

statistical methods are required to further the understanding of the contributions of genetic variation to complex traits. Indeed, the studies mentioned above were successful due to the development of a large range of statistical methods and analysis pipelines designed for NGS studies [17, 40, 54, 61]. However, to date there is no single pipeline that works uniformly the best in all NGS studies; the quality and quantity of variant discovery depends heavily on the statistical methods and bioinformatics tools used [90, 133]. While the curation of variants is very important to minimize false discoveries, the optimal variant discovery strategy should also minimize false negatives in order to make the best use of the data. Moreover, errors generated along the multiple steps of a sequencing experiment can accumulate and be misinterpreted as rare variants. In fact, NGS techniques are believed to have relatively higher error rates than chip genotyping or capillary sequencing techniques [94, 98, 107]; therefore the older technologies are still frequently used to validate selected variants from NGS studies [65, 82, 136].

In this dissertation, we proposed novel statistical methods and analysis plans to address the challenges in NGS data. We considered different variant calling algorithms for obtaining the most complete set of variants in large-scale sequencing datasets. We devised a strategy for estimating separate sources of sequencing errors in an NGS experiment. We proposed and implemented a method to correct for the dependence in overlapping pairs of sequence reads. Finally, we designed and carried out a whole-genome deep-sequencing study to understand the evolutionary trajectories of virulence genes in pathogens. The dissertation concludes with a discussion of the implications and significance of our methods for the development of NGS studies. In the sections below, we provide a more detailed overview of each chapter.

Variant calling in large-scale NGS studies

Variant discovery, or variant calling, plays a central role between raw NGS data and applications. Capillary sequencing directly reads the alleles at each position, and genotyping chips start with a set of known variant sites and type each individual sample's alleles at these sites; in contrast, NGS determines the alleles at each genomic position statistically based on the combined evidence from the short reads spanning the position [58, 59, 76, 86]. Therefore, variant calling typically improves with high coverage, because of the increased information at each position.

Sequencing experiments targeted at exonic regions or specific genes are popular because these regions potentially harbor most functional variants or variants of interest [6]. At a fixed cost, sequencing only these genomic regions allows the expansion of sample size. By examining a lot of individuals, these sequencing studies of exonic regions aim to identify rare variants contributing to complex traits [6, 67, 115]. With high coverage and large sample size, these studies tend to apply simple variant calling algorithms, which typically call genotypes from the reads covering each position per individual (individual-based caller) [58, 59]. However, coverage is often heterogeneous due to the uneven capturing of the targeting technology [68, 77]. Therefore, sites with insufficient coverage may benefit from sophisticated calling algorithms used in low-coverage sequencing studies, which call genotypes based on population allele frequency (population-based caller) [17, 40, 49] or based on genotypes in linkage disequilibrium (LD) in the sample (LD-aware caller) [61, 70].

In chapter 2, we evaluated the potential benefits of different calling strategies by performing a comparative analysis of variant calling methods on exonic data from multiple large-scale sequencing datasets [82, 136]. We call variants using individual-

based, population-based and LD-aware methods with stringent quality control. We measure genotype accuracy by the concordance with on-target GWAS genotypes and between 80 pairs of sequencing replicates. We validate selected singleton variants using capillary sequencing.

Using these strategies, we found that individual-based analyses identified the most high quality singletons. However, individual-based analyses generated more missing genotypes than population-based and LD-aware analyses. Individual-based genotypes were the least concordant with array-based genotypes and replicates. Population-based genotypes were less concordant than genotypes from LD-aware analyses with extended haplotypes. Therefore, we recommend population-based analyses for high quality variant calls with few missing genotypes. With extended haplotypes, LD-aware methods generate the most accurate and complete genotypes. Finally, individual-based analyses should complement the above methods to obtain the most singleton variants.

Segregating sources of sequencing error in NGS reads

Variant calling in NGS studies are often complicated by sequencing artefacts and errors; such artefacts and errors typically lead to low-quality variant calls [86, 91]. Rare variant discovery requires particular care because sequencing errors are often mistaken as rare variants [12, 52, 55]. Methods have been developed to detect and filter variant calls that are likely false positives [20, 40]. An additional set of methods have been developed to remove the effect of sequencing artefacts from the reads [40, 76, 91] and to adjust for known errors [53] prior to variant calling. Despite these efforts, not all error sources in a sequencing experiment are known and accounted for [100]. In particular, variant calling relies on the base quality score of each sequenced

base, but the score reported by the sequencing machine typically reflects multiple sources of errors, and requires recalibration to reflect the empirical error rates [76]. To properly adjust for these errors, we need to understand and disentangle error sources in an NGS experiment.

In chapter 3, we proposed a statistical model for estimating two common sources of error in an NGS experiment, namely machine error which arises from the sequencing machine and fragment error which occurs in the DNA fragment prior to base-calling. These two errors can be separately estimated from overlapping read pairs, arising from short fragments. Overlapping read pairs replicate any errors in the underlying fragment sequenced. However each read in the pair is an independently read by the sequencing machine, hence generating two estimates of the machine errors. We proposed models for machine errors and fragment errors based on concordance and discordance of overlapping bases, using base quality scores and read cycles as predictors. We designed a Markov chain Monte Carlo algorithm to sample from the posterior distribution of the errors, and analyzed 10 samples with over half of the reads overlapping.

We found that machine errors were mainly predicted by reported base quality scores, while they were mostly constant across read cycles, with only a slight increase in error rates at the last few cycles. These error rates were uniform across samples from different plates and lanes, suggesting that machine errors are consistent within a sequencing experiment. As for fragment errors, they were also uniform across samples. However, we found that fragment errors were predicted by the base quality scores of concordant overlapping bases, as opposed to the previous assumption that fragment error rates are uniform across the genome [20]. Therefore, our models demonstrated the utility of overlapping reads for better understanding the error

sources in sequencing experiments.

Resolving overlapping reads in NGS data

Overlapping read pairs are common sequencing artefacts in many NGS studies. Since these read pairs replicate fragment error at the overlapping regions, overlapping reads need to be treated before variant calling; falsely assuming independence would lead to overestimation of genotype calling accuracy, inflating the number of singleton variants. To address this problem, some studies have soft-clipped (discarded) one of the overlapping reads in each pair. This solves the dependence problem, but is an overcorrection because each read still contain independent machine errors. In chapter 4, we proposed an algorithm, RESCORE, to retain the combined machine information from both reads while removing the fragment dependence. We represented the combined machine error estimate from a pair of overlapping bases as a temporary quality score, and utilized base quality score recalibration to map the temporary scores to reflect the empirical mismatch rate at each position.

We then applied RESCORE to analyze 40 samples from a whole-genome sequencing study with 8x coverage, where each sample had over half of its reads overlapping. RESCORE increased the recalibrated base quality scores for the majority of overlapping bases when compared to soft-clipping the overlapping regions. This increment led to an almost 20-fold decrease in the estimated false positive rate of novel variants, which would result in the discovery of 0.027% additional variants that were likely to be genuine. Therefore we recommend incorporating RESCORE as a standard data processing step in NGS analysis pipelines.

Whole-genome sequencing of uropathogens

Genetic variation among pathogenic bacterial strains can be informative of the evolution and diversity of infectious disease phenotypes [8, 66, 102]. In Chapter 5, we presented an application of whole-genome sequencing for understanding the evolutionary history of uropathogenic *Escherichia coli* (*E. coli*). Uropathogenic *E. coli* (UPEC) are phenotypically and genotypically very diverse. This diversity makes it challenging to understand the evolution of UPEC adaptations responsible for causing urinary tract infections (UTI). To gain insight into the relationship between evolutionary divergence and adaptive paths to uropathogenicity, we sequenced at deep coverage (190x) the genomes of 19 *E. coli* strains from urinary tract infection patients from the same geographic area. Our sample consisted of 14 UPEC isolates and 5 non-UTI-causing (commensal) rectal *E. coli* isolates.

We developed a novel pipeline for the phylogenetic analysis of the *E. coli* strains. We identified strain variants using *de novo* assembly-based methods. Based on pairwise sequence differences across the whole genome, we clustered the strains using a neighbor-joining algorithm. We examined evolutionary signals on the whole-genome phylogeny and contrasted these signals with those found on gene trees constructed based on specific uropathogenic virulence factors.

The whole-genome phylogeny showed the divergence between UPEC and commensal *E. coli* strains without known UPEC virulence factors happened over 32 million generations ago, which is equivalent to 107,000- 320,000 years [88]. Pairwise diversity between any two strains was also high, suggesting multiple genetic origins of uropathogenic strains in a small geographic region. Contrasting the whole-genome phylogeny with three gene trees constructed from common uropathogenic virulence

factors, we detected no selective advantage of these virulence genes over other genomic regions. These results suggest that UPEC acquired uropathogenicity a long time ago and used it opportunistically to cause extraintestinal infections.

Conclusion

Next-generation sequencing technologies are rapidly advancing, as are statistical analysis and methodology developments, in order to efficiently and accurately process the high-throughput data. This dissertation develops methods in the upstream processing of NGS data and provides an application of NGS analysis for understanding the evolution of bacterial virulence. These studies together provide novel directions and promising perspectives for analyzing NGS datasets, thereby improving the confidence in the interpretation of NGS data.

CHAPTER II

Comparing Variant Calling Algorithms for Target-Exon Sequencing in a Large Sample

2.1 Introduction

With rapid advances in sequencing technology, large-scale sequencing studies enable discovery of rare polymorphisms. Exome and targeted sequencing studies are especially popular in the studies of complex traits. These designs focus on small genome regions likely to be enriched for functional variants [6, 67, 115], achieving higher coverage of an important subset of the genome and facilitating larger sample sizes [41, 68]. While variant calling typically improves with increasing read coverage [7], exome and targeted experiments tend to generate uneven coverage. For studies averaging 40x to 120x, empirical coverage per targeted position per sample can range from less than 5x to over 150x [11, 75, 84, 136]. At high coverage, genotypes can be called with high precision using basic calling strategies [6]. However, at regions with local low coverage, calling genotypes accurately is challenging, leading to more errors and missing data [17]. In studies with low mean coverage, advanced variant calling algorithms compensate by combining read information with linkage disequilibrium (LD) information across large samples [61, 125]. However, it is unclear if such algorithms substantially improve genotypes in datasets with heterogeneous coverage.

This chapter is published as Lo, Y. et al. 2015. *BMC Bioinformatics*, 16(1), 75.

To address this question, we evaluated the performance of advanced variant calling algorithms in targeted sequencing experiments. Our goal was to provide specific guidelines for applying variant calling algorithms to these studies.

Variant calling algorithms fall into three major categories depending on how information from shotgun sequencing data is aggregated across individuals and genomic positions [86]. The first category involves individual-based single marker callers (IBC), which assign genotypes based on aligned reads from a single individual at a single position [28, 58, 59, 60, 76]. These callers are typically applied to high-depth exome sequencing data [11, 84]. The second category of algorithms is population-based single marker callers (PBC), where reads per position from all samples jointly determine polymorphism and allele frequencies. Based on estimated allele frequencies, these methods then call genotypes using per individual read data [17, 49]. PBC is typically used in low-pass sequencing studies [61, 116, 117]. The third category of calling algorithms utilizes linkage disequilibrium (LD) information across several hundred kilobases flanking each variant base identified by an IBC or PBC [61, 70]. Similar to widely used imputation algorithms [9], these LD-aware calling methods (LDC) phase existing variant calls into haplotypes, then update genotypes according to the joint evidence across similar haplotypes. LDC, though computationally demanding, have been used in combination with PBC to successfully interpret low-coverage, genome-wide data such as that in the 1000 Genomes Project [116, 117].

To compare the performance of the three types of algorithms in large-scale sequencing datasets with high coverage, we analyzed 7,842 European individuals, each sequenced at 202 targeted genes [82]. The average per targeted site per individual coverage was 24x, but with a wide range from 0 to $>$ 75x (Figure 2.2). Genotype data from previous genome-wide association studies (GWAS) provided long haplo-

types for LD-aware genotype calling. We generated four sets of variant calls from this dataset, using (1) IBC, (2) PBC, (3) LDC based on only the sequencing data and (4) LDC after combining the sequencing data with flanking GWAS data. We focused on a fixed number of variants per call set after ranking the variants by quality control metrics, and assessed the quality of each filtered call set by transition to transversion ratio and the percentage of called variants confirmed in SNP databases. Moreover, we evaluated genotype accuracy by collating 80 pairs of experimental replicates and by comparing sequencing calls with on-target genotypes from previous GWASs. We further validated a subset of caller-specific singletons at the heterozygous individuals with an independent capillary sequencing experiment. Finally, to ensure applicability of our comparison findings to other studies, we investigated our dataset using alternative approaches of IBC and PBC. We also generated IBC variant calls from an additional dataset with average coverage of 127.5x, sequenced at 57 genes from 3,142 individuals [136], and compared these calls with their existing PBC call set.

We found that at a fixed number of variant sites, IBC identified a larger proportion of extremely rare variants of high quality, particularly singletons, while capturing most of the common polymorphic sites that were identified by the other callers. We replicated the result in the additional high-coverage dataset and by using different variant caller implementations. However, IBC genotypes at common variants were of the lowest quality by all measures. They were the least concordant with GWAS genotypes and within sequencing replicate pairs. Moreover, the IBC call set contained 4.72% missing genotypes, due to low coverage or low quality calls. In the PBC set, the percentage of missing genotypes dropped to 0.47% by using a population allele frequency prior. PBC also showed improved heterozygous concordance with on-target GWAS genotypes as well as between replicates. Without flanking

markers, LDC achieved similar genotype accuracy with PBC, while further reducing the missing genotypes to 0.17%. With extended haplotypes from flanking GWAS markers, LDC achieved the same level of missing genotypes (0.17%) and the highest genotype concordance among all callers.

2.2 Methods

2.2.1 Data description

To understand the strengths and limitations of individual-based, population-based and LD-aware variant calling methods, we analyzed sequence read data from 7,842 unrelated European individuals. The next-generation sequencing data was part of a large-scale targeted sequencing experiment generated for the purpose of identifying variants associated with 12 common diseases and cardiovascular and metabolic phenotypes, previously described in Nelson et al. [82]. This experiment targeted 2,218 exons of 202 genes of potential drug interest, covering 864kb ($\approx 1\%$) of the coding genome. Each exon was captured to include the coding sequence plus UTR and 50 bp flanking sequence on each end. Each sample had on average 0.6 million 100 bp paired-end Illumina reads, with overall average depth of 24x, but depth averaged per individual per targeted site ranged from 0x to over 75x (Figure 2.2a). In particular, six genes had low mean coverage (< 10x) across all exons; the mean coverage across gene regions and across individuals spanned a range of 7x to 35x (Figure 2.2b).

Among the 7,842 individuals considered, 80 were independently sequenced twice. All 7,842 individuals had been previously typed on one of Illumina (300k, 550k, 610k) or Affymetrix (500k, 6.0) chips for genome-wide association studies (GWASs). Prior to variant calling, we aligned reads using BWA 0.5.9 (`http://bio-bwa.sourceforge.net`) [56] with human genome build 36 as reference. We removed duplicate reads using Picard (`http://picard.sourceforge.net/`). We recalibrated base quality

scores using Genome Analysis Toolkit (1.0.5974) from the Broad Institute [76]. We combined the GWAS genotype data from various chips using PLINK [97] (See Section 2.6).

2.2.2 Variant calling

We used likelihood-based models for genotype and SNP calling, as outlined in Li et al. [61]. For each of the 10 possible genotypes (AA, AC, AT, AG, CC, CT, CG, TT, TG, GG) at each locus, the model computes genotype likelihood $Pr(reads|genotype)$. These likelihoods are calculated per genomic position with aligned reads. Base quality scores of the reads are refined using the base alignment quality (BAQ) adjustment to account for base calling error rates and mapping uncertainty [53]. Using Bayes' rule, these likelihoods are combined with a model-specific prior on the genotype $\pi(genotype)$ to generate posterior probabilities $Pr(genotype|reads)$. We considered 3 categories of calling algorithms that reflect how information is aggregated across individuals and positions.

Individual-based single marker caller (IBC)

IBC applies an individual based prior which assumes each allele has a probability $\theta = 0.001$ of being different from the reference. For variant sites, we assigned uniform prior probabilities for transitions and transversions to avoid bias in the evaluation based on transition to transversion ratio (Ts/Tv). By computing the genotype likelihoods using aligned reads per individual, the model assigns the most likely genotype when the posterior probability reaches a threshold of 99%; genotypes with lower posterior probability are marked as missing. We used glfSingle (`http://genome.sph.umich.edu/wiki/GlfSingle`) to call genotypes. By calling also the reference homozygous genotypes, we obtained the union set of all variant

sites and genotypes across all individuals.

Population-based single marker caller (PBC)

PBC uses a two-step procedure to call variants [61]. First, upon observing at least one read carrying a non-reference allele, the model applies a population genetic prior that estimates the probability of the site being polymorphic as a function of sample size, with per base pair heterozygosity of $\theta = 0.001$ under the stationary neutral model [126]. As with IBC, the model assumes a prior with uniform Ts/Tv. Second, per polymorphic site, PBC estimates the population allele frequency f using aligned reads from all individuals, assuming a biallelic site in Hardy-Weinberg equilibrium. These allele frequency priors combine with the likelihoods calculated per individual to generate posterior genotype probabilities. We used the PBC implemented as glfMultiples (`http://genome.sph.umich.edu/wiki/GlfMultiples`), which also generated variant calls for NHLBI GO Exome Sequencing Project (ESP) and contributed to 1000 Genomes Project analyses [114, 116, 117].

In this study, we used a posterior probability threshold of 99% for the most likely genotype, which was the same threshold as for the ESP [114]. To maintain independence between experimental replicates, we generated two call sets, each including 7,762 unique samples plus 80 samples, one from each sequence replicate pair.

LD-aware caller (LDC)

Starting from a set of variant calls, LDC updates the genotype of each individual at each marker using a Hidden Markov Model derived from the haplotype-based model used in the imputation software MACH [62]. The LDC algorithm starts with randomly phased haplotypes for each individual. Per iteration, the algorithm compares one sequenced sample with a randomly picked subset of haplotypes. It

updates each genotype or imputes missing genotypes, based on the similarity of the sample haplotype to the reference haplotypes. In addition to identifying the most likely genotype, LDC calculates the expected number of reference alleles carried by each individual (dosage). Per variant site, LDC also estimates the correlation coefficient R^2 between true allele counts and estimated allele counts, as a measure of imputation quality. This caller, previously used in low-pass sequencing studies [61, 116], has been implemented as ThunderVCF (`http://genome.sph.umich.edu/wiki/ThunderVCF`).

We used LDC to refine each of the two PBC call sets described above. We applied the standard setting of 30 iterations and 200 reference haplotypes per iteration. We considered two scenarios with different haplotype information: First, we applied LDC on short haplotypes, which consisted only of the PBC variant calls at the sequences captured in the sequencing experiment. Second, we created long haplotypes by combining PBC variant calls with GWAS-genotypes from flanking markers within 500 kb from both ends of each target gene. In both scenarios, we masked GWAS genotypes within the target regions and used these markers as measures of genotype quality.

2.2.3 Variant quality control

To remove potentially false variant calls caused by technical artifacts, we followed the filtering and support vector machine (SVM) approach used in the ESP [114] and Zhan et al. [136]. Initial filtering included quality metrics based on read alignments, nearby indels and excess heterozygosity (See Section 2.6.2). For LD-aware calls, we imposed an additional R^2 quality control criterion, which filters sites with $R^2 < 0.7$.

SVM generates a summary score for each site based on the initial quality metrics, classifying good and bad calls with respect to training call sets (See Section 2.6.2).

We ranked these scores and selected the 27,500 top-ranked variants per call set for comparison. We set the cutoff to compare only variants with positive SVM scores.

After selecting 27,500 top-ranked variants per call set from SVM classification, we filtered individual genotypes to discard those with more than 1% estimated error. From IBC genotypes, we removed and marked as missing the genotypes with PHRED quality score less than 20 or with genotype depth less than 7x. As the quality of PBC genotypes is less affected by individual genotype depth, we only filtered with PHRED quality < 20. Analogously, we filtered LD-aware genotypes with a posterior probability ratio $< 99 : 1$ between the genotypes with the highest and the second highest posterior probability. Comparing call sets

We compared 4 sets of 27,500 variants, generated using IBC, PBC, LDC without flanking haplotypes and LDC with flanking haplotypes. First, we evaluated the overall quality of each call set by calculating transition to transversion ratios (Ts/Tv), stratified by variant type as annotated by ANNOVAR (hg19, gencodeV7, `http://www.openbioinformatics.org/annovar/`) [123] and by minor allele count. Second, we compared our call sets to the Single Nucleotide Polymorphism database (dbSNP, release 135, `http://www.ncbi.nlm.nih.gov/SNP/`), a recent public archive of confirmed variants.

We then characterized IBC-specific variants and PBC-specific variants by their Ts/Tv and read coverage. Most of the IBC- and PBC-specific variants were singletons. We performed an independent capillary sequencing experiment on 32 IBC-specific and 41 PBC-specific singleton variants, sampled from individuals from the CoLaus study [82] carrying the singleton heterozygous genotypes (See Section 2.6.3). Error rates from this validation provided estimates of false discovery rates of caller-specific singletons. Finally, we extended the validation to 51 caller-specific singletons

with SVM scores below the cutoff, to assess the quality of discarded sites from each set.

We assessed genotype quality of each call set by four summary statistics: (1) The percentage of missing genotypes from no calls and filtered genotypes (2) The pairwise heterozygote mismatch rates (h_e) between our genotype calls from sequencing and the genotypes from GWAS chips at the on-target markers. h_e is defined as the number of genotypes called as heterozygous in one set but homozygous in the other, divided by the total number of heterozygous genotypes in both sets. (3) h_e for the 80 sequence replicate pairs, at variant sites where at least one individual per pair is heterozygous. (4) The shared variants between each pair of call sets and calculated the h_e between every pair of callers.

To investigate the effect of sample size on the difference in performance between IBC and PBC, we performed down-sampling analyses on our original dataset, evaluating the ability of the PBC caller to identify variants called as singletons by IBC. For simplicity, we focused on variants that were called as singletons in the full dataset of 7,842 individuals (IBC singletons). We generated random samples of 50, 100, 500, 1,000, 2,500 and 5,000 individuals from the original dataset by sequentially adding individuals and used PBC to call variants in each of the samples. For each down-sampled dataset, we calculated the proportion of IBC singletons identified by PBC and recorded the genotype quality of these PBC singletons. We repeated the full random sampling experiment 10 times and averaged the results.

To assess if our results were driven by the specific choice of calling algorithms, we applied the individual- and population-based settings of GATK UnifiedGenotyper (version 3.1.1-g07a4bf8) [17] to our original dataset. The UnifiedGenotyper follows the same genotype likelihood framework described above for variant calling. In par-

ticular, it uses the same model for individual- and population-based calling, where it estimates simultaneously the population allele frequency and most likely genotypes. To generate individual-based calls, population size is set to 1. We generated individual- and population-based variants for our targeted exon data with 7,842 samples. We compared the two resulting call sets, focusing on the singletons specific to each analysis.

To replicate our results in a second dataset with higher sequencing coverage, we considered an additional dataset obtained from the AMD Consortium, which sequenced 3,142 individuals at 57 genes from 10 age-related macular degeneration loci [136]. The average coverage was 127.5x, but 10% of the genes suffered from low average coverage of around 10x (See Section 2.6.4, Figure 2.3). We generated IBC variant calls and compared them with existing PBC variant calls of this dataset, obtained from the project investigators. We evaluated the IBC-specific singletons, particularly those at sites with local low coverage, and contrasted them with singletons identified by IBC and PBC.

2.3 Results

2.3.1 Summary of variant call sets

In the complete call sets of 7,842 individuals, the individual-based single marker caller (IBC) generated 31,970 variants while the population-based single marker caller (PBC) generated 29,147 variants. The LD-aware caller (LDC) modified genotypes from PBC, hence it generated the same number of variants. We filtered each call set separately and ranked the variants using a support vector machine (SVM). We observed 30,297 IBC, 27,690 PBC variants and 27,535 LDC variants with positive SVM scores. To compare call sets for a fixed call rate, we focused on the top 27,500 variant sites from each set. In the IBC set, 59.4% of the calls were singletons (MAF

Call set	All SNPs					Singletons		%Missing genotypes
	#SNPs	%dbSNP	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	#SNPs	Ts/Tv	
IBC	27500	25.72%	3.02	2.54	2.71	16325 (59.36%)	2.57	4.71
PBC	27500	26.87%	3.02	2.45	2.59	15877 (57.73%)	2.44	0.47
LDC	27500	26.85%	3.01	2.45	2.59	15857 (57.66%)	2.44	0.17
LDC+F	27500	26.81%	3.00	2.45	2.58	15869 (57.71%)	2.44	0.17

Table 2.1: Summary statistics of 27,500 top-ranked SNPs per call set and quality assessed by transition-to-transversion ratio (Ts/Tv). Abbreviations: IBC = individual-based single marker caller, PBC = population-based single marker caller, LDC = LD-aware caller without flanking haplotypes, LDC + F = LD-aware caller with flanking haplotypes. Expanded table showing quality of call sets broken down by variant class is included in Section 2.6 Table 2.4.

= 0.06%), while 57.7% of the PBC and LDC calls were singletons (Table 2.1). Over 81% of variants in each call set had minor allele counts ≤ 5. Most of these rare variants were novel; only 26-27% of variants from each call set were recorded in the dbSNP database (Table 2.1).

Combining our four filtered call sets each of 27,500 SNPs, our analyses generated a total of 29,652 autosomal SNPs. We identified 1,035 variants not previously found in the Nelson et al. analyses of the same dataset [82]. Among these, 509 (48.16%) were IBC-specific, while 445 (42.10%) were in all call sets. The IBC call set had the highest percentage of missing genotypes (4.72%), while the PBC call set had a substantially lower percentage (0.47%) (Table 2.1). The LDC call set had the lowest percentage of missing genotypes (0.17%). Typically LDC genotypes have no missing data; in our analysis, missing genotypes in LDC were a result of filtering genotypes with more than 1% uncertainty.

2.3.2 Overall quality of variant call sets

We assessed the quality of the variants included in the four call sets by calculating the transition-to-transversion ratio (Ts/Tv). A Ts/Tv > 2 is expected for intergenic sites; Ts/Tv is typically much higher in coding regions due to purifying selection

[114]. In our data, Ts/Tv of the unfiltered IBC call set was 2.27, and Ts/Tv of the unfiltered PBC and LDC call sets were both 2.46. Ts/Tv of all call sets increased after SVM classification at the 27,500 variant cutoff (Table 2.1), indicating reasonable quality control. We then focused on the quality of these SVM top-ranked call sets. As Table 1 shows, the IBC call set attained the highest Ts/Tv of 2.71, while PBC and LDC without flanking haplotypes had a Ts/Tv of 2.59. LDC with flanking haplotypes had a Ts/Tv of 2.58.

Comparing Ts/Tv between known variants and novel variants, we observed that known variants (in dbSNP) generally had higher Ts/Tv than novel variants (Table 2.1). Singletons had slightly lower Ts/Tv compared to the corresponding overall call set, as singletons represent recent mutations that are less affected by purifying selection [105]. Analogously, known variants had a higher Ts/Tv because such variants are typically older and have been subjected to purifying selection for longer.

At exonic variants, all call sets attained Ts/Tv greater than 3, with nonsynonymous variants having lower Ts/Tv than synonymous variants (Table 2.4). The coding variants had higher Ts/Tv than non-coding variants in all call sets, because coding sequences contains higher proportion of CpG sites enriched for transitions compared to non-coding regions, and because transitions are enriched at degenerate sites within coding regions. Intergenic and flanking variants had Ts/Tv around 2 in all call sets, consistent with expectations (Table 2.4).

2.3.3 Evaluating singleton variants

Most caller-specific variants were singletons. We found 4,203 caller-specific variants out of 29,652 in the union call set. Of these, 1,850 (44.02%) were IBC-specific, 1,787 (96.59%) being singletons with Ts/Tv 1.97. On the other hand, 1,731 (41.18%) variants were shared between PBC and LDC sets, but not found by IBC. We consid-

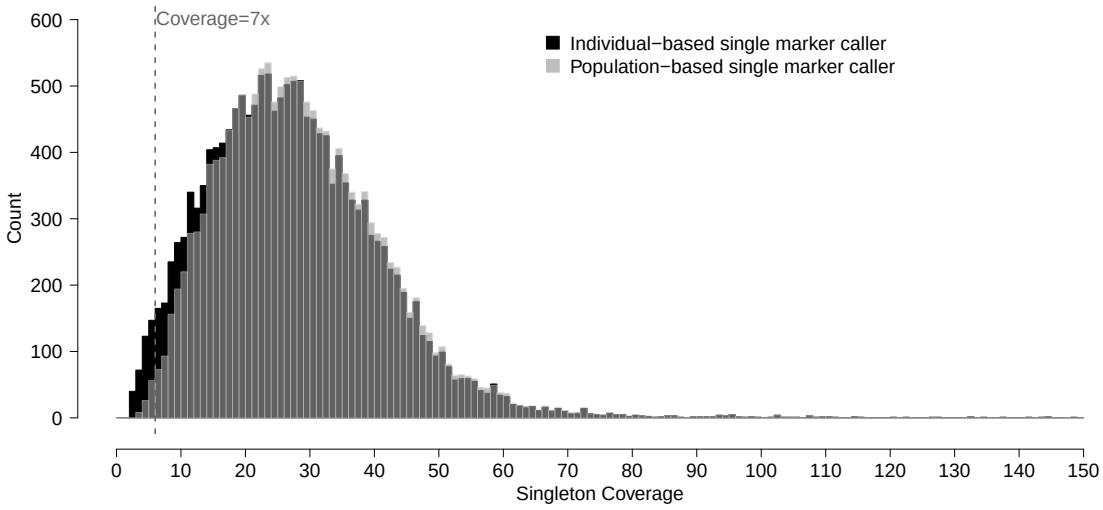


Figure 2.1: Distribution of coverage at the individual carrying the singleton alternative allele. We compare the distribution of coverage at called singleton variants between individual-based caller (black) and population-based caller (light gray). The overlap of the two distributions is in dark gray. Here we show all singleton variants after SNP filtering and genotype filtering on quality < 20. We keep individual-based single marker calls at low genotype coverage for this comparison, with the vertical dash line indicating genotype coverage filter at 7x.

ered sites in this category as PBC-specific since LDC did not introduce new sites, but only modified genotypes at sites called by PBC. Of these PBC-specific sites, 1,260 (72.79%) were singletons with Ts/Tv 1.08.

IBC identified more singletons at low coverage than PBC, even after an additional filtering of all genotypes with less than 7x coverage (Figure 2.1). Independent capillary sequencing experiment validated 30 out of 30 (100%) IBC-specific singletons, and 38 out of 41 (92.68%) PBC-specific singletons (Table 2.5). This difference in validation rates was not statistically significant (Fishers exact p-value $= 0.258$). Relaxing the SVM threshold to 29,000 SNPs per call set, IBC-specific and PBC-specific singletons still had comparable validation rates, at 91.30% (42/46) and 92.45% (49/53) respectively.

Notably, 99.13% of PBC-specific sites were in the IBC unfiltered (complete) call set of 31,970, including all 471 sites with minor allele count > 1. On the other hand,

only 177 (9.57%) IBC-specific sites were in the PBC complete call set of 29,147; the majority was undiscoverable using PBC. Therefore, we extended the validation experiment to IBC-specific singleton calls ranked below 29,000, where no singletons from PBC could be sampled from the CoLaus subset. Capillary sequencing showed that these IBC-specific singletons at the lowest ranks had a validation rate of 81.82% (18/22; Table 2.5).

To compare the performance of singleton calling between IBC and PBC in a different dataset with higher average coverage, we repeated these analyses on a targeted sequencing dataset of 3,142 individuals sequenced at a mean coverage of 127.5x [136]. We generated an IBC call set which contained 33,615 variants with Ts/Tv 2.12, while the existing PBC call set contained 31,527 variants with Ts/Tv 2.10. Comparing these two call sets, IBC called 1,913 more singletons than PBC. These additional singletons had Ts/Tv 1.63. Interestingly, the additional singletons with high quality were located in regions with low coverage. At depth $<$ 10x and with an extra genotype quality filter of $>$ 10, IBC identified 864 additional singletons with Ts/Tv 2.18. At the same genotype depth and quality thresholds, IBC and PBC shared 911 singleton variant calls with Ts/Tv 2.13 (Figure 2.4). When we relaxed the genotype depth threshold to $<$ 20x, IBC identified 1,360 additional singletons with Ts/Tv 1.90, while IBC and PBC shared 2,745 singletons with Ts/Tv 2.07.

We evaluated the impact of sample size on the difference in performance between IBC and PBC by down-sampling the data to sample sizes of 1, 50, 100, 500, 1,000, 2,500 and 5,000 and calling variants in these smaller datasets using PBC. We compared the PBC singletons from each down-sampled set to high-quality IBC singletons from the original dataset of sample size 7,842. We observed that for sample sizes $>$ 1, PBC failed to identify all IBC singletons. The proportion of IBC singletons called by

		Heterozygous mismatch rate			
		IBC	**PBC**	**LDC**	**LDC+F**
(a)	All samples at 378 GWAS markers	0.82%	0.38%	0.39%	0.32%
(b)	80 sequence replicate pairs at all called variants	1.01%	0.34%	0.36%	0.20%
(c)	Pairwise comparison of callers				
	vs PBC	0.42%	–	–	–
	vs LDC	0.93%	0.35%	–	–
	vs LDC+F	1.01%	0.41%	0.30%	–

Table 2.2: Heterozygous mismatch (a) between sequence calls and GWAS genotypes at 378 on-target GWAS markers, (b) between 80 sequence replicate pairs and (c) between pairs of algorithms.

PBC decreased as sample size increased. The quality score of the singletons called by PBC also decreased with sample size. At sample size = 100, PBC called 89.6% of the IBC singletons with average quality score of 73.7; at sample size = $5,000$, the percentage dropped to 84.0% with average singleton quality score 69.5 (Figure 2.5).

2.3.4 Evaluating non-singleton variants

We assessed genotype quality of common variants by comparing genotypes at 378 on-target variants shared between all call sets and the GWAS data from the same individuals (Table 2.2a). The IBC call set had the highest discordance with GWAS genotypes, with heterozygous mismatch $h_e = 0.82\%$ discordant genotypes. While heterozygous mismatch rates were comparable between PBC and LDC with no flanking haplotypes, at $h_e = 0.38\%$ and 0.39% respectively, the rate was lower for LDC with flanking haplotypes, at 0.32% (Table 2.2a).

Genotype concordance between sequencing replicate pairs provided a second metric of robustness of each calling algorithm (Table 2.2b). h_e at replicate pairs followed the same qualitative trend as the GWAS comparison (Table 2.2a), where IBC had the highest $h_e = 1.01\%$ at replicate pairs. The heterozygous mismatch rates were 0.34% for PBC and 0.36% for LDC without flanking haplotypes. With flanking haplotypes, $h_e = 0.20\%$ between experimental replicates of LDC. This mismatch rate was lower than the h_e with GWAS genotypes, suggesting that the error rate of

chip-based genotyping was higher than the error rate for LDC genotypes.

The non-missing genotypes between each pair of call sets had less than 1% heterozygote discordance (Table 2.2c). IBC and PBC call sets had low discordance, with $h_e = 0.42\%$. The PBC and LDC call sets also had similar discordance, with $h_e = 0.35\%$ and 0.41% respectively. The two LDC call sets were the least discordant, with $h_e = 0.30\%$. IBC and LDC call sets had higher heterozygous discordance, with $h_e = 0.93\%$ between IBC and LDC without flanking haplotypes, and $h_e = 1.01\%$ between IBC and LDC with flanking haplotypes. These mismatch rates were consistent with the above comparisons with GWAS genotypes and between sequence replicates (Table 2.2).

Complex calling algorithms called additional genotypes at sites that had missing calls at less complex calling algorithms (Table 2.3a). To evaluate specifically the quality of these additional sites, we calculated the heterozygous mismatch rates with GWAS genotypes (Table 2.3b). Comparing each algorithm with progressively more complex alternatives at the 378 on-target variant sites with GWAS information, we observe that the PBC call set contained 15,727 (5.68%) more heterozygous genotypes than the IBC call set, with $h_e = 0.85\%$. Thus PBC generates high-quality genotypes at most sites that cannot be called with IBC. LDC without flanking haplotypes generated 3,113 (1.06%) while LDC with flanking markers generated 3,664 (1.25%) more heterozygous genotypes than PBC. Mismatch rates in these extra genotypes varied widely between the two settings; calls from LDC without flanking markers had a mismatch rate of 2.41% while calls from LDC with flanking markers had an error rate of 0.71% (Table 2.3b).

	All samples at 378 GWAS markers		IBC	PBC	LDC	LDC+F
(a)	Number of heterozygous genotypes (hets)		276,761	293,730	298,220	298,531
	Heterozygous mismatch		0.82%	0.38%	0.39%	0.32%
(b)	Number of additional hets and heterozygous mismatch	not in IBC		15,727 (0.85%)	17,937 (1.23%)	18,308 (0.47%)
		not in PBC			3,113 (2.41%)	3,664 (0.71%)
		not in LDC				1,145 (0.87%)

Table 2.3: Heterozygous mismatch (a) between each call set and GWAS genotypes at 378 on-target markers, and (b) between additional heterozygous genotypes in more complex algorithms and the GWAS markers.

2.3.5 Alternative implementations of variant callers

To evaluate the consistency of these observations across other implementations of variant callers, we analyzed the same dataset using GATK UnifiedGenotyper. We generated individual-based (G-IBC) and population-based (G-PBC) call sets. The G-IBC call set contained 34,704 variants with Ts/Tv 2.21, while the G-PBC call set contained 33,696 variants with Ts/Tv 2.23. Each call set contained about 32% singletons: the G-IBC call set contained 11,001 singletons with Ts/Tv 2.13, and the G-PBC call set contained 10,678 singletons with Ts/Tv 1.77. The proportion of singletons was substantially higher in our IBC call set (59.36%) generated using glfSingle and PBC call set (57.73%) generated using glfMultiples, as well as in previous analyses of the same dataset (60.32%) [82] using the SOAP caller [59]. Since a high proportion of singletons identified by glfSingle and glfMultiples have been experimentally replicated or validated (see above), GATK UnifiedGenotyper is conservative when calling singletons. Nevertheless, G-IBC identified about 3% more singletons than G-PBC and these had significantly higher Ts/Tv, replicating the pattern observed in our analyses using glfSingle and glfMultiples.

2.3.6 Multi-allelic variants

IBC identified 523 on-target SNPs with more than one non-reference allele. Of these, 513 SNPs (1.87% of 27,500 IBC SNPs) had two non-reference alleles (triallelic)

and 10 had three non-reference alleles. Following the population genetics calculations used in Nelson et al. [82], we predicted that $\sim 0.9\%$ of variants would be triallelic and that a third allele would be called at 0.5% of biallelic sites due to sequencing error. Under a model of homogeneous mutation rate, we would thus expect a proportion of $\sim 1.4\%$ observed triallelic SNPs. Similar to others [30, 82], we observed an excess of triallelic SNPs.

Most of the triallelic variants were rare: 205 (38.53%) had two singleton non-reference alleles, and 253 (47.56%) had one singleton non-reference allele and one more common non-reference allele. For the 10 SNPs showing all four alleles, 8 had at least one singleton non-reference allele. Nelson et al. [82] validated 10 out of 10 singleton triallelic variants from the same dataset. Among the 523 multi-allelic variants called by IBC, PBC called 509 biallelic, identifying the non-reference allele with more information (higher allele frequency or higher read depth). PBC identified the remaining 14 multi-allelic SNPs as monomorphic.

2.3.7 Computational burden

The computational burden of variant calling increases when the algorithm aggregates more information across individuals and sites. Hence IBC is the fastest algorithm and LDC is the slowest. IBC used about 250 CPU-hours to generate all variants for all 7,842 individuals, while PBC used 400 CPU-hours. For IBC, each individual at a specific genomic region can be analyzed in parallel. For PBC, all individuals have to be considered jointly, but genomic positions are independent and can be analyzed in parallel. In terms of memory usage, IBC consumed negligible memory since it only needed to read in the genotype likelihoods for one position per individual. For PBC, memory consumption increased roughly linearly with sample size. To analyze our dataset with 7,842 individuals, the maximum memory usage was

7.9 Gb. In our down-sampling analyzes, sample sizes of 1,000 and 5,000 consumed 1.1 Gb and 5.3 Gb memory respectively.

The LDC model considers all haplotypes jointly, with run time increasing in quadratic scale with the number of haplotypes included in the reference panel, which is the state space of the underlying Hidden Markov Model. Other factors affecting run time included length of each haplotype, number of iterations, and total sample size. We performed LD-aware calling per gene for 15,684 haplotypes at 202 genes, using a reference panel size of 200 for 30 iterations. After running PBC, LDC with flanking haplotypes took about 3000 CPU-hours. Without flanking haplotypes, LDC took about 2000 CPU-hours. To speed up the process while retaining sufficient LD information, LDC can be run in parallel on larger genomic regions, such as a 1Mb region or a chromosome. Memory usage increased linearly with the number of variants in the gene: each gene contained 300 to 2,000 variant sites after adding GWAS flanking genotypes, with the memory required for running LD-aware algorithm ranging from 45 Mb to 300 Mb.

We performed all analyses on a Dell C6100 blade server with four discrete dual 6-core Intel Xeon X5660 CPUs at 2.80 GHz. 128 GB RAM and 1 TB of local SATA disk were available on this system.

2.4 Discussion

We performed an extensive comparison between calling algorithms of various complexity on a large sequencing dataset capturing exons of 202 drug-targeted genes with mean coverage of 24x. As a result of the capturing process necessary for targeted sequencing, we observed a wide range of coverage per targeted position, echoing the outcomes of other exome sequencing studies aiming at high coverage [83, 136]. Thus,

our work provides general guidelines for using variant calling algorithms on exome and targeted sequencing datasets.

Existing calling algorithms aggregate different levels of information from sequence reads. We considered three major groups of likelihood-based models: (1) Individual-based single marker caller (IBC) uses aligned reads at each marker per individual, (2) population-based single marker caller (PBC) uses aligned reads at each marker for all samples to estimate population allele frequency, (3) LD-aware genotype refinement caller (LDC) uses linkage disequilibrium information from loci surrounding each called variant. Many different approaches exist for each model; each uses a variation of individual-based, population-based or haplotype-based priors. Previous studies have shown comparable performance between glfSingle/glfMultiples and earlier versions of the GATK UnifiedGenotyper [64]. By comparing sets of IBC and PBC from the same developer, we observed excess high-quality singletons in individual- over population-based algorithms.

Comparing filtered call sets of identical size (27,500) for each caller, IBC discovered more rare variants than PBC. In particular, at lower coverage, IBC was able to identify more high-quality singletons than PBC. We replicated this result twice, in a second dataset with higher coverage and in the original dataset using a different approach of the callers. We observed that the ability of PBC to detect singletons depended on sample size: With increasing sample size, PBC identified fewer singletons, and the quality of the identified singletons decreased. This advantage of IBC over PBC can be partly explained by the fact that in larger samples, singletons have an allele frequency < 0.001. Hence the prior for a site being a singleton is stronger in the individual-based caller and less evidence is required to call a singleton.

While we found significant differences between caller-specific sites, IBC and PBC

call sets had > 99% concordance at the high-quality, non-missing heterozygous genotypes. Our validation experiment confirmed all selected IBC-specific singletons, with very few unconfirmed singletons in the PBC call set. Moreover, most PBC-specific singletons were in the IBC unfiltered (complete) call set. We observed the same trend of IBC generating an augmented set of singletons in high coverage sequencing data (> 120x), where IBC almost doubled the number of high-quality singletons at sites with local low coverage (< 10x).

Furthermore, only IBC was capable of identifying polymorphisms with more than one non-reference allele, which led to discovery of an additional 1.9% of rare alleles in the sample. The excess of triallelic sites over the theoretical prediction of 1.4% is likely the result of heterogeneity of mutation rate due to sequence context and genomic environment. Existing associations between multiallelic variants and disease phenotypes [16, 31] suggest that properly accounting for such variants can increase the power of a sequencing study.

While IBC had strengths in identifying singletons, PBC generated better overall genotype quality. At common variants, PBC genotypes overcame low coverage at specific samples, achieving fewer missing genotypes and higher accuracy than IBC calls. The discordance between IBC and GWAS genotypes was low (0.82%), but more than two times higher than the GWAS discordant rates of the other call sets.

LDC achieves even higher genotype accuracy than IBC and PBC by using haplotype information to impute missing genotypes from an existing single-marker call set. Imputation is typically more effective with longer haplotypes. In our study, we created long haplotypes by combining sequencing data with SNPs from previous GWAS genotyping chips. LDC with such flanking haplotypes achieved the highest accuracy and the least missing genotypes. As targeted sequencing studies might not

have chip data to generate long haplotypes, we studied if LDC would still improve genotype accuracy with haplotypes based only on the sequencing data. Without flanking haplotypes, LDC had fewer missing data at the common variants over PBC, yet with a slightly higher mismatch rate. In particular, the additional heterozygote genotypes at common GWAS markers had a high mismatch rate of 2.43%, despite an overall mismatch rate of 0.39%. This suggested that using LDC on short haplotypes to impute missing genotypes created a relatively large number of imputation errors. Comparison between sequence replicates further demonstrated that LDC without flanking haplotypes had minimal benefit over PBC. As LDC imposes a considerable computational burden, it seems questionable whether this caller should be used when flanking haplotypes are not available.

2.5 Conclusions

In summary, while IBC generated high quality unique singletons, as well as multiallelic variants, its resulting call set contained more missing genotypes and genotyping errors at common variants. PBC calls showed a substantial decrease in the number of missing genotypes and errors over IBC calls at these variants. Only when flanking haplotypes were available, LDC calls showed noticeable refinement of PBC genotypes, resulting in a call set with the highest concordance with GWAS genotypes and between experimental replicates. Therefore, IBC had strengths in calling extremely rare variants, while PBC combined with LDC had strengths in calling the more common variants.

Based on these results, we recommend a two-fold calling strategy for targeted sequencing studies with medium to high coverage in a large sample. We recommend first to use a population-based single marker caller to generate accurate common

variants and most of the rare variants. Second, we recommend using individual-based single marker caller to enrich the call sets with additional singletons. If flanking markers around targeted regions are available, despite the computation burden, we recommend using LD-aware caller to refine and impute population-based calls at high accuracy, resulting in a complete call set.

2.6 Appendix

2.6.1 Pre-variant calling data processing

Sequence read data

We aligned reads using BWA 0.5.9 (`http://bio-bwa.sourceforge.net`) with human genome build 36 as reference. Average mapping rate was 99.7%; 98.5% of reads were properly paired. Using Picard (`http://picard.sourceforge.net/`) we identified and removed 21% duplicate reads. We recalibrated the base quality scores using GenomeAnalysisTK-1.0.5974 (`http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration`)[17].

Genotype data

We combined genotype data from previous GWASs typed on Illumina 300k, 550k, 610k and Affymetrix 500k and 6.0 using PLINK [97]. We identified 378 GWAS variants on the targeted regions. At these variants, we confirmed reference allele for A/T and G/C variants using sequencing calls and respective allele frequencies, and there were no strand flip issues. We discarded a small number of genotypes at the flanking regions based on ambiguous strand information.

2.6.2 Variant quality control

Initial filtering

We applied to each call set initial filters, which were based on read alignments at variant sites and summary statistics of each site. In particular, at each polymorphic site, we computed several Z-score test statistics of read alignments, including strand bias, allele balance and alternative allele inflation, with the detailed statistical tests described below. A SNP with extreme Z-scores indicate bias from mapping or sequencing artefacts which likely lead to false positive calls. Cutoffs for each filter

followed from the ones used in the NHLBI GO Exome Sequencing Project [114]. We further imposed an indel filter, which filtered SNPs located within 5 base pairs of known insertions or deletions from 1000 Genomes low-coverage CEU data (July 2010 release). We detected sites with excess heterozygosity than expected under Hardy-Weinberg equilibrium, calculated using inbreeding coefficient, described below. For LD-aware calls, we imposed an additional R^2 quality control criterion by filtering the sites with estimated squared correlation less than 0.7 between true allele counts and estimated allele counts. Here we describe in detail the filters used:

1. Strand bias: Conditioned on the site being biallelic, strand bias refers to higher than expected frequency of observing the alternate allele on the forward or the reverse strand. Specifically, the strand bias filter counts the number of reference and alternate alleles on each strand as a 2-by-2 contingency table. Under the null hypothesis, a genuine polymorphism should have the alternate allele observed equally often from forward and reverse strands. Therefore, the strand bias filter discards sites with normalized Z-score greater than 10 or absolute correlation greater than 0.15, which suggest strong association between strand and the allele observed.

2. Allele balance: Allele balance measures the ratio between allele counts from genotype calls and estimated allele counts calculated from individual sequence depth and likelihoods (`http://genome.sph.umich.edu/wiki/Genotype_Likelihood_Based_Allele_Balance`). A small ratio indicates bias towards certain alleles at a called polymorphic site, which is likely to be false positives. We imposed a lower bound of 67% on the allele balance ratio for good quality SNPs.

3. Alternate allele inflation: Alternate allele inflation is a composite measure of

base quality inflation and alternate allele quality inflation. We count the number of third and fourth alleles observed at a biallelic site and test it against the expected value where third and fourth alleles only occur due to sequencing error. Large normalized Z-score of this test indicates there are more non-reference, non-alternate bases than expected by base quality, suggesting that base quality is over-recalibrated due to alignment artefacts. Alternate allele quality inflation measures the normalized deviation of the number of alternate allele calls from the actual number of alternate bases observed from the reads. A small Z-score provides stronger support of the site being polymorphic, as the alternate base is observed much more frequently than the other two bases besides the reference. The composite alternate allele inflation statistic is the sum of the two Z-scores described above. We filtered out sites having absolute composite score greater than 5, which means they are called polymorphic because of alignment artefacts that lead to inflated quality scores.

4. Excess heterozygosity: We measured deviation from Hardy-Weinberg equilibrium (HWE), in particular the excess of heterozygotes, by calculating the inbreeding coefficient F for each marker, where

$$F = 1 - \frac{\text{Observed number of heterozygous genotypes}}{\text{Expected number of heterozygous genotypes}}$$

The expected number of heterozygotes comes from assuming HWE, such that

$$E(het) = 2p(1-p)N$$

Here N denotes the sample size. F ranges from $(-\infty, 1]$, with positive values representing markers with fewer heterozygotes than expected. $F = 0$ means the marker is in perfect HWE. Negative F denotes an excess heterozygotes at that

marker. We set the cut-off at -0.1, meaning that we discard markers with more than 10% heterozygotes observed than expected under HWE.

Support Vector Machine (SVM) filtering

Second, based on the initial set of variant quality metrics, we used a support vector machine (SVM) approach to generate a summary quality score for each variant site [40]. This approach was also applied in filtering and generating consensus calls in ESP and 1000 Genomes Project [114, 117]. The SVM identifies a hyperplane separating a training set of good calls and bad calls and scores each variant site to reflect the distance of the SNP from this hyperplane. Good calls and bad calls are classified by contrasting the initial quality statistics between the SNP calls and the SNPs in positive and negative training sets respectively. We used HAPMAP3 and OMNI variant sites as positive training sets, and the calls that did not pass more than two initial quality metrics as the negative training set.

Genotype filtering

Third, after selecting a fixed call rate of 27,500 top-ranked variants per call set from SVM classification, we applied filters to individual genotypes to ensure quality of all genotypes under comparison, given each top-ranked variant site. From genotypes called by individual-based single marker caller (IBC), we removed and marked as missing the genotypes with PHRED quality score less than 20; we also removed genotypes with genotype depth less than 7x. The quality of genotypes called by population-based single marker caller (PBC) is less affected by genotype depth, hence we only filtered based on PHRED genotype quality < 20. Analogously, we filtered LD-aware genotype calls with a posterior probability ratio $< 99 : 1$ between the genotype with the highest posterior probability and the genotype with the second

highest posterior probability.

2.6.3 Validation experiment

We performed an independent Sanger capillary sequencing to validate singleton variants identified by IBC and PBC. We considered the singletons carried by individuals from the CoLaus study.6 Within this subset of individuals, we sampled from the top-ranked 27,500 variants 32 singletons called by only IBC and 41 singletons called by only PBC. We further extended the experiment to sequence some caller-specific singletons beyond our defined SVM ranking cutoff. For variants ranked between 27,501 and 29,000 in each call set, we sampled 16 IBC-specific singletons and 12 PBC-specific singletons from the CoLaus individuals. Since IBC called more variants than PBC, we sampled an additional 23 IBC-specific singletons at the tail of the SVM rankings ($> 29,000$). We performed capillary sequencing on these 124 singletons on the individuals carrying the heterozygous genotype.

After PCR amplification of sequences of the 124 singletons using designed primers, we performed Sanger sequencing on the PCR products. We performed both steps at the University of Michigan facilities. We designed PCR primers using NCBI Primer-BLAST program. In case the program was not able to pick the primers, we manually designed primers sequences and ran them through BLAST search for specificity. We amplified PCR amplicons using OneTaq hot start 2X mixes (NEB, USA) with standard or GC buffer depending on the GC contents of the sequences. For samples that did not amplify in the first round, we assigned them new primers before repeating amplification. We set up PCRs using GeneAmp PCR System 9700 (Applied Biosystems, USA). We ran aliquots of the amplicons on 1% TBE agarose gel with Sybr Safe DNA Gel Stain and viewed them in UVP or Typhoon 9000 to visualize the amplicons and to check the quality and the quantity of the amplified

bands. We ran other PCR amplicons on the Agilent Bioanalyzer 2200 TapeStation (Agilent, USA) using the D1K screen tapes. We diluted amplicons before performing Sanger sequencing with the selected primers. We verified sequencing chromatogram data using Sequencher 5.1 demo (CGC, USA). We reported alleles by inspecting peaks on each chromatogram.

Among the 124 reactions performed, 3 failed. Among the 121 successful reactions, 71 were expected heterozygotes that passed our SVM quality control threshold. In this category, 3 out of 41 (7.32%) PBC-specific singletons were found to be homozygous reference, while all 30 IBC-specific singletons were confirmed. This difference in error rates between IBC and PBC was not statistically significant (Fisher's exact p-value = 0.258). Beyond our defined quality control threshold, at variants ranked between 27,501 and 29,000 in each call set, 4 out of 16 IBC-specific and 1 out of 12 PBC-specific singletons were not confirmed. At the tail of the SVM-ranked IBC call set, 4 out of 22 IBC-specific singletons were found to be homozygous reference, corresponding to a calling accuracy of about 82% for IBC at the sites of lowest quality (Table 2.5).

2.6.4 Evaluation of singletons on additional dataset

We applied individual-based variant calling (IBC) on 3,142 individuals from the AMD Consortium targeted sequencing dataset [136]. This sample was sequenced at 57 genes at 10 AMD loci, at 127.5x. Despite high average coverage, we observed highly heterogeneous coverage across targeted genes (Figure 2.3). Several genes are covered at less than or close to 10x. The population-based variant calling (PBC) of the same sample were previously performed and published by Zhan et al.[136]. After filtering the IBC call set using the same initial filters as in the PBC analyses, we compared the singleton calls identified by IBC and PBC.

Across the dataset, IBC called 1,913 additional singletons with genotype quality > 10 compared to PBC. These additional singletons had a Ts/Tv ratio of 1.63. Interestingly, the additional singletons with high quality were located in regions with low coverage. We found that at coverage < 10, and with an additional genotype quality filter of > 10, IBC identified 864 additional singletons not found in the PBC call set, with Ts/Tv = 2.18 (Figure 2.4 top). At the same genotype depth and quality thresholds, IBC and PBC shared 911 singleton variant calls with Ts/Tv = 2.13 (Figure 2.4 bottom). When we relaxed the genotype depth threshold to $<$ 20x, IBC identified 1,360 additional singletons with Ts/Tv = 1.90. At the same thresholds, IBC and PBC shared 2,745 singletons with Ts/Tv = 2.07.

2.6.5 Supplementary figures and tables

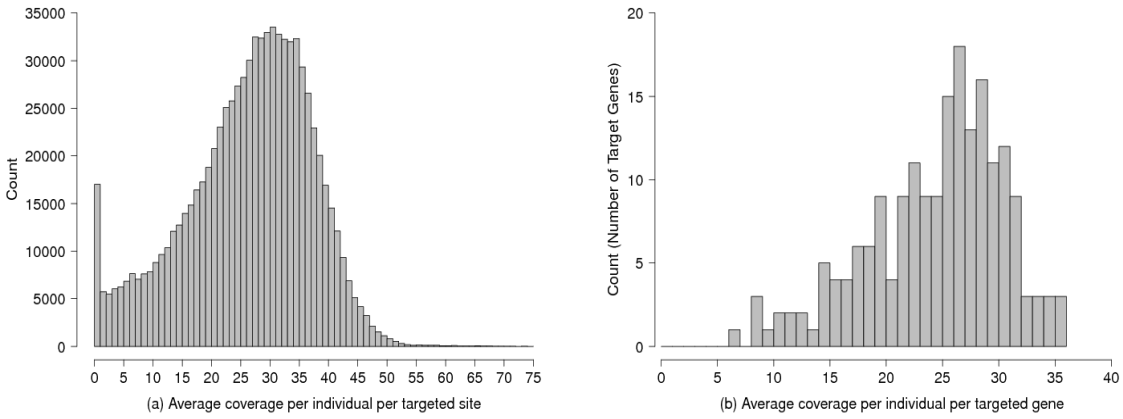


Figure 2.2: Distribution of average coverage of sequence read data from 7,842 unrelated European individuals. The next-generation sequencing data was part of a large-scale targeted sequencing experiment generated for the purpose of identifying variants associated with 12 common diseases and cardiovascular and metabolic phenotypes, previously described in Nelson et al [82]. This experiment targeted 2,218 exons of 202 genes of potential drug interest, covering 864kb (1%) of the coding genome (a) per individual per targeted genomic position, (b) per individual per targeted gene. The overall mean coverage is 24x.

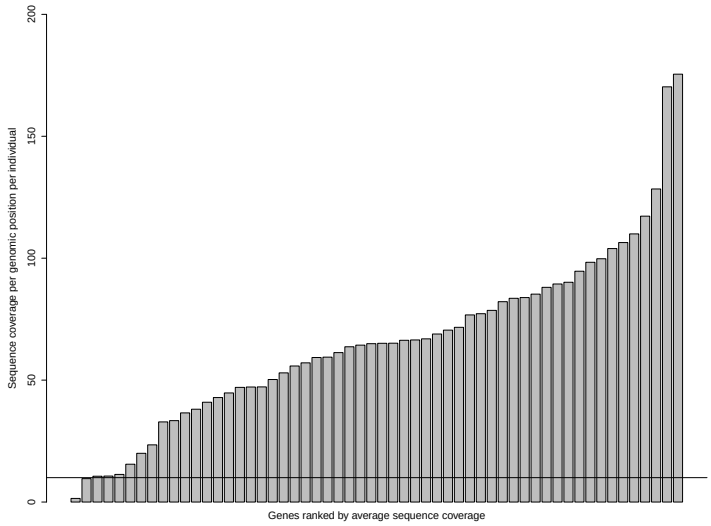


Figure 2.3: Average coverage at the 57 targeted genes in the AMD sequencing study [136], ranked by average coverage per individual per gene position. Overall average coverage across the whole targeted region was 127.5x. Horizontal line denote 10x coverage.

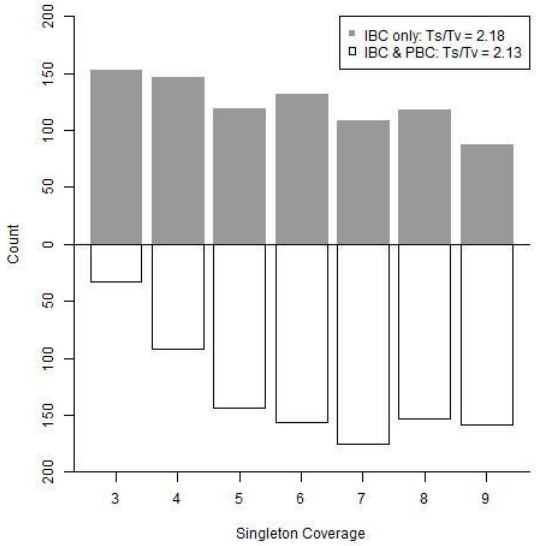


Figure 2.4: Distribution of AMD [136] singleton site coverage at the singleton-carrier at coverage < 10. All singletons shown in the figure have genotype quality >10. Top (gray): singletons identified by IBC only, Ts/Tv = 2.18. Bottom (white): singletons identified by both IBC and PBC, Ts/Tv=2.13.

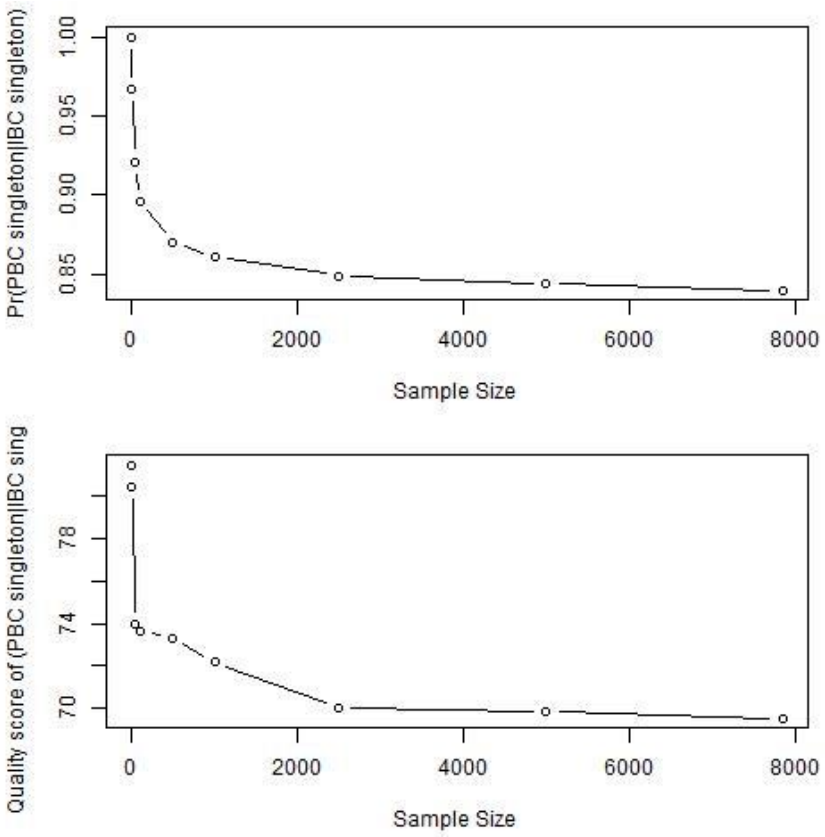


Figure 2.5: Proportion of IBC singletons identified by PBC at different sample sizes (top) and average quality score of IBC singletons identified by PBC at different sample sizes (bottom).

		All SNPs					Singletons	
Caller	Class	#SNPs	%dbSNP	Known Ts/Tv	Novel Ts/Tv	Overall Ts/Tv	#SNPs	Ts/Tv
IBC	**Total**	27500	25.72%	3.02	2.54	2.71	16325	2.57
	Nonsynonymous	6522	26.37%	2.92	2.30	2.50	4264	2.34
	Synonymous	4363	37.06%	14.90	5.31	5.60	2461	5.34
	Splice	130	16.15%	1.10	1.53	1.45	86	1.46
	Stop	163	0.25%	1.54	1.89	1.88	126	1.74
	UTR	10261	22.06%	2.50	2.30	2.39	5915	2.24
	Intronic	5610	23.46%	2.58	2.41	2.50	3213	2.49
	Flank*	341	24.93%	2.04	2.05	2.08	189	2.10
	Intergenic	110	14.55%	1.29	2.03	2.00	71	1.73
PBC	**Total**	27500	26.87%	3.02	2.45	2.59	15877	2.44
	Nonsynonymous	6547	27.19%	2.81	2.24	2.38	4222	2.25
	Synonymous	4377	38.15%	14.80	5.11	5.33	2415	5.16
	Splice	117	17.95%	1.33	1.74	1.66	76	1.81
	Stop	157	21.02%	2.30	1.88	1.96	119	1.77
	UTR	10285	23.11%	2.47	2.22	2.28	5759	2.12
	Intronic	5558	24.88%	2.65	2.29	2.37	3057	2.30
	Flank*	349	31.52%	2.33	1.91	2.03	160	1.86
	Intergenic	110	14.55%	1.67	2.03	1.97	69	1.76
LDC	**Total**	27500	26.85%	3.01	2.45	2.59	15857	2.44
	Nonsynonymous	6574	27.17%	2.78	2.24	2.37	4235	2.24
	Synonymous	4375	38.08%	15.05	5.13	5.37	2419	5.19
	Splice	119	17.65%	1.33	1.65	1.59	77	1.75
	Stop	157	21.02%	2.30	1.88	1.96	119	1.77
	UTR	10273	23.10%	2.46	2.22	2.28	5741	2.13
	Intronic	5549	24.82%	2.60	2.30	2.37	3044	2.31
	Flank*	342	32.75%	2.29	1.91	2.03	152	1.81
	Intergenic	111	14.41%	1.67	1.97	1.92	70	1.69
LDC+F	**Total**	27500	26.81%	3.00	2.45	2.58	15869	2.44
	Nonsynonymous	6570	27.12%	2.78	2.25	2.38	4235	2.25
	Synonymous	4378	38.05%	15.00	5.14	5.36	2419	5.20
	Splice	120	17.50%	1.33	1.61	1.55	78	1.69
	Stop	157	21.02%	2.30	1.88	1.96	119	1.77
	UTR	10265	23.08%	2.47	2.21	2.27	5742	2.11
	Intronic	5558	24.79%	2.60	2.29	2.36	3053	2.29
	Flank*	341	31.67%	2.38	1.88	2.02	153	1.73
	Intergenic	111	14.41%	1.67	1.97	1.92	70	1.69

Table 2.4: Quality of call sets assessed by transition-to-transversion ratio (Ts/Tv), broken down by variant class and frequency. Ts/Tv of each set at the top-ranked 27,500 SNPs via SVM classification were higher than the respective unfiltered call sets. Under a uniform Ts/Tv prior for all algorithms, IBC call set attained a higher Ts/Tv than the other call sets. Ts/Tv was higher at exonic variants and at known variants than at intronic variants and novel variants. Variants were classified using the ANNOVAR nomenclature (`http://www.openbioinformatics.org/annovar/annovar_gene.html`), with Splice including the splicing only sites, while the splice sites that lead to a stop codon were in the Stop class. *Flank refers to the upstream/ downstream variants within 50bp of the transcription site, as designed in the capture experiment described in Nelson et al. [82].

SVM ranking	Caller	Total reactions	Failed reactions	Confirmed	Not confirmed
$\leq 27,500$	IBC-specific	32	2	30	0
	PBC-specific	41	0	38	3
27,501-29,000	IBC-specific	16	0	12	4
	PBC-specific	12	0	11	1
$> 29,000$	IBC-specific	23	1	18	4

Table 2.5: Validation experiment results.

CHAPTER III

Markov Chain Monte Carlo Estimation of Sequencing Errors Using Overlapping Reads

3.1 Introduction

Next-generation sequencing (NGS) is a high-throughput, massively parallel process that generates thousands to millions of sequences of DNA simultaneously [72, 107]. NGS methods involve fragmenting genomic regions and sequencing each fragment [5, 77]. Typically, each fragment is read from both ends, generating paired-end short reads. These short reads are then reconstructed into longer sequences using alignment or assembly methods [57, 78]. Compared to traditional capillary sequencing approaches, which examines one base pair at a time, NGS methods makes the sequencing of full genomes highly feasible in terms of time and cost [63, 104, 107]. However, each sequencing step in a NGS experiment introduces a certain chance of error, as each short read is an imperfect realization of the underlying genome segment [100]. Thus, data generated using NGS methods tend to have lower accuracy than those generated using capillary sequencing, which can achieve per-base accuracy as high as 99.9999% [94, 98, 107].

In this chapter, we focus on single-base substitution errors occurring in NGS experiments. These errors can be broadly categorized into two types: First, for each position along a read, the sequencing machine calls a base and generates an

estimated base-calling error (machine error), which represents the uncertainty in the fluorescent detection of the bases [80, 112]. Second, prior to base-calling, fragmented DNA may contain errors from signal amplification via PCR or upstream preparation steps (fragment error) [3, 100].

While machine and fragment errors arise from completely different mechanisms and different steps of the sequencing process, they both show up as erroneous bases in the read, most commonly in the form of substitution errors. From the base calls and base quality scores generated by the sequencing machine, machine and fragment errors are indistinguishable. To account for both error sources in downstream variant calling, Flannick et al. [20] modeled fragment errors as a constant and machine errors from base quality scores. However, it is not clear whether fragment errors are constant within and between sequencing experiments. Multiple methods have been developed to use aligned read information to estimate the combined sequencing errors [32, 103, 124, 130]; however, due to the indistinguishability of the error types, to the best of our knowledge, there are no existing methods to separately estimate machine and fragment errors from the sequence reads. Knowing how much each source of error contribute to the observed mismatches can help identify areas of improvement in sequencing technologies. Understanding error sources can also provide insight into future sequencing experimental designs, particularly library designs that can isolate these errors from sequenced reads.

While normal single-end and paired-end reads cannot distinguish these errors, overlapping paired-end reads from short fragments provide an opportunity for separating fragment error from machine error. Overlapping read pairs occur when the pair of reads sequences the same part of the fragment twice. A normal read pair has the sum of read lengths greater than the length of the fragment sequenced. The

distance in terms of the physical genomic position spanned by the pair of reads is called the insert size [5, 63]. At an overlapping read pair, insert size is smaller than the sum of read lengths. Per overlap position, the pair of bases replicates the fragment error. However, an overlapping pair contains more than completely redundant information. The sequencing step is independent of fragment origin, meaning that each overlapping base gives two independent observations of base calls and quality scores. Therefore, in the simplest case, if two bases at a position of an overlapping read pair disagree, one of them must be due to machine error, because a fragment error would have shown up in both bases. On the other hand, if two bases at a position of an overlapping read pair agree, but they are different from the reference genome and they are not a known variant, it is very likely that there is an error in the underlying fragment.

In this chapter, we designed a statistical method to separately estimate machine and fragment errors from overlapping read pairs. We modeled machine and fragment errors using the counts of discordant and concordant pairs of bases from overlapping read pairs, with read cycle and base quality score as predictors. We used a Markov chain Monte Carlo (MCMC) algorithm to estimate the posterior distribution of machine errors from discordant pairs of overlapping reads. We then applied the posterior mean estimates of machine errors to estimate fragment errors from concordant pairs of overlapping reads, given the reference genome and a list of known variants, with base quality score as a predictor.

We evaluated our MCMC algorithm using simulated error rates, then applied the algorithm to analyze chromosome 20 of 10 samples from a whole-genome sequencing study with average 8x coverage; each sample had over half of its reads overlapping. We showed that machine error was largely predicted by base quality scores of the

discordant overlapping read pairs, while being slightly increasing with read cycle. Fragment errors were also predicted by the base quality scores of concordant overlapping bases, as opposed to the previous assumption that fragment error rates are uniform across the genome [20]. Both machine and fragment errors were consistent across the 10 samples, regardless of the sequencing lane and plate of preparation, suggesting that these errors are consistent within the same sequencing experiment.

3.2 Methods

Here we present our statistical models for estimating machine and fragment errors using overlapping reads. Our approach focuses on substitution errors, which are the most common type of errors found in sequencing studies. Overlapping read pairs are defined as the paired-end reads with insert size smaller than the sum of read lengths. Assuming perfect alignment, at each position of the overlapping region of such read pairs, the overlapping bases b_1, b_2 have quality scores q_1, q_2, and come from read cycles c_1, c_2 respectively. The subscripts 1 and 2 per pair of bases denote the first and second read from the overlapping read pair (which can be in arbitrary order).

3.2.1 Model for estimating machine errors

If $b_1 \neq b_2$, the pair of overlapping bases is discordant. A discordant pair of overlapping bases must reflect at least one error from the sequencing machine, regardless of the reference base at this position and the underlying true base, because the two base calls are read from the same fragment. A discordant pair of overlapping bases occurs for one of the following three reasons: (1) b_1 is a machine error, (2) b_2 is a machine error, (3) both b_1 and b_2 are machine errors that result in different bases. Therefore, the probability (θ) of observing a discordant pair of bases among all overlapping pairs of bases can be expressed as a function of machine errors from b_1 and

b_2:

$$\theta = \theta_1(1-\theta_2) + \theta_2(1-\theta_1) + \frac{2}{3}\theta_1\theta_2, \tag{3.1}$$

where θ_1, θ_2 are the machine errors from a pair of overlapping bases. We assume that there is an equal probability of the machine error calling the underlying nucleotide as one of the other three bases; thus, two-thirds of the time two machine errors at the same base will result in a discordant pair, as reflected in the last term of Equation 3.1.

We model the machine error as a joint function of read cycle c and quality score q, therefore we can write $\theta_k = e_M(c_k, q_k)$, $k = 1, 2$. Given a dataset X of pairs of overlapping bases, we stratify them by the read cycles and quality scores of the two bases, represented by the tuples (c_1, q_1, c_2, q_2). Denote the total number of overlapping bases per stratum by $T(c_1, q_1, c_2, q_2)$, then each pair of overlapping bases is either concordant or discordant. Let $D(c_1, q_1, c_2, q_2)$ denote the pairs of discordant bases, then the log-likelihood of the data x can be expressed as a binomial function in terms of the discordance probability θ:

$$\begin{aligned} l(\mathbf{X} = (\mathbf{D}, \mathbf{T})|\theta) \propto & \sum_{(c_1,q_1,c_2,q_2)} D(c_1, q_1, c_2, q_2)\log(\theta) \\ & + [T(c_1, q_1, c_2, q_2) - D(c_1, q_1, c_2, q_2)]\log(1-\theta) \end{aligned} \tag{3.2}$$

Here we consider an additive model for machine errors with respect to read cycles and quality scores, where $e_M(c, q) = e_c(c) + e_q(q)$ for all c, q. In this construction we assume the effects of read cycle and quality score on machine error are independent. Let C denote the range of possible read cycles and Q denote the possible quality scores. Current sequencing technology typically generates 100bp reads with quality scores from 2 to 40, i.e. $C = \{1, 2, \ldots, 100\}$ and $Q = \{2, 3, \ldots, 40\}$. To estimate $e_M(c, q)$ for all $c \in C, q \in Q$, given the high-dimensionality $(|C| + |Q|)$ of the param-

eter space, and that the probability θ in the binomial likelihood is a function of two machine errors, we apply a Markov chain Monte Carlo (MCMC) approach to sample from the posterior distribution of the parameters.

3.2.2 Markov chain Monte Carlo algorithm for machine errors

The posterior distribution of θ is given by

$$p(\theta|\mathbf{X} = (\mathbf{D}, \mathbf{T})) \propto l(\mathbf{X} = (\mathbf{D}, \mathbf{T})|\theta)p(\theta). \tag{3.3}$$

Assuming an uninformative prior for the machine error rates, the likelihood is proportional to the desired (posterior) distribution of the machine error rates. Therefore, we apply a MCMC Metropolis-Hastings algorithm to explore the likelihood space. By definition, the Metropolis-Hastings algorithm utilizes a Markov process, which guarantees that the accepted sampled likelihoods will reach a unique stationary distribution that equals the posterior distribution of the machine errors. By proposing new parameter values in each iteration, the algorithm generates posterior probabilities of all parameters upon converging to the stationary distribution.

In the Metropolis-Hastings algorithm, we first provide initial conditions $e_c^0(c)$, $e_q^0(q)$ for all c, q. Then, to propose an update, we randomly select a pair of c_u, q_u from the ranges of possible cycles and scores and propose $e_c^*(c_u)$, $e_q^*(q_u)$ by drawing from a proposal function $g(e^*|e)$. We design $g(e^*|e)$ so that it allows mostly small-range jumps and occasionally long-range jumps, because we expect the error rates to be small, and small-range jumps can generate more precise estimates. Specifically, we write $g(e^*|e)$ as follows:

$$g(e^*|e) \sim 0.7 \times Unif(|e - j_1|, e + j_1) + 0.3 \times Unif(|e - j_2|, e + j_2), \tag{3.4}$$

where $e = e_c(c_u), e_q(q_u)$ and $j_1 < j_2$. Note that we impose absolute values at the lower bounds of the uniform distributions to ensure that the proposed error probabilities

stay greater than zero. We can then express the Metropolis-Hastings acceptance probability as

$$A(e \to e^*) = min\left(1, \frac{l(\mathbf{X}|e^*)g(e|e^*)}{l(\mathbf{X}|e)g(e^*|e)}\right). \tag{3.5}$$

We assess the convergence of the chains using the potential scale reduction factor (PSRF) of each parameter [25], which estimates the variance of each parameter as the weighted sum of the within-chain and between-chain variance, normalized by the between-chain variance. When multiple chains of the same sample converge to the maximum likelihood, the PSRFs for all parameters will be close to 1. We run the algorithm until the likelihood converges and the posterior estimates of the machine errors are stable, i.e. have reached their stationary distributions.

3.2.3 Model for estimating fragment errors

If $b_1 = b_2$, the pair of overlapping bases is concordant. If the concordant pair of bases are different from the reference base, assuming perfect mapping, one of the four cases has to hold:

1. The pair of bases reflect a fragment error (with no machine errors);
2. Both bases are machine errors that result in the same base (with no fragment error);
3. Both bases are machine errors that result in the same base, with underlying fragment error;
4. The pair of bases reflect a true variant.

We model fragment errors based on these scenarios and the machine error estimates from the MCMC algorithm described above. We assume each fragment error is predicted by the quality scores of the two concordant bases and is independent of

read cycles. Therefore, let $S(q_1, q_2)$ denote the number of pairs of concordant bases with quality scores q_1, q_2 respectively, that are different from the reference allele and the position is not a known variant. Let $N(q_1, q_2)$ denote the total number of pairs of overlapping reads with quality q_1, q_2. Then, fragment error $e_F(q_1, q_2)$ is evaluated by the following equation:

$$\begin{aligned}\frac{S(q_1,q_2)}{N(q_1,q_2)} =& e_F(q_1,q_2)[1-\hat{e_M}(\cdot,q_1)][1-\hat{e_M}(\cdot,q_2)] \\ &+ \frac{1}{3}\hat{e_M}(\cdot,q_1)\hat{e_M}(\cdot,q_2) \\ &+ \frac{1}{3}e_F(q_1,q_2)\hat{e_M}(\cdot,q_1)\hat{e_M}(\cdot,q_2) \\ &+ Pr(TV).\end{aligned} \tag{3.6}$$

Here $\hat{e_M}(\cdot, q)$ denotes the posterior mean of the machine error at q, averaged across the read cycles. $Pr(TV)$ denote the probability of the position being a true variant, which can be estimated from existing datasets and known variant databases. Therefore, fragment error is the only unknown in the equation above and hence can be evaluated analytically.

3.2.4 Simulation

To assess the properties of the MCMC for estimating machine error, we simulated sets of overlapping base pairs. We specified machine error for each combination of base quality scores and read cycles, $e_M(c, q)$. Then, given the total number of overlapping read pairs $T(c_1, q_1, c_2, q_2)$ per cycle and quality combinations, obtained from empirical data, the probability of observing $D(c_1, q_1, c_2, q_2)$ discordant pairs of overlapping bases is given by $p(c_1, q_1, c_2, q_2)$, where

$$\begin{aligned}p(c_1,q_1,c_2,q_2) =& e_M(c_1,q_1)(1-e_M(c_2,q_2)) \\ &+ e_M(c_2,q_2)(1-e_M(c_1,q_1)) \\ &+ \frac{2}{3}e_M(c_1,q_1)e_M(c_2,q_2).\end{aligned} \tag{3.7}$$

Then the simulation of discordant pairs of bases followed from a binomial draw, where

$$D(c1, q1, c2, q2) \sim \text{Binomial}(T(c1, q1, c2, q2), p(c1, q1, c2)). \tag{3.8}$$

We simulated discordant pairs of overlapping bases from the empirical total number of overlapping bases from chromosome 20 from one individual in the BRIDGES Consortium, a whole-genome sequencing dataset with average coverage 8x. The sequenced genome had 2.3 million reads (61.2%) overlapping on chromosome 20. Using the total number of overlapping bases stratified by read cycle and quality score from this individual, we generated discordant pairs based on two different sets of underlying machine error rates. Figure 3.1 shows the two sets of simulated e_c and e_q expressed in PHRED (logarithmic) scale [18]. The first set of simulated error rates came from the intermediate parameter estimates of the real data of the same individual sample, where the e_c estimates were roughly uniform across all $c \in C$, and the e_q estimates were inversely proportional to $q \in Q$. We generated a second set of error rates based on a smoothed form of theoretical expectations, with the read cycle and quality score components parameterized as follows:

$$\begin{aligned} e_c(c) =& 0.0007 - 0.0005c + 10^{-5}c^2 \\ e_q(q) =& 10^{-q/10} \end{aligned} \tag{3.9}$$

For each underlying error rate, we simulated three replicate datasets. For each resulting dataset, we performed our MCMC to estimate each error parameter, starting from six different sets of initial conditions, to assess the chain behavior starting from various locations of the likelihood space. We examined the convergence of likelihoods and parameter trace plots to determine burn-in and thinning parameters for the real sequencing datasets.

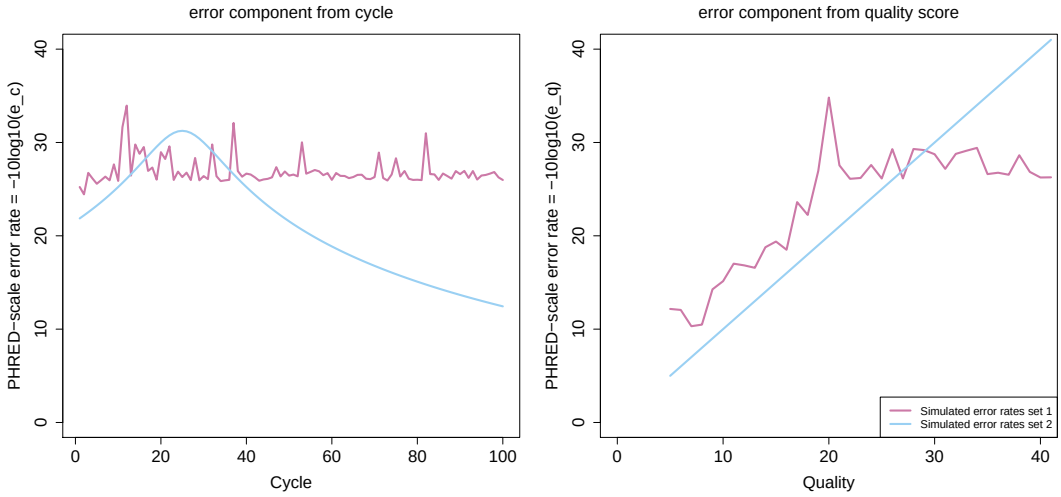


Figure 3.1: Two sets of machine error rates used in simulations, represented in PHRED scale. In our model, each machine error rate is the sum of the error probability contributed by read cycle (left) and the error probability contributed by quality score (right).

3.2.5 Sequencing data analysis

We estimated machine and fragment error from chromosome 20 of 10 individuals from the BRIDGES Consortium. We selected these 10 samples such that they all came from different sequence lanes (Table 1). These individual samples came from three plates of the sequencing experiment that had high proportions of overlapping reads on chromosome 20, ranging from 41.25% to 66.53%, with a median of 61.1%, which corresponds to an average of over 1.8 million pairs of overlapping reads per sample (Table 3.1). The average length of the overlapping region was 44 bp out of the 100 bp reads; therefore, the number of overlapping pairs of bases on chromosome 20 per sample was about 75 million. The base quality scores of these samples ranged from 2 to 41; as quality control, we discarded all overlapping bases with quality scores < 5 in our analyses. We performed MCMC for estimating machine errors on the overlapping reads data from chromosome 20 of each of the 10 samples, starting at 10 different randomized initial conditions. Each independent chain ran for 1,000,000

Sample ID	Sequence Plate	Sequence Lane	Percentage of Overlapping Reads on Chr 20	Number of Overlapping Reads on Chr 20
001	1	1	61.19%	2,328,750
013	1	2	57.25%	1,545,153
026	1	3	41.25%	1,621,221
038	1	4	54.64%	1,768,333
097	2	1	66.53%	2,100,776
113	2	2	64.63%	1,827,837
124	2	3	64.86%	1,853,273
181	3	1	61.05%	1,602,989
193	3	2	61.41%	1,703,547
205	3	3	60.87%	1,779,878

Table 3.1: Samples analyzed by the proposed error rates model. The samples come from three sequencing plates, all sequenced at different lanes.

iterations. We applied a burn-in of 20,000 iterations and a thinning parameter of 200 when calculating the posterior distribution of error rates. We used PSRF to assess convergence among the independent chains for each sample.

We then obtain the posterior mean estimates of each parameter $\hat{e_c}(c)$, $\hat{e_q}(q), \forall c, q$ by averaging across the converged chains of the same sample; estimates of machine errors $\hat{e_M}(c, q)$ follow from the sum of the respective $\hat{e_c}(c) + \hat{e_q}(q)$. We marginalize the machine error estimates by cycle and quality to assess the effect of each predictor on the errors.

To estimate the fragment error $e_F(q_1, q_2)$ as a function of the quality scores described in Equation 3.2, we apply the marginal quality error estimates $\hat{e_M}(\cdot, q)$. To obtain the fraction of concordant overlapping pairs that are different from the reference allele, we consider only the positions that are not present in dbSNP release 144, which removes the pairs different from the reference as they are known variants. Conditional on the position not being a known variant, we estimate its probability of being a true novel variant using the proportion of non-dbSNP variants in the BRIDGES Consortium final data freeze call set (Adam Locke, personal communication). Finally, we assess the correlation and variance of machine and fragment errors

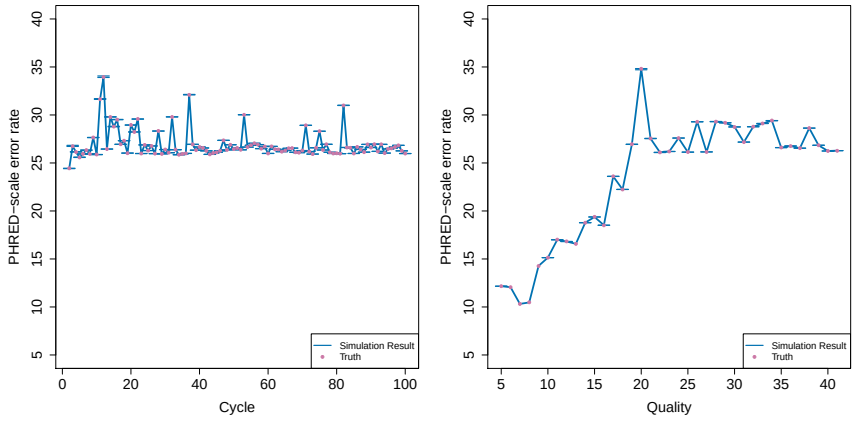
across samples.

3.3 Results

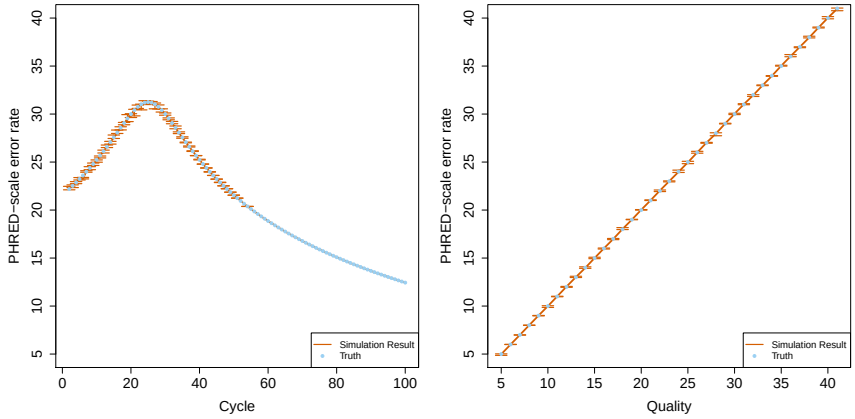
3.3.1 Simulation results

We simulated overlapping reads from two sets of error rates, generating three replicate simulated datasets from each set. For each replicate we start the chains at six different initial conditions; therefore, we have a total of 18 independent chains per set. For both sets of underlying error rates, each MCMC chain converged to the close neighborhood of its simulated true likelihood, regardless of the initial condition; the chain that started furthest from the truth took about 20,000 iterations to converge (Figure 3.7). We assessed the convergence of each parameter estimate using the proportional scale reduction factor (PSRF, [25]) which measures the ratio of within-chain variance to between-chain variance; therefore, PSRF ≈ 1 denotes convergence. Across all replicates and all initial conditions of the first set of error rates, the mean PSRF was 1.016, with variance 2.73×10^{-5}. Similarly, across all replicates and all initial conditions of the second set, the mean PSRF was 1.014, with variance 2.38×10^{-5}. Trace plots of representative parameters across a broad range of the parameter space showed good mixing of the chains around the true values, with a minor degree of autocorrelation between subsequent iterations, suggesting that thinning is needed for obtaining independent estimates (Figure 3.8).

Figure 3.2 shows the posterior mean estimates and 95% credible intervals for error sets 1 and 2. The posterior mean of each parameter and the 95% credible intervals are based on the parameter estimates across all 18 chains, with 4,000 observations per chain after burn-in and thinning. All estimates converged to the true values, with narrow credible intervals, on the order of 10^{-5}, while all parameter values were less than 10^{-4} ($<$40 in PHRED scale).



(a)

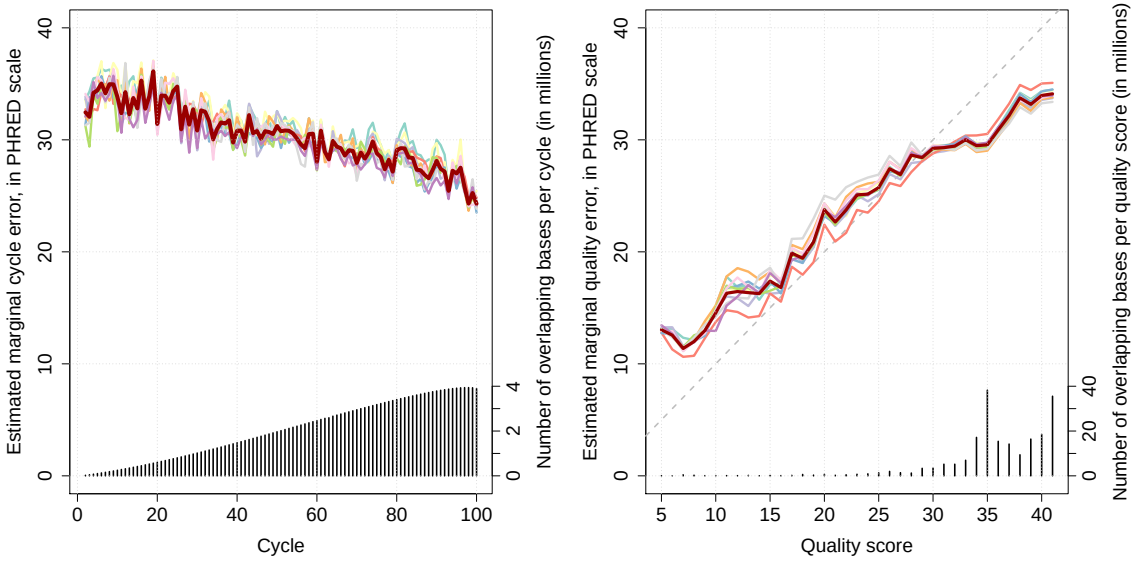


(b)

Figure 3.2: (a) Posterior mean error estimates of the cycle component, e_c (left) and quality component, e_q (right) of the machine errors simulated from error set 1. (b) Posterior mean error estimates of the cycle component, e_c (left) and quality component, e_q (right) of the machine errors simulated from error set 2. Vertical bars indicate 95% credible intervals of the parameter estimates.

3.3.2 Sequencing data results

We analyzed 10 samples with a median of 61% of overlapping reads on chromosome 20 (Table 3.1) using our MCMC algorithm for estimating machine errors. Considering the 10 independent chains per sample, the average PSRF across all machine error parameters was between 1.002 and 1.005, with variance on the order of 10^{-5} (Figure 3.9). This indicates that for each sample, all chains converged to the same posterior distribution for all parameters. We computed the posterior mean and credible intervals based on 4,900 observations per parameter, after burn-in of 20,000

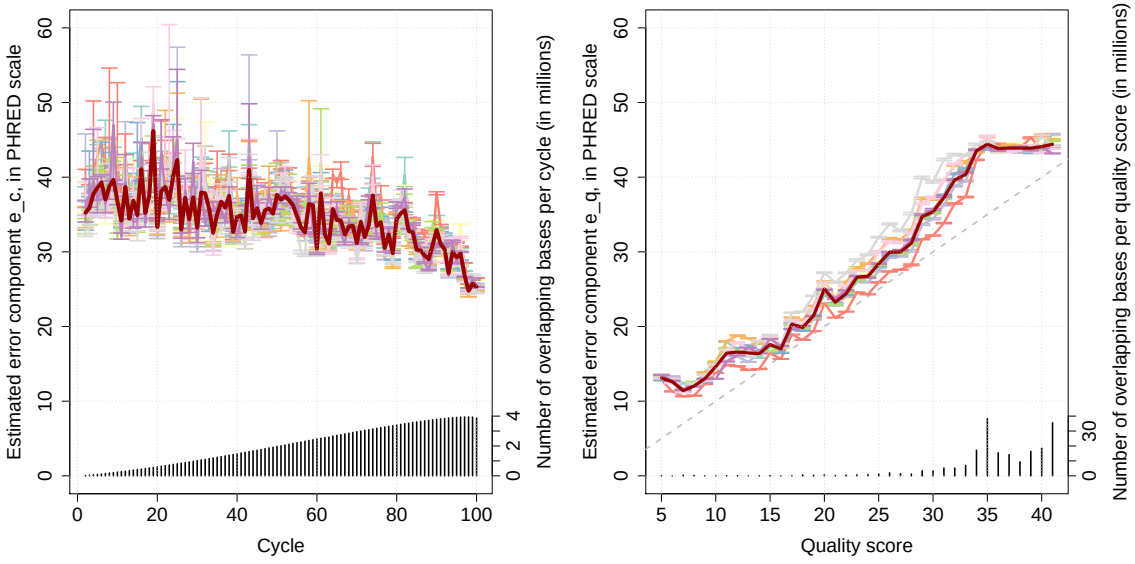


(a) (b)

Figure 3.3: Marginal distribution of estimated machine errors (a) by read cycle and (b) by quality score. Each color represents a sample, with the bold red line representing the mean across all samples. Histogram shows the distribution of overlapping bases in each covariate bin.

iterations and thinning of every 200 iterations.

We evaluated the marginal distributions $e_M(c, \cdot)$ and $e_M(\cdot, q)$ (Figure 3.3), for $c \in [1, 100]$ and $q \in [5, 41]$. Here we present the parameter estimates in PHRED scale, where PHRED $= -10 \log_{10}(e)$. Therefore, the higher the PHRED score, the lower the error rate. For each sample, the marginal machine error with respect to cycle showed minimal deviation from their mean marginal distribution. The marginal machine error was less than 0.001 (PHRED scale > 30) in the first 40 cycles (Figure 3.3a, Figure 3.10a). The marginal machine error increased with cycle, reaching an error rate close to 0.004 (PHRED scale 24) at the last cycles. When the reported quality score was less than 30, the estimated marginal error was lower than the reported error rate, as indicated by the quality score covariate. When the reported quality score was higher than 30, the estimated marginal error was higher than the reported error rate; the maximum PHRED-scale marginal error was 35, while the



(a) (b)

Figure 3.4: (a) Read cycle and (b) quality score components of machine errors estimated from the MCMC algorithm. Each color represents a sample, with corresponding vertical bars representing the 95% credible intervals of the error estimates. The mean across all samples is denoted by the bold red line. Histogram shows the distribution of overlapping bases in each covariate bin.

reported quality score was 41.

We assessed the contribution of quality score and cycle components of machine errors. Figure 3.4 shows the posterior mean estimates and 95% credible intervals for all parameters of machine error, namely the read cycle components $e_c(c)$ and the quality score components $e_q(q)$, for $c \in [1, 100]$ and $q \in [5, 41]$. We found that in all samples, e_c in PHRED scale showed a slight decrease with c, and the decrease was more significant in the last 10 cycles, indicating that the machine error contributed by read cycle is higher towards the end of a read. In the early cycles, the 95% credible intervals were wide for some samples with relatively few observations of overlapping bases (Figure 3.4a). Nonetheless, the 95% credible intervals of all samples overlapped. In the later cycles, the number of overlapping bases increased, reaching a maximum of 3.8 million at cycle 98 averaged across the 10 samples, thereby resulting in significantly narrower credible intervals.

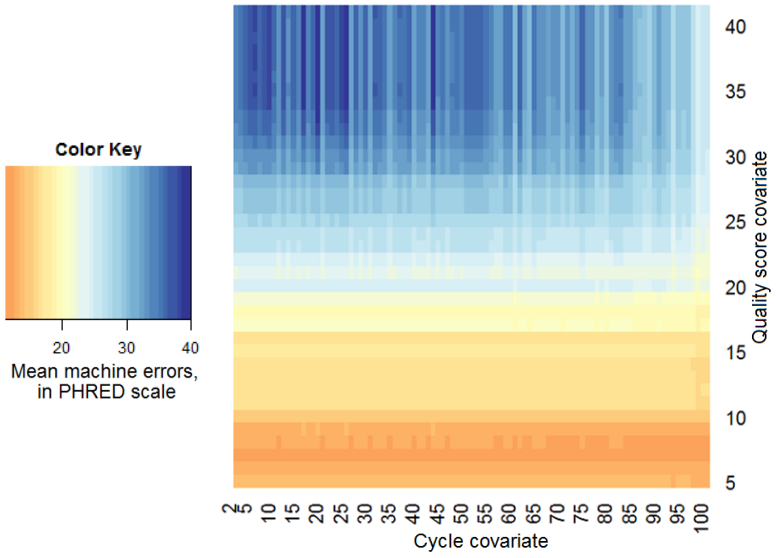


Figure 3.5: Machine error (e_M) estimates averaged across 10 samples.

The posterior mean estimates of e_q in PHRED scale increased with q, the reported quality scores of the overlapping bases (Figure 3.4b). The PHRED-scale e_q was consistently higher than the corresponding reported quality score. The 95% credible intervals for all quality score parameters were narrow because there were over 1 million observations per quality score bin.

Overall, the 10 samples showed highly consistent parameter estimates, with an average pairwise correlation of over 93% for the cycle parameters and 98% for the quality score parameters. We computed the average parameter estimates across 10 samples to obtain the average machine error $e_M(c, q)$ as an additive function of $e_c(c)$ and $e_q(q)$. We observed that quality score was the dominant predictor of machine error (Figure 3.5). The estimated machine error was the lowest (highest in PHRED scale) at the earliest cycles and when base quality scores were high (> 30).

Finally, we evaluated fragment error as a function of the reported quality scores of concordant overlapping bases different from the reference allele. Figure 3.6 shows the estimated fragment errors averaged across 10 samples. The variance of fragment

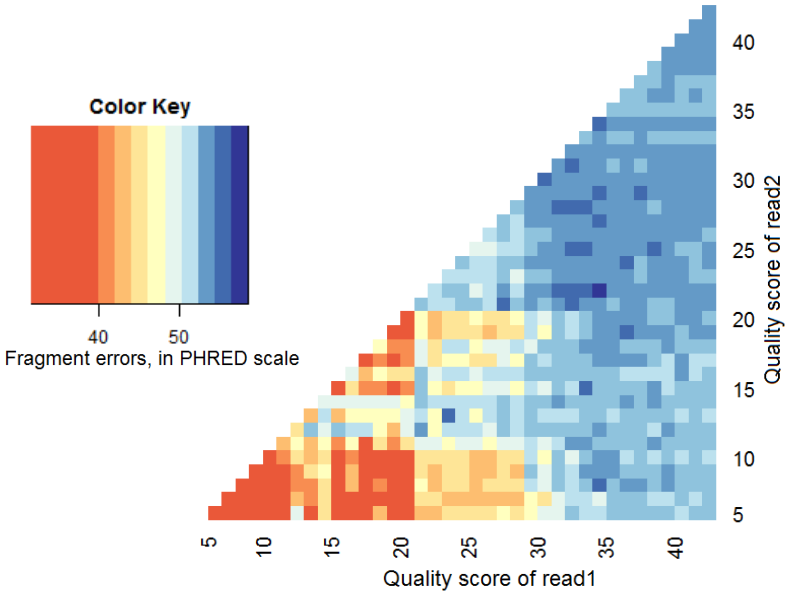


Figure 3.6: Fragment error (e_F) with respect to quality scores of pairs of overlapping bases, averaged across 10 samples.

errors across samples was in the order of 10^{-5}, suggesting that the fragment error estimates were stable across samples, regardless of their sequence lanes and plates. Low quality scores in both overlapping bases predicted high estimated fragment error (< 40 in PHRED scale). When one overlapping base had higher quality score (> 20), the estimated fragment error was low, mainly between 40 and 50 in PHRED scale. When both overlapping bases had high quality, fragment error was the lowest, at over 50 in PHRED scale, corresponding to an error rate of less than 10^{-5}.

3.4 Discussion

We designed error models for machine and fragment errors by utilizing the dependence property of overlapping reads pairs. Since an overlapping region sequenced the same part of a fragment, any discordance in a pair of overlapping bases would be due to machine error. Conversely, any concordant pair of overlapping bases different from the reference allele would likely be due to fragment error. To the best of our

knowledge, this is the first attempt to separately estimate fragment and machine errors from read data. We applied our models to a dataset with high proportion of overlapping reads; on chromosome 20 alone there were over seventy million overlapping bases, allowing enough observations for the high dimensional parameters in our error models.

We estimated that machine errors was largely predicted by the reported base quality scores. In particular, when the reported quality scores were less than 30, the estimated marginal machine error was lower than the reported error rate as indicated by the quality score. Conversely, when the reported quality scores > 30, the estimated marginal machine error was higher than the reported error indicated by the quality score. This could be explained by the fact that the reported quality scores were not recalibrated to reflect empirical mismatch rates. In fact, due to the fragment dependence, overlapping reads cannot be recalibrated the same way as normal reads; any mismatches to the reference at the overlapping regions should not be double counted. We will discuss in the next chapter a method to properly recalibrate overlapping reads. Nonetheless, these reported scores served as good predictors of machine errors although they did not directly represent machine errors.

Our results showed that fragment errors were predicted by the pair of quality scores in overlapping bases. While the majority of overlapping bases had quality scores > 20, which predicted fragment errors between 10^{-4} and 10^{-5}, the estimated fragment errors were at least one order of magnitude higher when both overlapping bases had low quality scores. Therefore, fragment errors spanned a wide range. While further experiments are required to assess the robustness of these findings to model specification, our results suggested the possibility that variant calling algorithms based on a fixed fragment error rate, typically 10^{-4} [20], might lead to inaccurate

estimation of genotype quality. In addition, we also observed that machine and fragment errors were consistent across samples in different sequence lanes and plates for the same experiment. It would therefore be useful to compare the machine and fragment error rates across sequencing experiments and platforms to assess the variability of these error rates.

Our error models relied on several assumptions that could potentially be relaxed or modified. First, we assumed the simplistic addictive contribution by read cycle and reported base quality score to machine error. While this model allows straightforward interpretation of the covariates contribution to machine error, the true interactions between cycle and quality score covariates is unknown. To potentially improve the fit of our model to the data, we propose modeling the interactions between covariates in different ways, for example by applying a multiplicative model for read cycle and quality score, or by incorporating higher order interactions in form of polynomials. Furthermore, previous studies suggested that errors in sequencing experiments could depend on sequencing context [2, 80]. Therefore, a logical extension of our error models would be to include sequencing context (dinucleotide or longer motifs) as a covariate. However, the number of parameters would increase when adding covariates and interactions. In this case we expect that overlapping reads data from the whole genome or from across samples would be needed to obtain stable estimates for all parameters.

Second, in our error estimation algorithm, we assumed independence of all parameters. Our results showed considerable correlation between neighboring estimates for both cycle and quality score parameters. Thus, we could potentially improve the estimation of error parameters with relatively few observations by taking into account the correlation of neighboring estimates. To improve the efficiency of the MCMC

algorithm, we could also incorporate the expected correlation between parameter estimates as a prior.

In summary, overlapping reads in a sequencing experiment have typically been discarded. In this chapter we have presented a utilization of overlapping read pairs for better understanding error sources in sequencing experiments. In the following chapter, we will continue addressing the problem of overlapping reads and provide a solution to resolve fragment dependence.

3.5 Appendix

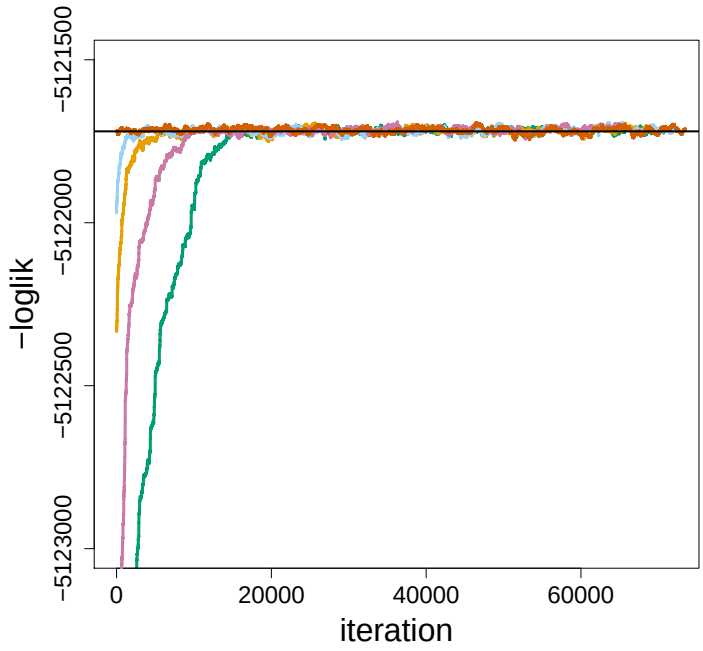


Figure 3.7: Likelihood convergence plot for simulated error rate set 1. The true likelihood is indicated by black horizontal line. Chains starting at different initial conditions converged to the true likelihood. The chain starting at the true likelihood (orange) stays around the true likelihood. We observe the same pattern for all replicate datasets for both sets of underlying error rates.

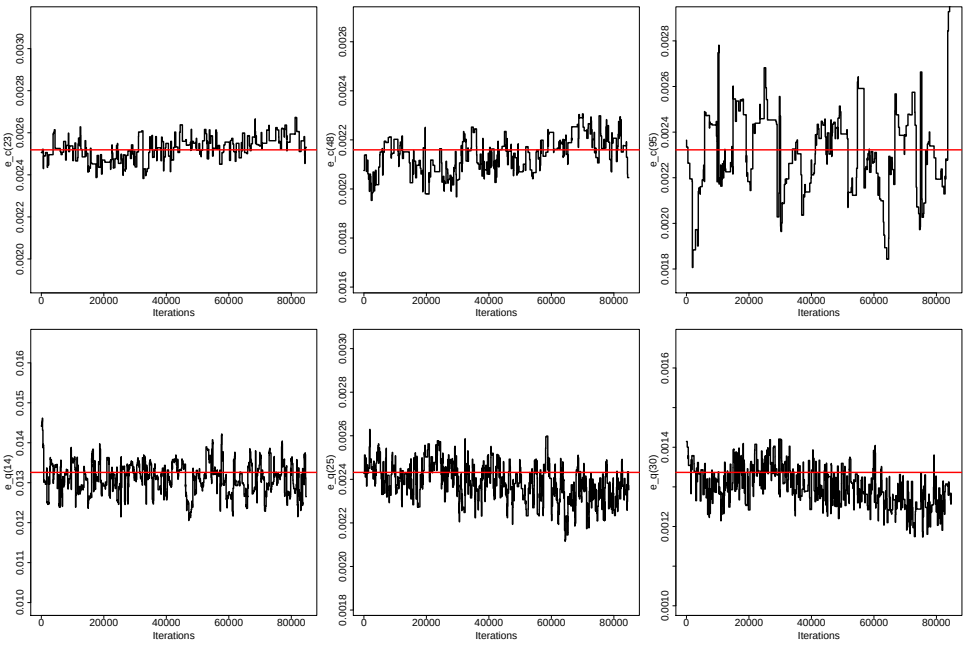


Figure 3.8: Trace plots of representative parameters over the range of parameter space. Red line on each plot indicates true error rate. Top panels show trace plots for the cycle parameters $e_c(23)$, $e_c(48)$ and $e_c(95)$. Bottom panel shows trace plots for quality score parameters $e_q(14)$, $e_q(25)$ and $e_q(30)$. The number of iterations on x-axis denote the iterations after 20,000 iterations of burn-in.

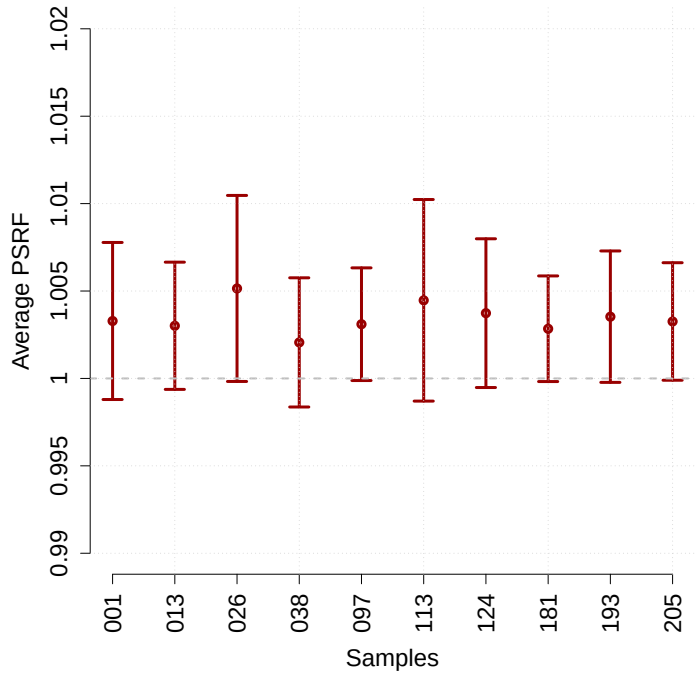
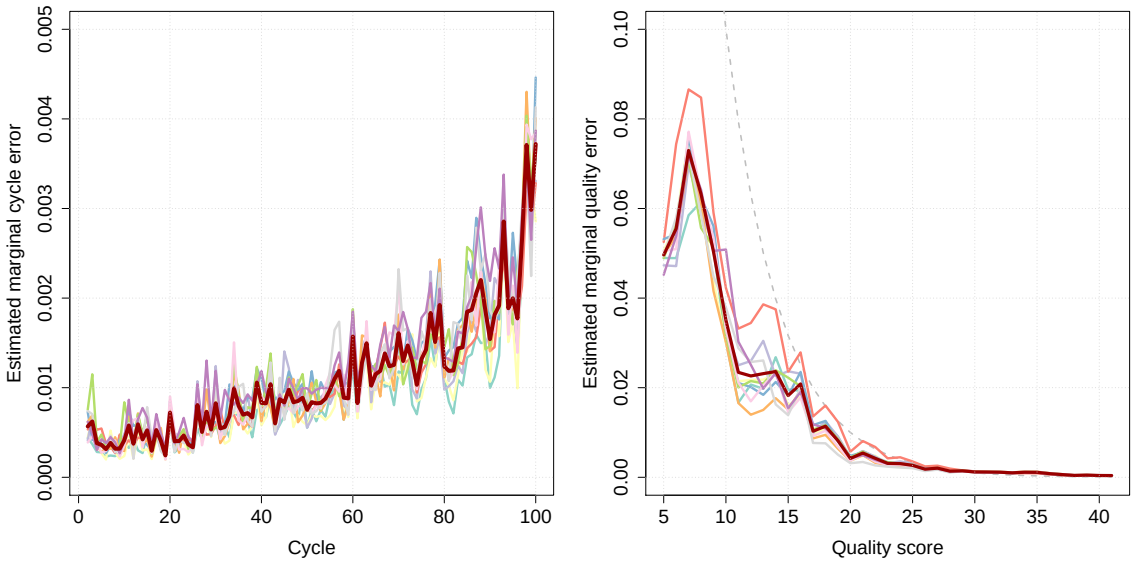


Figure 3.9: Potential scale reduction factor (PSRF) averaged across all parameters per sample. Vertical bars denote the standard deviation. All PSRF values were close to 1, suggesting that the 10 chains per sample all converged to the same parameter values.



(a) (b)

Figure 3.10: Marginal machine errors by (a) read cycle and (b) quality score. Each color represents one sample. Bold red line indicates mean marginal machine error across the 10 samples. Gray dotted line on (b) denotes the error rates as calculated from the reported quality scores.

CHAPTER IV

RESCORE: Resolving Overlapping Reads Dependence in Next-Generation Sequencing Data

4.1 Introduction

In Chapter 3 we introduced overlapping read pairs from short fragments in next-generation sequencing (NGS) experiments. When performing paired-end sequencing, each fragment is read by the sequencing machine from both ends [73]. An overlapping read pair occurs when the pair of reads sequences the same part of the fragment twice, resulting in the sum of read lengths longer than the insert size between the reads. At the overlapping region, any fragment errors, defined as the substitution errors on the DNA fragment, are replicated in both reads. As discussed in chapter 3, estimated fragment and machine errors are represented as base quality scores. These base quality scores and base calls generated by the sequencing machine are collectively used to determine the genotype at each genomic position sequenced. Genotype calling algorithms assume that reads covering a position come from independent fragments [76, 86], and then use cumulative base calls from reads and a base-calling error model [17, 53] to make genotype calls. Thus, falsely assuming independence of such error probabilities in genotype calling causes overestimation of genotype accuracy.

Overlapping reads are common in many studies, and may frequently occur for

specific experimental protocols. For example, in the 1000 Genomes Exome Phase 1 release, we observed an average of 28.8% overlapping reads across 185 European samples, with an average 28bp overlapping region from the 100bp reads [117]. In a recent study sequencing the whole genome of bipolar patients at an average coverage of 8x (BRIDGES Consortium), 806 samples contained an average of 40% overlapping reads (Figure 4.2). Sequencing studies of ancient DNA also tend to generate overlapping reads due to the short fragments of available DNA [45, 96]. In addition, new sequencing technologies are generating longer reads while insert sizes are about the same [63], which also leads to overlapping reads.

To avoid false positive calls from overlapping reads, existing solutions include "soft-clipping"(ignoring) one read at the overlapping region (clipOverlap) [26]. At the overlapping region, clipOverlap ignores the read segment with lower average base quality. This method is straightforward to implement, but the clipped read contains sequencing information independent of fragment origin, specifically replicates of the underlying estimates of machine errors, represented by base quality scores, at the clipped bases. Therefore, ignoring one read of an overlapping pair leads to loss of data and overcorrects the dependence problem. In the case of low-coverage sequencing of overlapping reads, with already limited information per genomic position, ignoring one overlap compromises the sequencing information from the clipped read and thereby undermines resulting genotype quality.

An alternative approach of correcting for overlapping reads is to use fragment-based likelihood calling implemented in the GATK variant calling pipelines [20]. Fragment-based calling calculates likelihood at each fragment instead of each read, by assuming a fixed fragment error rate and treating the base quality scores as machine error rates. While this method has been shown to reduce the number of

false positive singleton calls [20], we discussed in Chapter 3 that fragment error is not fixed across all bases. Fragment error can also be variable across different sequencing experiments. Moreover, this fragment aware method is not a stand-alone step and must be used as part of the GATK variant calling pipelines, making it less flexible when a different variant caller is desired for specific experiments.

In this chapter, we present our algorithm, RESCORE, for resolving the dependence in overlapping reads while retaining the information from the sequenced bases and machine errors at each overlap position. RESCORE identifies overlapping regions from aligned paired-end reads. Per overlapping position, RESCORE assigns a consensus base call, and a quality score reflecting the combined evidence for machine errors from both reads.

We applied RESCORE to 40 individuals from the BRIDGES Consortium with an average of 48.8% overlapping read pairs per individual. Overall, compared to soft-clipping one overlapping read at random or soft-clipping the read with lower average quality, RESCORE improved the recalibrated quality scores at the overlapping regions. RESCORE showed the most substantial gain over soft-clipping when both overlapping bases have high quality scores (> 25), because soft-clipping discards one high-quality base. The benefit of applying RESCORE was also reflected in the increased number of high-quality novel variants, with a ten-fold lower false positive rate, resulting in about 70 more novel variants on chromosome 20, which is a 0.027% increase in total number of variant calls over soft-clipping methods. In summary, RESCORE resolves the fragment error dependence and efficiently utilizes data at the overlapping reads by retaining sequencing information from both reads, which results in an accurate and high-quality call set.

4.2 Methods

4.2.1 The RESCORE algorithm

We consider a paired-end sequencing experiment with reads aligned to a reference and stored as a SAM/BAM file. Focusing only on properly paired reads, we identify overlapping read pairs in the data file using the physical positions of the alignments. In general, an overlap occurs when the sum of the read lengths in a pair is greater than the absolute value of the insert size. Since the alignment procedure introduces uncertainty and not all reads are matched exactly to the reference genome, we do not apply treatment to the overlapping region if the region is poorly aligned to the reference, as indicated by the non-match operations in the cigar string of the read. Since most fragment and machine errors are substitution errors, we do not apply treatment to reads with insertions or deletions in overlapping regions.

The overall idea of RESCORE is to assign a consensus base and new base quality score at each overlapping position, then recalibrate these scores to reflect the underlying error rates based on empirical mismatch. RESCORE consists of three steps for combining the base calls and base quality scores for each pair of overlapping bases.

The first step involves creating consensus base call and an intermediate score per pair of overlapping bases. Consider a pair of overlapping bases b_1 and b_2, with respective quality scores q_1 and q_2. Base quality scores are expressed in PHRED scale [18] which represents in logarithmic scale the estimated error at the sequenced base. We assign a combined base b_c, and an intermediate score s based on concordance of the base calls. If $b_1 = b_2$, $b_c = b_1 = b_2$ and $s = q_1 + q_2 + 100$ (For example, the first pair of overlapping bases C and C in Figure 4.1a). If $b_1 \neq b_2$, b_c is assigned to the base with higher quality score, and $s = |q_1 - q_2| + 200$ (For example, the third pair of overlapping bases G and T in Figure 4.1a). In the case where the overlapping

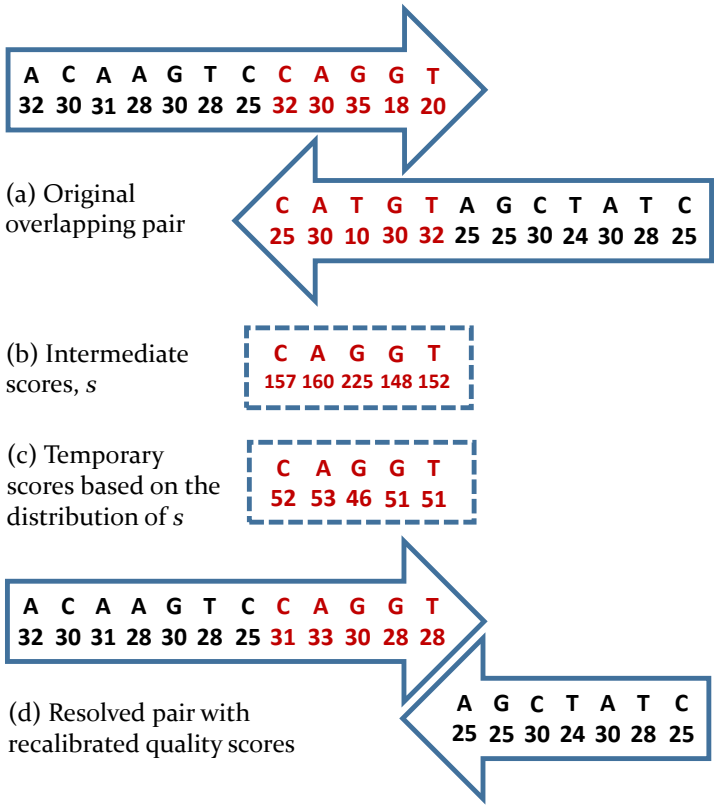


Figure 4.1: Cartoon description of the RESCORE algorithm: a) Original pair of overlapping reads. Red characters denote overlapping bases and respective base quality scores. b) The first step of RESCORE generates a consensus base and an intermediate score per pair of overlapping bases. c) The second step of RESCORE maps intermediate scores from previous step onto empty PHRED-scaled quality score bins. d) The third step of RESCORE concatenates the overlapping region to the read of higher average base quality, then applies base quality score recalibration such that the recalibrated scores reflect base calling error.

bases are discordant but the quality scores are equal, we assign b_c by flipping a fair coin. Note that these intermediate scores are not representative of the actual error probabilities at the bases, but are considered a summary of q_1, q_2. Adding constants to the summary of q_1, q_2 ensures that the resulting scores are outside of the range of quality scores generated by sequencers and generates disjoint intervals for concordant and discordant bases. In this construction, we assumed that two pairs of concordant bases with the same summed quality scores are the same, regardless of the original two scores. Typical NGS reads generated from Illumina machines have

base quality scores 0 to 40, hence s will ranges from 100 to 240. In practice, the exact mathematical operation on a pair of scores is flexible, as long as it results in disjoint ranges of intermediate scores for concordant and discordant bases.

The second step of RESCORE involves distributing the intermediate scores into a range of bins recognizable by base quality recalibration tools (Figure 4.1b). We identify the maximum base quality score (bq_{max}) in the data set, then assign the intermediate scores into score bins $> bq_{max}$ to avoid mixing the base qualities of overlapping and normal reads. To ensure optimal performance of recalibration, each bin must have a large number of observations and the number per bin is approximately equal. On the other hand, we need to avoid having a wide range of intermediate scores being grouped together, because this would obscure the evidence from the combined base quality scores from the overlaps. Therefore, we distribute the intermediate scores s into bins sequentially, such that:

1. Bases with the same s must be in the same bin;
2. The total number of consecutive s per bin is no more than 10;
3. If condition 2 is satisfied, each bin has a maximum of 100 million data points. If condition 2 is not satisfied, the bin contains all counts from 10 consecutive scores, and the following score starts a new bin.

Note that typically there are very few discordant bases; hence RESCORE assigns all corresponding intermediate scores, ranging from 200 to 239, to the last bin.

After assigning score bins to all overlapping bases, we reconstruct pairs of reads by concatenating the overlapping region with consensus bases and temporary quality scores to the read with higher average base quality at the overlapping region (Figure 4.1c). The overlapping region on the read with lower quality is soft-clipped, in order

to replicate the construction of reads in clipOverlap. All overlapping reads are then printed into a new BAM file together with the other reads originally in the file. This procedure allows subsequent recalibration to work on overlapping reads at the same time as normal reads, and a flexible choice of base quality score recalibrator.

The third step recalibrates the base quality scores such that they represent empirical mismatch probability (Figure 4.1c). Given each reported quality score, the recalibration algorithm counts the number of empirical mismatches in the bases with this score, adjusting for read group effect, sequencing context (dinucleotide), and the position of the base within a read (cycle). Mismatch is defined as difference compared to the reference genome, excluding known sites of variation. The number of mismatches over the total number of bases in a category represents the actual base calling error rate in that category [27, 76]. In the case of low count in a bin, recalibration will add in a pseudo count to adjust for the empirical mismatch rate, which may lead to slight overestimate of error rate in that bin but will not affect the recalibration of other bins.

4.2.2 Evaluating RESCORE

We applied RESCORE to 40 samples from the BRIDGES Consortium, a whole-genome sequencing data set with average coverage 8x. This study has 806 samples with over 40% overlapping reads; our 40 samples came from one plate, with an average 49.7% overlapping reads per sample (Figure 4.2). The yield per sample ranged from 92 to 203 million reads, with a mean of 155 million reads. The mean number of overlapping reads per sample was 77 million. The mean length of the overlapping region in each 100bp reads was 44bp. Therefore, on average each sample had 3.4 billion pairs of overlapping bases. We applied RESCORE on each chromosome separately with the same bins for all chromosome. We then performed recalibration collectively

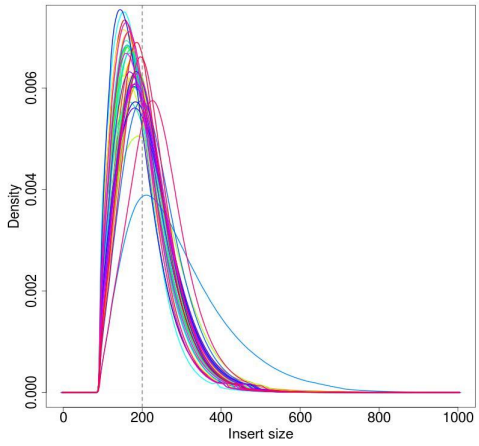


Figure 4.2: Insert size distribution of 40 samples from the BRIDGES Consortium. This sequencing experiment generated 100bp reads, hence all insert sizes < 200 (gray vertical dashed line) are overlapping.

on the whole genome using `dedup/recab` [27] that simultaneously removes duplicate reads and recalibrates scores.

We evaluated RESCORE by comparing it against the existing method of treating overlapping read pairs (clipOverlap) which applies soft-clipping to the read with lower average quality score in the overlapping segment [26]. To develop a baseline for our comparisons, we also applied to our samples the strategy of soft-clipping one of the overlapping reads at random. Following each of these three strategies, we performed the same recalibration step and then compared the recalibrated quality scores at the overlapping bases generated from these treatments. To assess the potential improvement of RESCORE over the other strategies, we compared the recalibrated scores from the resulting bases in overlapping pairs to the averages and to the differences between two original base quality scores. We also noted the difference in computation times among the strategies.

To determine the effect of resolving overlapping reads on variant calling, we next applied individual-based single-marker calling to the 40 samples, using reads with

overlapping regions (1) treated by RESCORE, (2) treated by clipOverlap, and (3) treated by soft-clipping one overlap at random. For all variant positions covered by overlapping reads, we compared the quantity and quality of novel variants identified in each set. Novel variants were defined as the ones not found in the most recent release of dbSNP 144. We curated the quality of the novel variants by applying a genotype quality filter at Q20, which was the PHRED-scale ratio of the posterior probabilities of the most likely and the second most likely genotypes. We then investigated how many of these passed variants were shared between datasets and how many were specific to one dataset. The quality of the filtered sets of novel variants was measured by transition-to-transversion ratio (Ts/Tv). We estimated false positive rates in these call sets using the observed and expected Ts/Tv. Assuming that random sites had Ts/Tv = 0.5, the false positive rate (e_{fp}) can be estimated by the equation

$$e_{fp} = \frac{\text{expected Ts/Tv} - \text{observed Ts/Tv}}{\text{expected Ts/Tv} - 0.5} \tag{4.1}$$

Finally, we assessed our RESCORE algorithm across various bin sizes applied in the pre-recalibration score allocation process (Step 2). Specifically, we applied RESCORE with maximum bin sizes between 1 million to 1 billion on one individual sample. The binning rules 1 to 3 stayed the same for all bin sizes. We compared the resulting the difference in recalibrated scores with respect to various bin sizes.

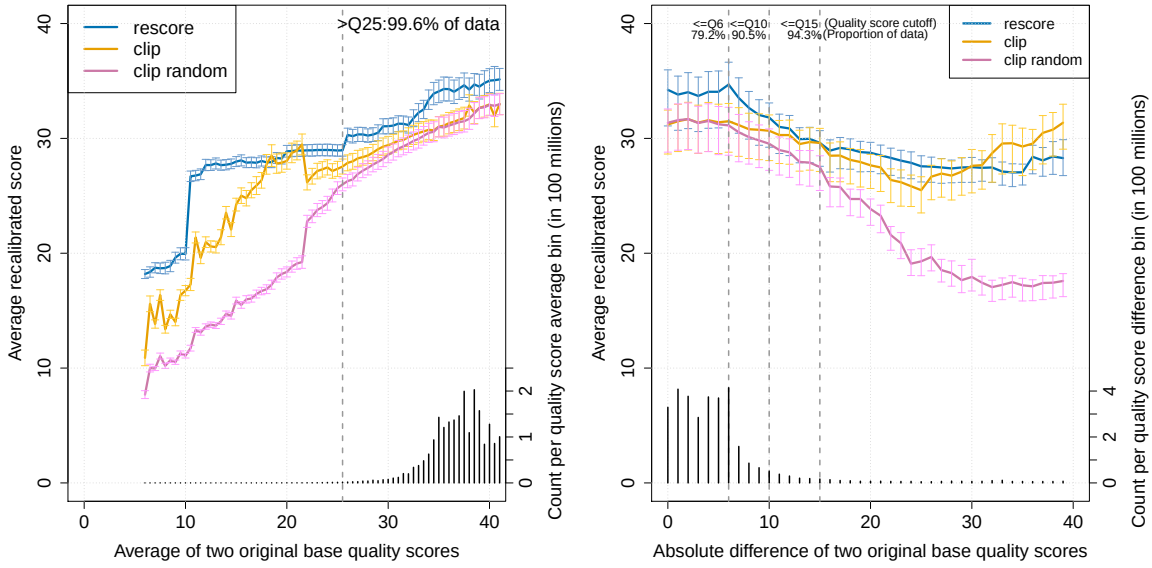
4.3 Results

4.3.1 Recalibrated scores comparison

Our proposed RESCORE algorithm utilizes combined base call and base quality information from an overlapping pair of bases, then assigns a consensus base and quality score that reflects the underlying mismatch rate. Figure 4.3 shows increased

recalibrated base quality scores in RESCORE over the soft-clipping methods, for the majority of the data. The maximum per-sample average recalibrated quality score attained by RESCORE was 37, while that attained by clipOverlap and clipping one read at random were both 34. Over 99% of the overlapping pairs of bases had average quality above 25, where RESCORE resulted in an overall increase of 2.46 in the recalibrated score compared to clipOverlap, and an overall increase of 2.65 compared to clipping one read at random (Figure 4.3a). The majority of overlapping bases had an average original quality over 35; here, clipOverlap performed no better than clipping one read at random, while RESCORE had significantly higher recalibrated scores of 2.60 (Figure 4.3a). In comparing the differences of scores, when original scores differed by less than 6, which was true for over 79% of the data, RESCORE had on average 2.63 higher recalibrated scores than clipOverlap and 2.69 higher than clipping at random (Figure 4.3b). Only in the rare case of a score difference of 30 or higher (2.3% of the data), meaning that one of the overlapping bases had a much higher original quality than the other, RESCORE was slightly more conservative than clipOverlap in the resulting recalibrated scores. Here RESCORE had an average of 1.80 lower recalibrated score than clipOverlap. However, RESCORE still showed a significantly higher recalibrated score of 10.26 than the null strategy of clipping at random.

Since most pairs of overlapping bases had small differences in the quality scores, we focused on the pairs of overlapping bases with the same original base quality scores. In this category, clipOverlap and clipping at random showed an almost identical trend of recalibrated scores, with respect to all original scores. RESCORE showed substantial improvement in recalibrated quality scores over both soft-clipping methods, particularly when the original scores were moderate (between 10 and 20),



(a) (b)

Figure 4.3: Recalibrated quality scores from overlapping read pairs processed by RESCORE (blue), soft-clipping the lower quality read (clipOverlap, yellow) and soft-clipping one read at random (pink), plotted against: (a) average original base quality scores of the overlapping bases, and (b) difference between original base quality scores of the overlapping bases. Vertical bars on each colored line denote the range of the recalibrated scores across 40 samples. The histogram in the plots represent the number of overlapping pairs of bases in each original quality score bin. Vertical dashed lines (gray) denote the partitions of the majority of data, with the respective cutoff and proportions of data labeled next to the dashed lines.

where RESCORE had on average 6.90 higher recalibrated scores than the soft-clipping methods (Figure 4.4). The maximum difference in recalibrated score between RESCORE and the soft-clipping methods was 10.57, when the original quality was 16.

4.3.2 Effect of bin size in RESCORE

We studied the effect of using different maximum bin sizes in step 2 of the RESCORE algorithm, where intermediate scores were grouped into PHRED-scale score bins (that were otherwise empty) for recalibration. When bin size was between 1 million and 10 million, at average original score < 25, the recalibrated scores from RESCORE were lower than those from clipOverlap (Figure 4.5). When bin size was

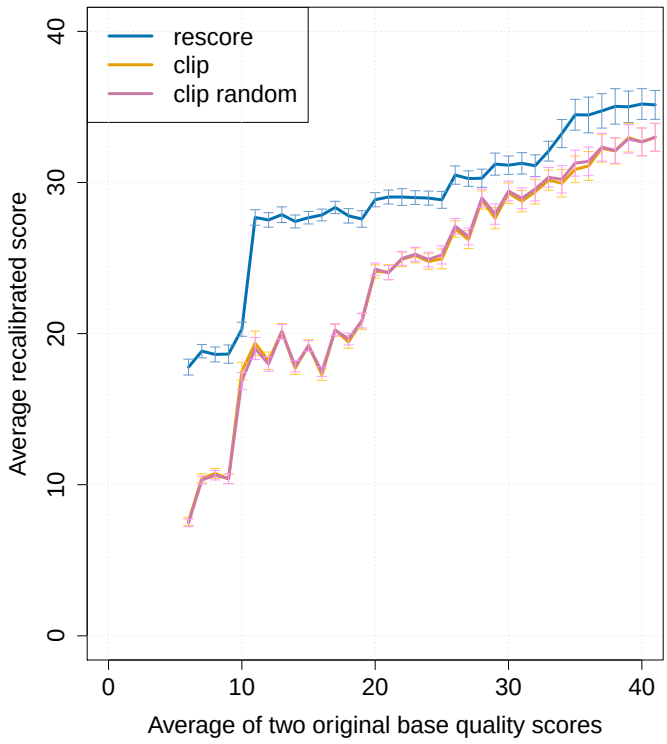


Figure 4.4: Recalibrated quality scores from overlapping read pairs processed by RESCORE (blue), soft-clipping the lower quality read (clipOverlap, yellow) and soft-clipping one read at random (pink), plotted against original base quality scores of the overlapping bases, when the two original scores were equal.

between 25 million and 100 million, the step-like recalibration scores in the original score range of $10 - 25$ indicated that most of those scores with relatively few counts were recalibrated together due to the increase in bin size. At average original quality > 25, which was the majority of the data, some original scores contained as many as 200 million observations (Figure 4.3a). Therefore, when bin size ranged from 1 million to 100 million, each original average score beyond 32 formed its own bin; as a result, the recalibrated scores were the same for these bin sizes. When bin sizes further increased to 200 million, 500 million and 1 billion, multiple original average scores grouped into a bin even at the highest original scores. The resulting recalibrated quality leveled off when original scores were greater than 35; these recalibrated scores from large bin sizes were lower than the ones from smaller bin sizes.

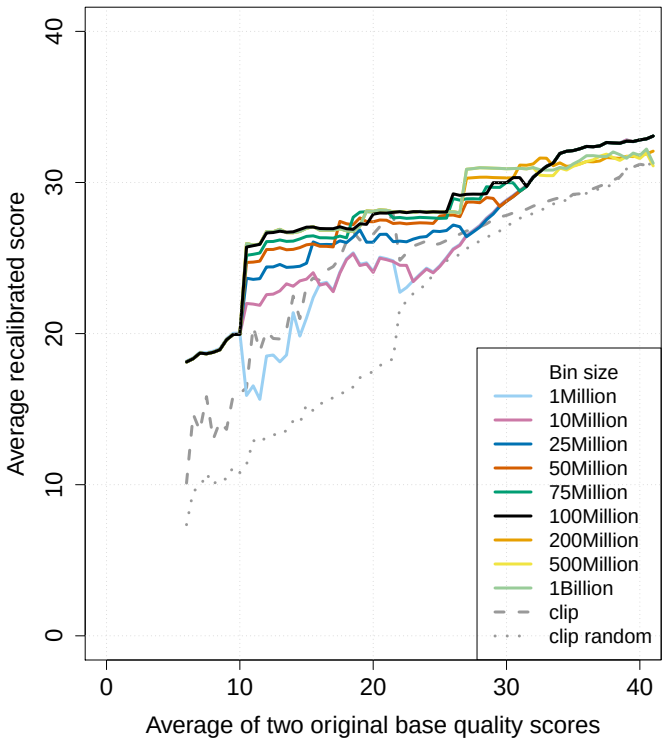


Figure 4.5: Recalibrated quality scores from overlapping read pairs of one sample processed by RESCORE plotted against original base quality scores of the overlapping bases. The different solid color lines represent a range of RESCORE bin sizes used in step 2 of the algorithm. Black line denotes the bin size of 100 million that resulted in the highest overall recalibrated scores. Gray lines represent the scores of the same sample processed by clipOverlap (dashed) and clipping and random (dotted).

Overall, the bin size of 100 million showed the highest resulting recalibrated scores.

4.3.3 Variant calling comparison

We evaluated chromosome 20 variant calls of all 40 samples generated by individual-based single-marker caller. A total of 247,666 variants (union set) on chromosome 20 were identified by the individual-based single marker caller performed on the dataset processed by RESCORE, clipOverlap and clipping at random. The majority of variants (219,591, i.e. 88.7%) was spanned by at least one overlapping read pair. Of these variants, 205,821 (93.7%) were found in the most recent database of known variants (dbSNP v144), and over 99% of these variants were concordant among the three datasets. The remaining 13,770 (6.3%) variants were not found in dbSNP and

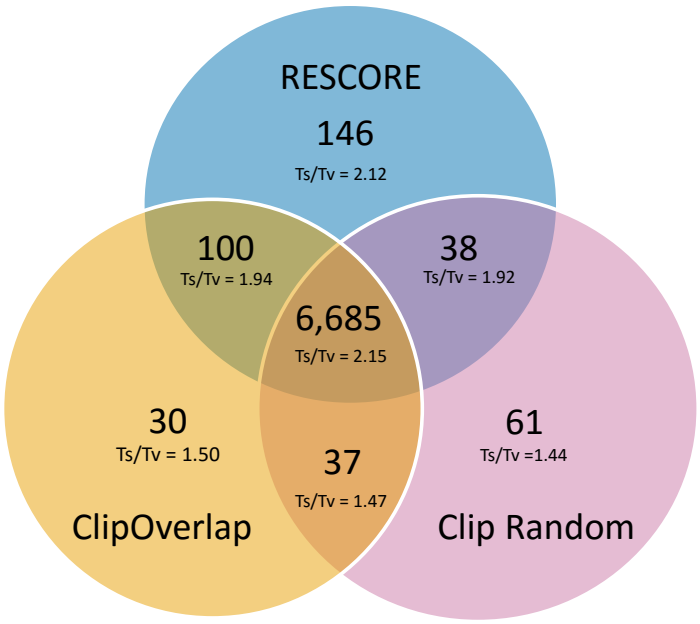


Figure 4.6: Venn diagram showing the number and Ts/Tv of >Q20 novel variants from each dataset.

hence were considered novel. After applying the genotype quality filter of Q20, 7,097 (51.5%) novel variants remained.

Of these >Q20 variants, 6,685 were shared by the three datasets generated by RESCORE, clipOverlap and clipping at random, with a transition-to-transversion ratio (Ts/Tv) of 2.15 (Figure 4.6). The remaining variants were either shared by two datasets or were specific to one dataset. RESCORE identified 146 variants not found in the other datasets, with a Ts/Tv of 2.12. RESCORE and clipOverlap shared 100 variants, with a Ts/Tv of 1.94; RESCORE and clipping at random shared 38 variants, with a Ts/Tv of 1.92. In total, RESCORE generated 6,969 >Q20 novel variants, the largest number among the three datasets.

In contrast, variants specific to clipOverlap and clipping at random had Ts/Tv lower than 2. clipOverlap generated 6,852 total variants, 30 of which were specific

to this set with a Ts/Tv of 1.50. Clipping at random generated 6,821 total variants, 61 of which were specific to this set with a Ts/Tv of 1.44. clipOverlap and clipping at random shared 37 variants that were not in the RESCORE set, with a Ts/Tv of 1.47 (Figure 4.6). Note that all 128 variants that were specific to the soft-clipping methods were identified by the RESCORE full call set, but were subsequently filtered out due to low quality.

Assuming that the novel variants shared among all three datasets were true, their Ts/Tv = 2.15 served as the expected value for calculating the false positive rates. Among the novel variants specific to RESCORE, which had a Ts/Tv of 2.12, we estimated the false positive rate to be 2.6%, meaning that 4 out of the 146 RESCORE-specific variants were likely false positives. Among the variants specific to clipOverlap and/or clipping at random, the combined Ts/Tv was 1.46. The false positive rate was 41.7%, meaning that 53 out of the 128 soft-clipping-specific variants were likely false positives.

4.4 Discussion

Overlapping reads from short genomic fragments is an on-going issue in many next-generation sequencing experiments. Many studies treat overlapping reads as normal read pairs, which leads to overestimation of genotype accuracy. On the other hand, soft-clipping one of the overlapping reads neglects base calling information independently available from the two reads. Our RESCORE algorithm solves both problems by keeping only one read from the overlapping regions while retaining the combined information from both reads in the base quality scores.

Overall, RESCORE generated higher base quality scores than soft-clipping methods, particularly when the two overlapping bases had comparable moderate to high

base quality scores. This can be explained by the fact that soft-clipping methods discarded one useful piece of sequencing information, while RESCORE combined the two scores into a higher recalibrated score than the recalibrated score from one read in the soft-clipping methods. In our dataset, the majority of overlapping base pairs fell into this category. In fact, we envision that many current sequencing studies will have overlapping base pairs where both overlapping bases have comparable quality, because the improved sequencing technologies are able to generate longer reads with stable base quality along the read. However, the fragment lengths are not proportionately longer, making overlapping reads an increasing problem. Therefore, our RESCORE algorithm will be effective in these new datasets.

In the rare scenario when the pair of overlapping bases consisted of one high quality base and one low quality base, RESCORE was slightly conservative when compared to clipOverlap [26]. This occurred because RESCORE incorporated information from the low-quality base into the consensus base and base quality, leading to somewhat lower recalibrated scores. Nonetheless, RESCORE showed substantial improvement over the baseline strategy of soft-clipping one overlap at random at all combinations of quality scores between the two overlapping bases.

We also studied how varying the bin size in the RESCORE mapping step affected the resulting recalibration quality. Base quality score recalibration was a crucial step in RESCORE. To allow recalibration to work effectively, RESCORE mapped the intermediate scores containing combined information from the overlapping pairs of bases to a set of bins at the PHRED-score range. The number of observations present in each bin for recalibration must be large, because recalibration further stratifies the observations by read cycle, sequencing context, and possibly other covariates. While more observations per bin gives more stable base error estimates, we

observed that for the bases with high original quality scores, the recalibrated quality was lower at large bin sizes than at small bin sizes. This can be explained by the fact that we artificially grouped together observations that represented a wide range of underlying mismatch rates, thereby lowering the recalibrated quality of the otherwise highest quality bases by incorporating the lower quality bases in the mismatch calculation. Our analyses were performed using the bin size of 100 million that resulted in the highest recalibrated scores among all the bin sizes we tested. While the empirical distributions of quality scores differ between datasets, with improved sequencing technology, we expect recent sequencing datasets have most overlapping bases falling in the higher quality score range, similar to the dataset we studied. Further experiments could be performed to assess if there is an optimal bin size and binning rules that maximize the improvement of RESCORE over soft-clipping strategies.

The increment of base quality scores in overlapping reads processed by RESCORE was reflected by the discovery of more high quality novel variants. While the majority of novel variants were concordant between the RESCORE and soft-clipping datasets, the variants that were only identified in the soft-clipping datasets had an estimated false positive rate of over 41%, while the variants that were RESCORE-specific had an estimated false positive rate of 2.6%. This suggested that RESCORE was able to identify 0.027% more variants on chromosome 20 that were likely to be true positives. This is equivalent to expanding the novel variant set on chromosome 20 by 1.00%.

RESCORE incurs a higher computational burden than soft-clipping methods, because RESCORE needs to store intermediate scores from overlapping reads, and parse the overlapping reads once more to assign temporary quality scores for recalibration. The run time of RESCORE scales linearly with the number of overlapping

reads in the dataset. In our dataset, RESCORE took on average 6 to 45 minutes to analyze the shortest chromosome to the longest chromosome, while clipOverlap took 4 to 24 minutes. Clipping at random used about the same amount of time as clipOverlap, because the comparison of read qualities in the overlapping pair requires negligible time. The additional computational burden could be alleviated if RESCORE is integrated with the recalibration algorithm, where we minimize the time spent on reading the input files by directly tabulating the intermediate scores from RESCORE into the recalibration table, thereby achieving higher efficiency and improving the usability of RESCORE.

We were not able to compare RESCORE directly to the fragment-aware calling method [20], because the fragment-aware method must be used as part of the GATK variant calling pipeline, which included base quality score recalibration and variant calling. Therefore, if we compared our existing results against results from the GATK pipeline, all algorithms leading to the results were different in the two sets, which would make us unable to isolate the difference between RESCORE and the fragment-aware step. In contrast to this limitation of fragment-aware calling, RESCORE can be flexibly incorporated into any sequencing processing pipeline since it utilizes standard BAM format as input and output, and the choice of subsequent recalibration software is also flexible. In fact, treating fragment as a unit for likelihoods is conceptually similar to RESCORE in adjusting for quality scores per fragment; thus, we expect comparable results from fragment-aware calling and RESCORE.

The performance of RESCORE may be limited by the compression of base quality scores in some recent NGS datasets. In order to reduce the burden of computer storage and data transfer, methods have been proposed to compress base quality scores [35, 46, 122, 134], because the full-range PHRED scale quality scores from 0

to 40 require almost three times as much storage space as base calls. Compression reduces the number of possible quality scores in the dataset, where each compressed score only represents a crude probability of machine error at the base. It requires further investigation how RESCORE and recalibration could be applied to these compressed scores without recreating the full-range scores and hence imposing the storage burden again.

While the proposed compression methods seem to have preserved the sensitivity and specificity of genotype calling [35, 134], in the published assessments of these data compression methods, none investigated the impact of score compression on downstream analyses that utilize genotype quality that are based on quality scores. We expect that score compression will lead to imprecise estimates of genotype quality because the compressed scores only represent a coarse probability of base error. Since scores are in logarithmic scale, a 5-point score difference corresponds to more than 3-fold difference in the underlying error probability. Therefore, regardless of the details of the compression algorithm, having a coarse range of base qualities means encapsulating a wide range of underlying base error rates into one single rate, which intuitively would undermine the precision of the resulting genotype qualities. Therefore, before further evaluation of the impact of data compression on downstream analyses, we still recommend keeping the full-range quality scores and apply RESCORE and recalibration to datasets.

In summary, RESCORE resolves the dependence in overlapping read pairs by combining base calling information from two reads into a consensus base and a summarized quality score, which reflects empirical mismatch after recalibration. Compared to soft-clipping methods for treating overlapping reads, RESCORE generates higher recalibrated quality scores, and as a result higher quality novel variants. We

recommend applying RESCORE as a standard read processing step, particularly low-coverage sequencing experiments with large number of overlapping reads.

CHAPTER V

Whole-genome sequencing of uropathogenic *Escherichia coli* reveals long evolutionary history of diversity and virulence

5.1 Introduction

Extraintestinal *Escherichia coli* (*E. coli*) are capable of causing various infections, including urinary tract infection (UTI) and meningitis [22]. Approximately 90% of all UTI cases are caused by *E. coli* capable of colonizing the urinary tract, collectively known as uropathogenic *E. coli* (UPEC) [138]. From an evolutionary perspective, UPEC together with other extraintestinal pathogenic *E. coli* (ExPEC) belong to the *E. coli* phylogroups B2 and D, characterizing their specific adaptations to colonize and cause infections outside of the gut [10]. Since the urinary tract presents a significantly different environment than the gut, UPEC carry virulence factors very different from diarrheagenic *E. coli* [42]. For example, UPEC possess adhesins to attach to epithelial cells of the urinary tract to overcome the frequent flow of fluids [89] and specific toxins for invading and replicating in the urinary tract [79]. These known uropathogenic virulence factors presumably have multiple functions, as there is no direct correlation between these factors and UTI symptoms [74]. UPEC display a high diversity of genotypes and phenotypes [47, 137], suggesting that UPEC have multiple origins [23, 128].

This chapter is published as Lo, Y. et al. 2015. *Infection, Genetics and Evolution*, 34, 244-250.

However, previous insights into the origins and spread of uropathogenecity were limited by their focuses on small regions of the bacterial genome that are well-conserved, such as genes used in multilocus sequence typing (MLST)[29, 74]. These regions provide limited insight in the evolution of pathogenicity as they do not contain any of the virulence factors. Marrs et al. [74] classified UPEC by grouping them into pathotypes based on virulence factors, analogous to the pathotypes for diarrheagenic *E. coli* [81]. However, they did not find direct correlation between pathotype and clinical presentation. Other attempts of grouping UPEC by virulence factors also failed to identify a correlation between virulence factors and UTI symptoms [111, 135]. These classification attempts suggest that UPEC virulence and genetic diversity cannot be captured by studying only a restricted set of genomic regions.

To allow a more complete understanding of the virulence and genetic diversity of bacterial strains, we examined full bacterial genomes in high resolution. To understand the evolution of uropathogenicity, we sequenced at over 190x coverage the genome of 19 *E. coli* strains isolated from UTI patients, 14 pathogenic strains from urine samples and 5 non-UTI-causing (commensal at the time of infection) rectal strains. We applied a *de novo* assembly-based algorithm to identify variants among the 19 strains, and constructed a whole-genome phylogeny based on these variants via a neighbor-joining algorithm.

In the whole-genome phylogeny, two commensal *E. coli* without typical combinations of pathogenicity genes formed the outgroup. This suggested that pathogenicity genes were present in infectious UPEC strains for a long time, with an estimated split from non-pathogenic *E. coli* over 32 million generations in the past. Even though our strains were collected in a small geographic area within a short period of time,

we found high pairwise genomic diversity between any two strains of *E. coli* in our sample, which was incompatible with recent epidemics of a subset of strains.

To contrast the evolutionary signature of the strains with the evolution of individual uropathogenic virulence factors, we constructed gene trees of the three most common virulence factors in our sample. Comparing the whole-genome phylogeny to gene trees of uropathogenic factors, we observed that the virulence gene trees displayed a topology distinct from the whole-genome tree, suggesting that the whole-genome phylogeny could not capture the specific evolutionary history of virulence factors. We identified no excess horizontal gene transfer at these virulence factors, as indicated by the observation that the topology of the virulence gene trees were not more different from the whole-genome tree than the topology of random region gene trees were from the whole-genome tree. Hence the uropathogenicity in our strains was not the result of a recent adaptation. Instead, uropathogencity appeared to be a maintained ability in a subset of *E. coli*. These UPEC carried uropathogenicity genes for a long time, and they used such virulence opportunistically.

5.2 Methods

5.2.1 Study design

We selected 19 *E. coli* isolates from 14 subjects with UTI, including 14 UPEC isolates from urine samples and 5 non-UTI-causing (commensal at the time of infection) rectal *E. coli*. The 5 commensal strains were isolated from the same individuals and at the same time point as 5 of the UPEC isolates The samples were selected from a large collection of previously isolated strains from female patients attending the same clinic for UTI between 1996 and 2007. Sampling from the collection was based on pathotypes as defined by Marrs et al. [74] based on common groupings of known uropathogenic factors (Table 5.1), in order to ensure a diversity of virulence

Pathotype	Virulence factors
Pathotype 1	*cnf1*, *hly*, $prsG_{J96}$
Pathotype 2	*cnf1*, *hly*, *sfa*
Pathotype 3	*aer*, *hly*, $papG_{ADIA2}$
Pathotype 4	*aer*, *kpsMT*, *ompT*, *drb*
Pathotype 5	*kpsMT*, *ompT*
Pathotype 0	all remaining strains

Table 5.1: Summary table of pathotype classification scheme, adapted from Marrs et al. [74]. Pathotypes are assigned hierarchically to over 800 UPEC strains by examining pairwise association of 10 known virulence factors.

factors in our sample. The 19 strains belonged to pathotypes 1 to 4 and pathotype 0, which is comprised of strains with no major groupings of uropathogenic factors. We employed a paired design, where each of the 5 commensal *E. coli* was isolated from an individual that also hosted one of the selected UPEC strains. We chose these pairs so that the commensal strain belonged to a different pathotype than the UPEC strain within each pair (Table 5.2).

We regrew the 19 *E. coli* isolates from $-80°$C stocks. We sequenced their genome using a single flow cell on Illumina HiSeq that produced 120 bp paired-end reads. The sequencing yield per sample ranged from 756 Mb to 1,328 Mb, totaling 19,535 Mb across all samples.

5.2.2 Variant calling

We employed a two-step, *de novo* assembly-based method (Cortex) to simultaneously reconstruct contigs and identify variants across multiple samples [37, 38]. This method is known to be conservative with high specificity [132]. Using a graph-building algorithm, Cortex constructs a colored de Bruijn graph from the sequence reads, where each color represents a sample. The resulting graph is error-cleaned by dynamically selecting a cleaning threshold from the coverage distribution. Divergence between samples exists as bubbles on the cleaned graph representation.

We employed the bubble calling algorithm in Cortex to detect variations between

Patient ID	Source	MLST Type	Pathotype	Code
1	Urine	ST 127	1	01U1
1	Fecal	ST 12	2	01F2
2	Urine	ST 12	1	02U1
2	Fecal	ST 131	4	02F4
3	Urine	ST 73	2	03U2
3	Fecal	ST 73	1	03F1
4	Urine	ST 144	3	04U3
4	Fecal	ST 2731	0	04F0
5	Urine	ST 404	4	05U4
5	Fecal	ST 10	0	05F0
6	Urine	ST 73	1	06U1
7	Urine	ST 127	1	07U1
8	Urine	ST 544	1	08U1
9	Urine	ST 127	1	09U1
10	Urine	ST 95	1	10U1
11	Urine	ST 73	2	11U2
12	Urine	ST 12	3	12U3
13	Urine	ST 131	4	13U4
14	Urine	ST 131	4	14U4

Table 5.2: Samples description of the pilot study. 19 *E. coli* isolates were selected from 14 female patients attending the same clinic for UTI. 14 isolates were UPEC from urine sample and 5 were commensal *E. coli* from rectal swab sample, paired with one of the 14 UPEC from the same individual. We listed the MLST type and pathotype [74] of each strain. We labeled each strain by a four-digit code: first two digits represent individual host ID (01-14), the third digit represents UPEC from urine (U) sample or commensal *E. coli* from fecal (F) sample. The last digit represents the pathotype of the strain.

samples. We used a low k-mer ($k = 31$) and a high k-mer ($k = 61$) to build assembly graphs because low k-mers allow discovery of variants at relatively lower coverage, and large structural variations and genome complexity are more accessible at high k-mers [37]. We combined variants called using the two k-mers into a joint call set. To get relative positions and to filter duplicate calls and overlapping sites, we aligned the assembled contigs, including each varying site and its flanking regions, with respect to each other. For the purpose of this study, we disregarded the complex variations including long segments of insertions, deletions or repeats and used only the single nucleotide polymorphisms (SNPs) for the following phylogenetic analyses.

We annotated all SNPs based on the genbank annotation of a uropathogenic *E. coli* reference strain (UTI89), using the coordinate-only method in Cortex [38]. We identified the phylogroup of each strain based on the presence and absence of three loci described in [14]. In addition, we used this annotation to tabulate the presence and absence of 23 virulence factors [108, 109, 135] in each strain.

5.2.3 Phylogenetic analyses

Using the SNPs identified from Cortex, we computed the pairwise sequence difference between samples and clustered them using a neighbor-joining algorithm [101]. We used *Escherichia fergusonii* (*E. fergusonii*) to root the phylogeny, since it is the closest species to *E. coli* [118]. To do so, we oriented the variants to *E. fergusonii* using the coordinate-only method in Cortex [38]. In this way, variant discovery was independent of the choice of rooting or reference genome. To measure the confidence of the whole-genome phylogeny, we employed a bootstrap algorithm to resample the sequences of variants from the samples 10,000 times and obtain bootstrap values of the branches. We applied Phylip [19] as neighbor-joining and bootstrapping algorithms. We studied clustering patterns on the phylogeny based on pathotype, host

individual, and sequence type (ST) defined by the University College Cork *E. coli* scheme [129]. To test the significance of specific clustering patterns, we calculated the probability of a cluster given the tree topology, under the null hypothesis that the labeling of the tree is completely random; a small probability indicates that the cluster is unlikely to occur by chance.

To understand strain divergence times, we scaled the tree branches using a calibrated substitution rate of *E. coli* from Wielgoss et al. [127]. The rate was inferred directly from tracking the accumulation of synonymous substitutions via whole-genome sequencing of 19 *E. coli* strains in a 40,000-generation evolution experiment. We compared this calibration with alternative substitution rates presented in the earlier literature that were based on comparing sequences with known divergence times [50, 88]. We categorized variants into synonymous and non-synonymous substitutions, and counted the number of synonymous and non-synonymous sites on the coding region, to estimate non-synonymous/synonymous rate ratio using a maximum likelihood method [131].

To contrast the evolution of the organism with the evolution of UTI virulence, we selected three UTI virulence factors: *hly, aer, kpsMT* [74]. Each was carried by over half of our sampled strains. We derived gene trees from the annotated variants called at each virulence factor and evaluated clustering by pathotype on these gene trees. The *hly* virulence factor consists of 4 genes: *hlyA*, *hlyB*, *hlyC* and *hlyD*, and the combined length is 7,281 bp. *aer* and *kpsMT* are 1,521 bp and 777 bp long, respectively. When constructing the respective gene tree, we considered only samples carrying the complete virulence factor. While *kpsMT* is the definitive virulence factor for pathotype 4, we discarded one pathotype 4 strain (14U4) in this construction due to low sequencing coverage at the region. The first 200 bp of the

777-bp region were sequenced at less than 10-fold for this particular strain.

To compare a virulence factors gene tree with the whole-genome tree, we reconstructed the whole-genome phylogeny based only on the samples carrying the virulence factor. We scaled branch lengths by the total number of variants on a tree. We then measured the similarity of the gene tree and the whole-genome tree using a topological score, generated by a branch-matching algorithm that searches for the optimal one-to-one transformation between two trees [87]. We contrasted the similarity score of each gene tree with an empirical distribution of similarity scores of trees containing the same number of leaves and same number of variants as the virulence gene tree. We generated this empirical distribution by randomly drawing sets of the same number of consecutive variants as each gene tree and generating trees based on these sets of variants. We then calculated the topological similarity score of each random tree to the whole-genome tree, which gave us an empirical distribution of similarity scores. A score at the extremes of the empirical distribution indicates that the gene tree is significantly more different from or more similar to the whole-genome tree than random regions of the genome.

5.3 Results

5.3.1 Whole-genome phylogeny

Using *de novo* assembly-based variant calling methods, we identified 68,396 SNPs with a transition-to-transversion ratio of 2.73. All our 19 strains belonged to the phylogroup B2. We oriented 24,568 of the variant set to the *E. fergusonii* outgroup sequence coordinates and constructed a rooted phylogeny (Figure 5.1). Most splits on this whole-genome phylogeny had bootstrap values of 100%, while two splits had 95-100% bootstrap values, and three had 65-95% bootstrap values.

Applying the *E. fergusonii* gene annotation to our variant set, we identified 11,216

synonymous mutations (45.7% of the variants), and counted 963,414 synonymous sites on the oriented genome. The maximum likelihood estimate of the ratio of the number of non-synonymous substitutions per non-synonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s) was 0.54, indicating purifying selection consistent with previous findings [39]. Using an estimated substitution rate of 8.9×10^{-11} per base pair per generation, based on the laboratory evolution of *E. coli* [127], the evolutionary time elapsed on the entire phylogeny was over 130 million generations. Based on this calibration, we expected 1 synonymous mutation per 11,600 generations. Alternative substitution rate estimates based on sequences with known divergence times were 5.2×10^{-11} per base pair per generation in Lecointre et al. [50], and 3×10^{-11} per base pair per generation in Ochman et al. [88], which led to 1.5- to 3-fold shorter evolutionary time.

Two non-UTI-causing (commensal) strains, both belonging to pathotype 0, formed the outgroup on the rooted phylogeny (Figure 5.1) ($p = 0.0058$). The split between this outgroup and the remaining phylogeny represents the time of divergence of UPEC strains from commensal strains occurred; we estimated a split time of 32 million generations ago.

We observed that the clustering of strains on our whole-genome phylogeny did not correspond to pathotype classification. Strains of pathotypes 1, 2 and 3 showed no distinct subclades and together formed a single large cluster, regardless of whether the strain was a commensal or uropathogenic *E. coli* (Figure 5.1). Similarly, applying the grouping methods of UPEC based on presence of several virulence genes described in Tarchouna et al. [111] (Grouping 1) and Yun et al. [135] (Grouping 2), we observed that none of the groups fell completely and distinctively into subclades (Figure 5.1, Figure 5.3). In pathotype and Grouping 2 classifications, each had one

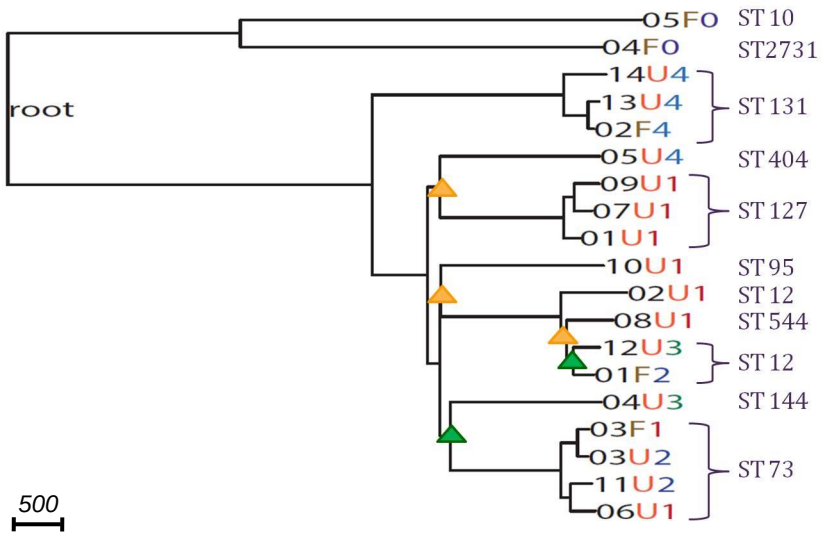


Figure 5.1: Phylogeny constructed from whole-genome assembly-based variants. The phylogeny is rooted by *E. fergusonii* (long branch not shown). Nodes are labeled by the sample codes as listed in Table 5.2. All strains are of phylogroup B2. Branches with circles represent 60-95% bootstrap values, and branches with triangles represent 95-100% bootstrap values. All unmarked branches have 100% bootstrap value. Scale shows length on branches representing 500 pairwise sequence differences.

type with four strains where three strains formed a significant subclade ($p = 0.0041$), but the remaining strain of the same type was far away from the subclade on the phylogeny. In the pathotype classification, three pathotype 4 strains of ST 131 clustered, but the remaining pathotype 4 strain (05U4) in our sample had a pairwise sequence difference of over 4,700 with the other pathotype 4 strains. The split of 05U4 with the other pathotype 4 subclade happened over 10 million generations ago. Among the three strains that clustered, the shortest external branch leading to 02F4 still represented over 320,000 generations, indicating long divergence times between pathotype 4 strains (Figure 5.1). Similarly, based on Grouping 2 classification, three type 6 strains of ST 127 clustered, with a mean pairwise sequence difference of over 500. However, the remaining group 6 strain (02U1) had a pairwise sequence difference

of over 3,800 with the other group 6 strains (Figure 5.3). The split of 02U1 with the other type 6 subclade happened over 7 million generations ago. Moreover, when tabulating the virulence genes present in each strain, we found that strains from the same ST did not necessarily carry the same set of virulence genes (Table 5.5).

To investigate if multilocus sequence typing (MLST) based on seven housekeeping *E. coli* loci was consistent with our whole-genome phylogeny, we identified the ST of each strain in our sample (Table 5.3). While most strains in our sample were singletons in the ST classification, four STs had three or more strains in our sample, namely, ST 12, ST 73, ST 127 and ST 131 (Table 5.3). We observed that organisms with the same ST mostly formed consistent subclades on the phylogeny (Figure 5.1), with the exception of the ST 12 cluster, which also contained one ST 544 strain. However, all splits defining this cluster had less than 100% bootstrap value. Nonetheless, strains within each ST still had remarkable diversity (Table 5.3): For ST 73, ST 127 and ST 131, each pair of strains within its respective ST had on average over 500 differences, reflecting a divergence of $> 360,000$ generations. ST 127 strains displayed the highest overall similarity with an average of 501.3 pairwise differences. The second most similar type was ST 73 with an average of 554.5 pairwise differences. ST 131 strains had an average of 597.3 pairwise differences. ST 12 strains were more diverse, with an average of 907 pairwise differences.

Finally, when evaluating the matched pairs of one commensal and one UPEC strain sampled from the same individual, we observed that only one pair clustered with a pairwise difference of 237 variants. The other four pairs were located very far apart on the tree (Figure 5.1); they did not cluster in the same subclade, indicating no significant excess of clustering ($p = 0.19$).

		ST 131			ST 127			ST 73				ST 12		
		14U4	13U4	02F4	09U1	07U1	01U1	03F1	03U2	11U2	06U1	02U1	12U3	01F2
	14U4	0	829	785										
ST 131	13U4		0	178										
	02F4			0										
	09U1				0	462	497							
ST 127	07U1					0	545							
	01U1						0							
	03F1							0	237	638	712			
ST 73	03U2								0	563	681			
	11U2									0	496			
	06U1										0			
	02U1											0	1195	1015
ST 12	12U3												0	512
	01F2													0

Table 5.3: Pairwise sequence differences of strains belonging to the same ST.

5.3.2 Phylogeny of virulence factors

We constructed 3 gene trees from known uropathogenic virulence factors [74] that are present in multiple strains in our sample: *aer*, *hly* and *kpsMT* (Figure 5.2). Each gene tree consisted of a different number of tree leaves, because not all virulence factors were found in all strains: All 19 strains carried *aer*, 13 strains carried the complete segment of *hly*, while 16 strains carried *kpsMT* (Table 5.4). After scaling branch lengths by the number of mutations on a tree, we compared each gene tree to a subtree of the whole-genome phylogeny consisting of the corresponding subset of strains. For this comparison, we applied a topological score [87] which summarizes the percentage of topological similarity between branches of two trees.

The *aer* gene tree, containing all 19 strains, possessed similar structural features as the whole-genome tree (Figure 5.2a) with a topological score of 0.635 (Table 5.4). The difference between the two trees was best illustrated by the two strains 04U3, 05U4 that segregated differently in this gene tree than on the whole-genome phylogeny. Some strains carried identical copies of the virulence gene. For example, pathotype 1 UPEC strains (01U1, 07U1, 09U1) displayed no pairwise sequence difference at this gene. The two fecal strains with virulence factors (03F1, 01F2) were also identical at this gene, so were three other UPEC (02U1, 03U2, 06U1). When we assessed similarity by comparing the topology score against to the empirical distribution of scores, we saw no signal that the similarity between the aer gene tree and whole-genome tree was higher or lower than other trees containing the same number of variants ($p = 0.318$, Table 5.4).

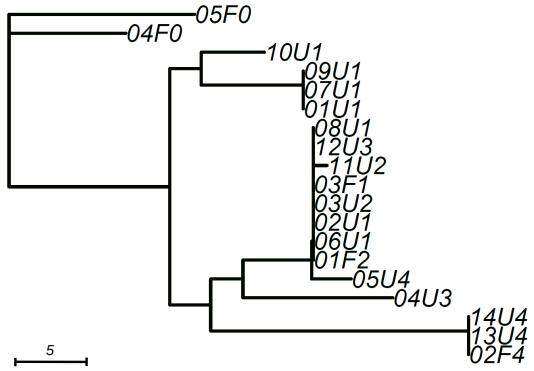
The *hly* gene tree showed more differences from the whole-genome tree with a topological score of 0.565 (Table 5.4). Only the 13 strains of pathotypes 1,2 and 3 carried the complete segment of *hly* (Figure 5.2b). On the *hly* gene tree (Figure

Virulence gene	Gene length	Number of carriers	Topological Similarity Score	p-value
aer	1,521 bp	19	0.635	0.318
hly	7,281 bp	13	0.565	0.185
kpsMT	777 bp	16	0.516	0.209

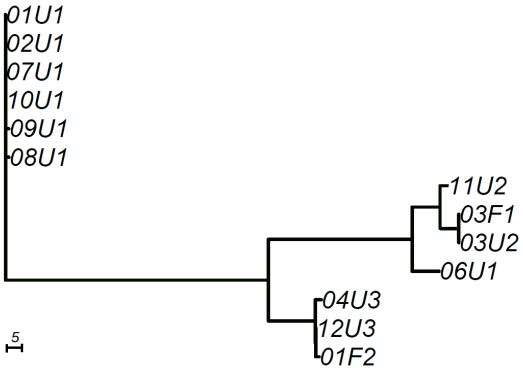
Table 5.4: Virulence gene trees: gene length, number of carriers (out of 19 samples) , topological similarity scores compared to whole-genome tree and p-values of these scores generated from the empirical distribution of scores from random trees.

5.2b), we observed that four pathotype 1 UPEC strains (01U1, 02U1, 07U1, 10U1) carried the identical copy of *hly* as the UTI89 reference strain. Two other pathotype 1 strains (08U1, 09U1) had a *hly* copy with only one base different from the UTI89 *hly*. The *hly* regions of the remaining two pathotype 1 strains were significantly different from the other pathotype 1 *hly* regions. 03F1 and 06U1 each displayed over 100 pairwise sequence differences at *hly* when compared to the UTI89 *hly*. These two *hly* were more similar to the pathotype 2 *hly*, as a result formed a cluster on the *hly* tree. The *hly* region of the pathotype 2 fecal strain were very similar to those of pathotype 3 strains, with 1 and 3 pairwise differences respectively. While the similarity score of the *hly* gene tree to whole-genome tree (0.565) was lower than that of the *aer* gene tree (0.635), there was no strong signal that the *hly* tree was less similar to the whole-genome tree than random genomic regions ($p = 0.185$, Table 5.4).

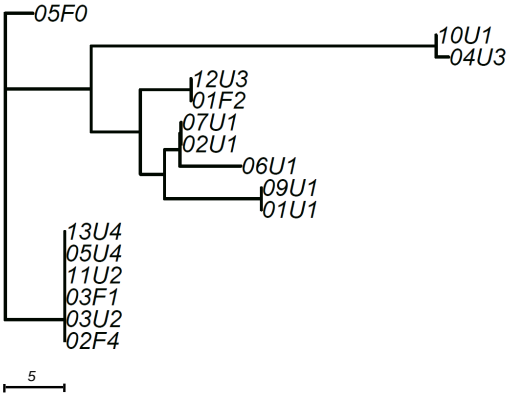
The *kpsMT* gene tree consisted of 16 strains and was the least similar to the whole-genome phylogeny among the three virulence factors studied, with a topological score of 0.516 (Figure 5.2c, Table 5.4). This factor was completely absent in a commensal pathotype 0 strain (04F0) and one pathotype 1 UPEC strain (08U1). The resulting gene tree showed that four UPEC and two commensal strains were identical at this gene The remaining strains displayed considerable diversity at this gene, as indicated by the longer branches. The longest branch leading to the 10U1 and 04U3 subclade contained 28 mutations. Notably, strains that clustered closely



(a) *aer*



(b) *hly*



(c) *kpsMT*

Figure 5.2: Unrooted trees derived from three selected virulence factors: (a) *aer*, (b) *hly*, and (c) *kpsMT* for uropathogenic and commensal *E. coli*. Each gene tree consists of a different number of tree leaves, as not all virulence factors occurred on all strains. Scale shows length on branches representing 5 pairwise sequence differences.

together in the previous two gene trees appeared to be further apart on this gene tree. However, the *kpsMT* gene tree was not significantly less similar to the whole-genome tree than random genomic regions ($p = 0.209$, Table 5.4).

5.4 Discussion

We studied the evolution of pathogenicity in uropathogenic *E. coli* (UPEC) with the goal of understanding the origin and spread of UPEC in the context of urinary tract infections. We sampled strains from a large collection of *E. coli* isolates with well-characterized virulence factors [74], in order to expand the diversity of virulence factors in our sample. This approach is different from most existing studies where sampling is based on clinical visits [111, 135]. This study design allowed us to study a broader spectrum of virulence factors in order to understand the evolution of uropathogenicity.

Using whole-genome deep-sequencing, we explored whole bacterial genomes at high resolution, allowing more detailed analyses than pathotype or MLST schemes that study only small regions of the genome. We employed a multi-sample *de novo* assembly algorithm Cortex that simultaneously assembles genomes and calls variants [37]. This method calls variants independently of a reference genome. Hence, the variant calls are unaffected by the choice of reference sequence, making this approach well-suited to a sample with high diversity such as ours. The variant calls generated from Cortex are known to be conservative with high specificity [132]. These high quality whole-genome variants allowed a more accurate investigation of the evolutionary pathways of uropathogenicity and the degree of diversity among strains.

Our phylogenetic analyses contrasting UPEC with non-pathogenic *E. coli* showed that their divergence happened over 32 million generations ago, which is equivalent to

107,000 to 320,000 years, assuming the *E. coli* had 100-300 generation per year [88]. Alternative estimates of substitution rates gave qualitatively consistent results, with an estimated 10.6-18.7 million generations on the whole-genome phylogeny. Between pathogenic strains, the whole-genome pairwise diversity was high, corresponding to a long divergence history of over 130 million generations, or 0.4-1.3 million years. Both of these estimates from the whole-genome phylogeny showed that within the small geographic region of our sample collection, UTI were caused by strains of multiple origins. In addition, commensal and UPEC strains from the same individual were as different from each other as from other strains in the sample, suggesting that the infection was unlikely caused by gut *E. coli* strains that recently acquired uropathogenicity factor within the human host.

Our phylogenetic analysis of the entire *E. coli* genome allowed us to evaluate more basic methods of *E. coli* classification such as the MLST scheme and pathotype groupings based on virulence factors. The pathotype classification we used for sampling did not capture the overall relationship of strains. In particular, pathotypes 1, 2 and 3 did not form distinctive clusters on the whole-genome phylogeny. When we applied to our sample alternative groupings of UPEC strains based on other distinct sets of virulence factors [111, 135], we also observed that most groups did not cluster well on the whole-genome phylogeny. For the pathotype and group that had three out of four strains forming a significant subclade, we observed high diversity among the three strains that clustered, and a deep split between the subclade and the remaining strain of the same type. Therefore, classification based on presence and absence of virulence factors did not appear to be meaningful for understanding the evolutionary history of UPEC strains.

On the other hand, classification based on the traditional MLST scheme did gen-

erally capture the evolutionary relationship of strains. However, our whole-genome phylogeny identified high diversity among strains that were classified into the same sequence type, something that was previously not well-studied. The divergence time of particular clusters, for example the pathotype 4 ST 131 cluster, was longer than suggested by previous studies [13, 85]. Long branch lengths of the ST 131 cluster reflected ancient origins and high diversity within the ST. Consistent with our findings, a recent study using a phenotypic microarray showed that ST 131 was not a distinct lineage of ExPEC [4]. We observed a similar level of diversity in other sequence types, with substantial variation among strains of the same ST of over 100 pairwise differences. Long divergence times among strains of the same ST suggests that they are not clonal, as they evolve to accumulate large number of genomic differences over time. Moreover, by tabulating the presence and absence of 23 potential uropathogenic virulence factors in our strains, we observed that strains of the same ST carried different sets of virulence factors. Therefore, classifying UPEC by sequence type is not sufficient for drawing inferences on the presence of UPEC virulence factors.

We further explored the evolutionary pathways of individual uropathogenic virulence factors by constructing gene trees of three virulence factors that were the most common in our sample. We aimed to identify evidence for horizontal gene transfer as horizontal gene transfer is the major mechanism for non-pathogenic *E. coli* to acquire uropathogenicity. If uropathogenic *E. coli* strains acquired pathogenicity via elevated rates of horizontal gene transfer, or preferential selection of horizontally transferred virulence genes, the corresponding gene tree would display a significantly different topology from the whole genome tree. Therefore, we contrasted the topology of the gene trees and the whole-genome phylogeny to see if virulence genes displayed

distinct evolutionary pathways. We found that each virulence gene tree had an evolutionary pathway distinct from the whole-genome phylogeny, indicated by the 50-65% topological similarity scores. Based on the empirical distribution of similarity scores of random gene regions, the virulence gene trees did not have extreme similarity scores, meaning these uropathogenic genes were not significantly more similar to the whole-genome phylogeny than other genomic regions. This suggests no signal of excess horizontal gene transfer and no selective advantage at the uropathogenic genes than elsewhere in the genome.

In summary, by quantifying the diversity of UPEC strains using whole-genome deep-sequencing and contrasting with commensal *E. coli*, we showed that UPEC had a long evolutionary history since their divergence from non-UTI-causing commensal *E. coli*. Our study illuminated the development of UTI and showed that UPEC are opportunistic, conserving uropathogenic virulence factors without signals of preferential selection or increased rates of horizontal gene transfer. Our results indicated that the phylogenetic relationship of UPEC provided only limited information about the presence of virulence factors and thus suggested that closely related UPEC may have dissimilar uropathogenic phenotypes. Further extensive sequencing of UPEC and commensal *E. coli* will allow deeper understanding of the genetic signals and mechanisms driving the epidemiology of uropathogenicity.

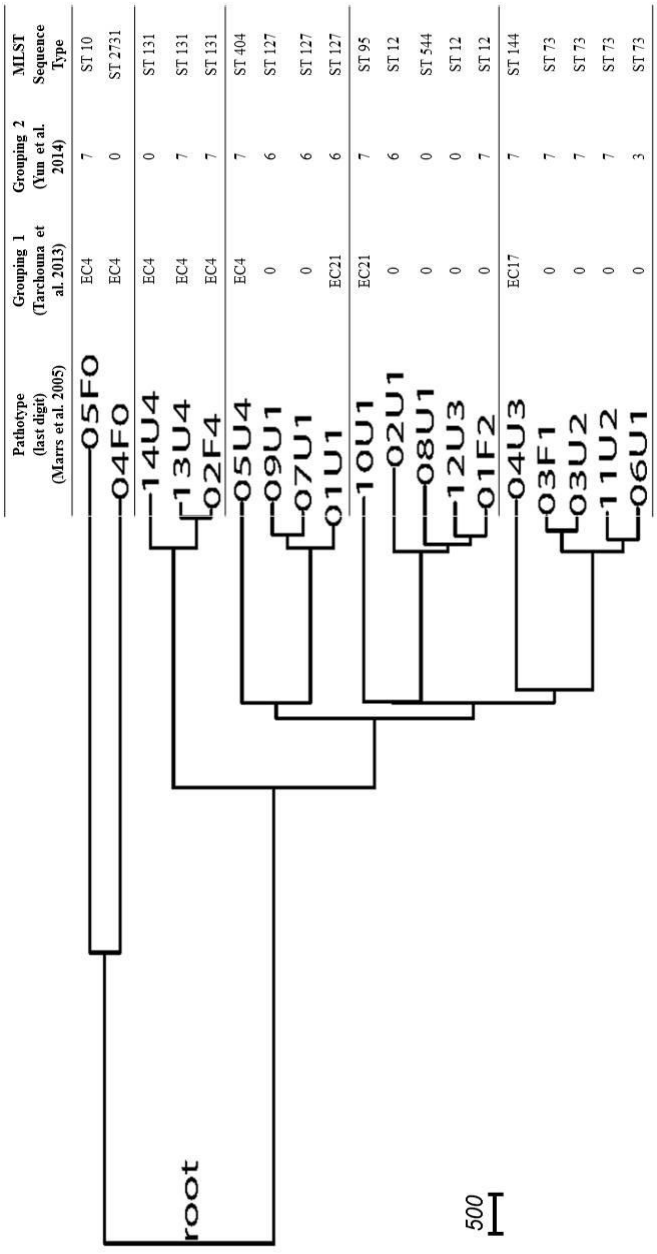


Figure 5.3: Classifications based on presence and absence of several virulence factors, described in Marrs et al. [74], Tarchouna et al. [111] and Yun et al. [135], and multilocus sequence typing (MLST) based on presence and absence of seven housekeeping genes. For Groupings 1 and 2, type "0" refers to the combinations of virulence factors in our sampled strains that were not found in the original studies. MLST types correspond well to the whole-genome phylogeny, while none of the groupings based on uropathogenic virulence factors were consistent with the whole-genome clustering. Three pathotype 4 strains formed a significant subclade, but the remaining pathotype 4 strain had a split with this subclade over 10 million generations ago. Similarly, using Grouping 2, three type 6 strains formed a significant subclade, while the remaining group 6 strain showed a deep split with this subclade over 7 million generations ago.

Sample Code	05F0	01F2	02U1	12U3	07U1	09U1	02F4	13U4	14U4	04U3	04F0	05U4	08U1	06U1	03U2	03F1	11U2	10U1
MLST Type	ST 10	ST 12	ST 12	ST 12	ST 127	ST 127	ST 131	ST 131	ST 131	ST 144	ST 2731	ST 404	ST 544	ST 73	ST 73	ST 73	ST 73	ST 95
aer	+	+	+	+	+ +	+	+	+	+	+	+	+	+	+	+	+	+	
afa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
aufA	-	+	+	+	+	+	-	-	-	+	-	+	+	+	+	+	+	+
chuA	-	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+
cnf1	-	+	+	-	+	+	-	-	-	-	-	-	+	+	+	+	+	+
feoB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
fimH	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
fyuA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
hlyA	-	+	+	+	+	+	-	-	-	+	-	-	+	+	+	+	+	+
iroN	-	+	+	+	-	-	-	-	-	+	+	-	+	+	+	+	+	+
iutA	+	+	-	+	-	-	+	+	+	+	+	+	-	-	+	+	+	-
kpsMT	+	+	+	+	+	+	+	+	-	+	-	+	-	+	+	+	+	+
ompT	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
papA	-	+	+	-	-	-	-	-	-	+	-	-	-	-	+	+	-	+
papC	-	+	+	+	+	+	-	-	-	+	-	-	+	+	+	+	+	+
papG	-	-	+	-	+	+	-	-	-	-	-	-	+	+	+	+	-	+
sfaS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
vat	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
yadN	-	+	+	+	+	+	-	-	-	+	-	+	+	+	+	+	+	+
yejE	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
yfcV	-	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+
ygiL	-	+	-	+	-	-	+	+	+	+	-	-	+	+	+	+	+	+
yniA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 5.5: List of virulence factors in each sampled strain. We annotated our sampled strains with the UTI89 reference. Based on the virulence factors compiled from Spurbeck et al. [108] and Yun et al. [135] and conditioning on their presence in UTI89, we listed the presence and absence of virulence factors in our sample. Sample code of each strain refers to: Individual host ID (first two digits), Urine/ Fecal sample (U/F in the third digit), and pathotype (last digit).

CHAPTER VI

Discussion

In this dissertation, we have demonstrated extensive methodological developments, statistical analyses and applications of next-generation sequencing (NGS), which is the state-of-the-art technology for genetic studies. We have shown that we can generate high quality rare variants by applying individual-based variant calling to heterogeneous coverage targeted data, and by resolving the dependence in overlapping read pairs. We have also provided a statistical model for separately estimating machine and fragment errors in a sequencing experiment. We have applied NGS techniques to sequence uropathogenic bacteria and to make inferences on the evolution of virulence. In this chapter, we discuss the significance of this dissertation and address future challenges of the continuous development of sequencing technologies.

In Chapter 2, we evaluated existing likelihood-based variant calling algorithms on targeted sequencing data and recommended the combined use of individual-based and population-based single marker callers. These two callers together generated the most complete call set with the highest quality, because the individual-based caller identified a larger number of high-quality rare variants, particularly singletons, while the population-based caller produced higher quality common variants than the individual-based caller. In practice, however, applying two callers on the same

dataset is computationally inefficient. Moreover, for large-scale sequencing datasets, storing and merging multiple call sets impose heavy burden in computer hardware requirements. Therefore, to optimize our recommended variant calling strategy for broader and more practical uses, we suggest a flexible calling approach, where we use the individual-based caller to obtain a union set of variable sites across all individuals in the sample. Then the population-based caller follows to generate genotypes per variant site in the union set. In this way, we are still able to identify the same number of singleton variants, while generating high-quality genotypes across all variant sites. Each genotype is only called once, thereby achieving computation efficiency and removing the need of merging across call sets.

After the development of our recommended variant calling strategy, the flexible approach was recently implemented as glfFlex (`https://github.com/statgen/glfFlex`). The BRIDGES Consortium applied glfFlex to analyze >2,500 bipolar cases and controls in their whole-genome sequencing study. Preliminary results showed that the flexible approach led to the discovery of about 20% more SNPs than using population-based calling. Most of these additional SNPs were singletons. While more detailed quality control and analyses are needed to evaluate the actual gain of high-quality SNPs, these results showed that our method of combined calling algorithms is effective not only in targeted-sequencing datasets, but also in whole-genome sequencing datasets, particularly those with homogeneously low coverage; in the case of the BRIDGES study the average coverage was 8x. As the individual-based caller showed the most improvement in identifying singletons over the population-based caller at low coverage, it is not surprising that our method can be directly extended to analyze large-scale whole-genome sequencing data.

Statistical methods involved in NGS analysis pipelines often need extra attention

to avoid false discoveries due to sequencing artefacts. Chapters 3 and 4 addressed the issue of overlapping read pairs. Overlapping reads are a common sequencing artefact, which has often been overlooked, but which leads to overestimation of genotype accuracy. At the same time, simply discarding overlapping reads can lead to data loss for downstream analyses. In these two chapters we presented novel methods of utilizing overlapping reads. In Chapter 3, we took advantage of the fragment dependence property in overlapping read pairs to separately estimate fragment and machine errors. We provided a Markov Chain Monte Carlo algorithm to estimate machine errors from discordant overlapping bases and modeled fragment errors based on concordant overlapping bases that are different from the reference. In Chapter 4, we designed the RESCORE algorithm to resolve the fragment dependence in overlapping read pairs, while retaining the machine error estimates independently provided by the two overlapping reads. Using the RESCORE algorithm increased the number of high-quality novel variant calls. Thus, we illustrated the utility of overlapping reads data that would otherwise have been discarded and furthered the understanding of error sources in a sequencing experiment. More importantly, our methods processed overlapping reads efficiently and showed promising improvement in novel variant calls.

We anticipate these methods can be applied to analyze more reads affected by other sequencing artefacts. One straightforward extension is to incorporate optical duplicate reads, i.e. overlapping read pairs that completely overlap each other. Duplicate reads are very common in sequencing studies [82, 114, 116]; even with the newest PCR-free sequencing approach that removes amplicon duplicate reads and has a 2-fold reduction in duplicate reads [33], we still observe a mean of 9.7% duplicate reads in recent large-scale sequencing datasets of over 5,700 individuals

(TopMed study, personal communication with Hyun Min Kang). These duplicates are routinely discarded before variant calling because they replicate the same fragment error , creating the same dependence problem as overlapping reads. Since methods for detecting and flagging or removing duplicate reads have been widely used in sequencing pipelines [40, 76], RESCORE can easily be modified to assign a new quality score to each duplicated base that reflects combined sequencing information from the flagged duplicates. We expect that by recovering the data lost by discarding duplicate reads, variant calls can achieve even higher accuracy. Moreover, by including duplicate reads in the error estimation models, we can generate more precise estimates for the machine and fragment errors.

As sequencing technologies improve, reads generated are longer. From our estimation of the marginal distribution of machine errors contributed by read cycle, which is consistent with previous knowledge, the quality of sequenced bases decreases along the length of a read. Therefore, given the success of RESCORE in resolving the overlapping bases, it poses an interesting question of whether we should intentionally create overlapping reads with the goal of improving the quality at the ends of the read pairs. While further assessment is required to quantify the impact of RESCORE on high-coverage datasets, we foresee a modest gain of novel variants from applying RESCORE. After RESCORE combines the overlapping reads and improves the quality at the read ends, the genotype likelihoods of all variants will be more accurate.

Chapters 2 to 4 of the dissertation collectively presented novel statistical methods for NGS studies to promote genetic discoveries, particularly the discovery of rare variants, which have extensive implications in the understanding of diseases and complex traits. Indeed, NGS is powerful for exploring any genome. However, with

most methods and tools designed mainly for analyzing the human genome, challenges are involved in applying these methods to bacterial sequencing data due to their genomic difference. For example, due to the frequent occurrence of repeated segments and multiple copies of the same genes in the bacterial genome, alignment of NGS reads is challenging. The diversity among bacterial strains also makes the choice of reference genome a non-trivial question. Genes that are laterally transferred to a specific bacterial strain may not be present in the chosen reference genome; hence the alignment procedure is inaccurate at reconstructing the genome of the sequenced strain. Furthermore, variant calling algorithms typically assume a diploid genome where there are three possible genotypes at a biallelic locus, while for a haploid genome there are only two possible genotypes.

In Chapter 5, we presented one application of NGS methods and analyses to understand the evolutionary pathways of virulence genes in uropathogenic *E. coli*. We carefully designed the analysis pipeline which uses tools that are robust to bacterial sequencing data, which enables the use of this pipeline on large number of sequences across different bacterial species. Our analyses showed that for three common virulence genes associated with uropathogenecity, the evolutionary pathways differed from that of the whole-genome phylogeny. However, these genes did not have elevated rates of horizontal gene transfer, suggesting that there was no selective advantage of these virulence genes over other genomic regions. We project that our analyses can be directly extended to study the evolutionary pathways of all annotated genes along the bacterial genome. By comparing each gene tree to the whole-genome phylogeny, we can identify genes that have significantly different evolutionary histories from the rest of the genome; these genes are likely important in the development of pathogenicity.

NGS studies have, and will continue to unravel numerous genetic discoveries. This dissertation presents our efforts in developing new methods and applications, and providing practical guidelines and strategies for analyzing a wide range of NGS datasets. We project that this work can be extended to accommodate new features in the continuously evolving technology, thereby contributing to further advances in the field of genetics and to our understanding of the genetic basis of diseases.

BIBLIOGRAPHY

[1] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[2] D. Aird, W.-S. Chen, M. Ross, K. Connolly, J. Meldrim, C. Russ, S. Fisher, D. Jaffe, C. Nusbaum, and A. Gnirke. Analyzing and minimizing bias in illumina sequencing libraries. *Genome biology*, 11(Suppl 1):P3, 2010.

[3] M. Akbari, M. D. Hansen, J. Halgunset, F. Skorpen, and H. E. Krokan. Low copy number dna template can render polymerase chain reaction error prone in a sequence-dependent manner. *The Journal of molecular diagnostics*, 7(1):36–39, 2005.

[4] A. Alqasim, R. Emes, G. Clark, J. Newcombe, R. L. Ragione, and A. McNally. Phenotypic microarrays suggest *Escherichia coli* st131 is not a metabolically distinct lineage of extra-intestinal pathogenic *E. coli*. *PLoS One*, 9(2):e88374, 2014.

[5] W. J. Ansorge. Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203, 2009.

[6] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.*, 12:745–755, 2011.

[7] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, Chiara, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O/'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano,

C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[8] M. A. Brockhurst, N. Colegrave, and D. E. Rozen. Next-generation sequencing as a tool to study microbial evolution. *Molecular ecology*, 20(5):972–980, 2011.

[9] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12:703–714, 2011.

[10] S. L. Chen, M. Wu, J. P. Henderson, T. M. Hooton, M. E. Hibbing, S. J. Hultgren, and J. I. Gordon. Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Science Translational Medicine*, 5(184):184ra60, May 0 2013.

[11] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkalovglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton. Genetic diagnosis by whole exome capture and massively parallel dna sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 106(45):19096–19101, 2009.

[12] E. T. Cirulli and D. B. Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010.

[13] G. Clark, K. Paszkiewicz, J. Hale, V. Weston, C. Constantinidou, C. Penn, M. Achtman, and A. McNally. Genomic analysis uncovers a phenotypically diverse but genetically homogeneous *Escherichia coli* st131 clone circulating in unrelated urinary tract infections. *J. Antimicrob. Chemother.*, 67(4):868–877, 2012.

[14] O. Clermont, S. Bonacorsi, and E. Bingen. Rapid and simple determination of the escherichia coli phylogenetic group. *Applied and Environmental Microbiology*, 66(10):4555–4558, 2000.

[15] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson, and H. H. Hobbs. Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science*, 305(5685):869–872, 2004.

[16] G. Curocichin, Y. Wu, T. W. McDade, C. W. Kuzawa, J. B. Borja, L. Qin, E. M. Lange, L. S. Adair, L. A. Lange, and K. L. Mohlke. Single-nucleotide polymorphisms at five loci are associated with c-reactive protein levels in a cohort of filipino young adults. *J. Hum. Genet.*, 56(12):823–827, 2011.

[17] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, and M. Hanna. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, 2011.

[18] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome research*, 8(3):186–194, 1998.

[19] J. Felsenstein. Phylip (phylogeny inference package) version 3.69, 2005.

[20] J. Flannick, J. M. Korn, P. Fontanillas, G. B. Grant, E. Banks, M. A. Depristo, and D. Altshuler. Efficiency and power as a function of sequence coverage, snp array density, and imputation. *PLoS computational biology*, 8(7):e1002604, 2012.

[21] J. Flannick, G. Thorleifsson, N. L. Beer, S. B. Jacobs, N. Grarup, N. P. Burtt, A. Mahajan, C. Fuchsberger, G. Atzmon, R. Benediktsson, et al. Loss-of-function mutations in slc30a8 protect against type 2 diabetes. *Nature genetics*, 46(4):357–363, 2014.

[22] B. Foxman. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *Am. J. Med.*, 113(1):5–13, 2002.

[23] B. Foxman and P. Brown. Epidemiology of urinary tract infections: transmission and risk factors, incidence, and costs. *Infect. Dis. Clin. North Am.*, 17(2):227–241, 2003.

[24] W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013.

[25] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[26] Genome Analysis Wiki. Bamutil:clipoverlap. `http://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap`. Last accessed: 2015-07-15.

[27] Genome Analysis Wiki. Bamutil:recab. `http://genome.sph.umich.edu/wiki/BamUtil:_recab`. Last accessed: 2015-07-15.

[28] Genome Analysis Wiki. glfsingle. `http://genome.sph.umich.edu/wiki/GlfSingle`. Last accessed: 2015-03-26.

[29] T. M. Gibreel, A. R. Dodgson, J. Cheesbrough, A. J. Fox, F. J. Bolton, and M. Upton. Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from northwest england. *Br. Soc. Antimicrob. Chemo.*, 67(2):346–356, 2012.

[30] A. Hodgkinson and A. Eyre-Walker. Human triallelic sites: evidence for a new mutational mechanism? *Genetics*, 184(1):233–241, 2010.

[31] C. Huebner, I. Petermann, B. L. Browning, A. N. Shelling, and L. R. Ferguson. Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: Mdr1 (abcb1) g2677/t/a as an example. *Cancer Epidemiol. Biomarkers Prev.*, 16(6):1185–1192, 2007.

[32] L. Ilie, F. Fazayeli, and S. Ilie. Hitec: accurate error correction in high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 27(3):295–302, 2011.

[33] Illumina. Data sheet: Sequencing - truseq DNA PCR-free sample preparation kit. `http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf`.

[34] Illumina. Specification sheet: Sequencing - HiSeq X series of sequencing systems. `http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf`.

[35] Illumina. White paper: Informatics - reducing whole-genome data storage footprint. `http://www.illumina.com/documents/products/whitepapers/whitepaper_datacompression.pdf`.

[36] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[37] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. *de novo* assembly and genotyping of variants using colored de bruijn graphs. *Nat.Genet.*, 44(2):226–232, 2012.

[38] Z. Iqbal, I. Turner, and G. McVean. High-throughput microbial population genomics using the cortex variation assembler. *Bioinformatics*, 29:275–276, 2013.

[39] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Microevolutionary genomics of bacteria. *Theor. Popul. Biol.*, 61:435–447, 2002.

[40] G. Jun, M. K. Wing, G. R. Abecasis, and H. M. Kang. An efficient and scalable analysis framework for variant extraction and refinement from population scale dna sequence data. *Genome Res.*, pages gr–176552, 2015.

[41] J. Kaiser. Affordable 'exomes' fill gaps in a catalog of rare diseases. *Science*, 330:903–903, 2010.

[42] J. B. Kaper, J. P. Nataro, and H. L. Mobley. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.*, 2(2):123–140, 2004.

[43] A. Keinan and A. G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *science*, 336(6082):740–743, 2012.

[44] A. Kiezun, K. Garimella, R. Do, N. O. Stitziel, B. M. Neale, P. J. McLaren, N. Gupta, P. Sklar, P. F. Sullivan, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, T. Lehner, Y. Y. Shugart, A. L. Price, P. I. W. de Bakker, S. M. Purcell, and S. R. Sunyaev. Exome sequencing and the genetic basis of complex traits. 44:623–630, 2012.

[45] M. Kircher. Analysis of high-throughput ancient dna sequencing data. In *Ancient DNA*, pages 197–228. Springer, 2012.

[46] C. Kozanitis, C. Saunders, S. Kruglyak, V. Bafna, and G. Varghese. Compressing genomic sequence fragments using slimgene. *Journal of Computational Biology*, 18(3):401–413, 2011.

[47] M. Landgren, H. Odén, I. Kühn, A. Österlund, and G. Kahlmeter. Diversity among 2481 *Escherichia coli* from women with community-acquired lower urinary tract infections in 17 countries. *J. Antimicrob. Chemother.*, 55(6):928–937, 2005.

[48] L. A. Lange, Y. Hu, H. Zhang, C. Xue, E. M. Schmidt, Z.-Z. Tang, C. Bizon, E. M. Lange, J. D. Smith, E. H. Turner, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol. *The American Journal of Human Genetics*, 94(2):233–245, 2014.

[49] S. Q. Le and R. Durbin. Snp detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, 21:952–960, 2010.

[50] G. Lecointre, L. Rachdi, P. Darlu, and E. Denamur. Escherichia coli molecular phylogeny using the incongruence length difference test. *Molecular biology and evolution*, 15(12):1685–1695, 1998.

[51] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014.

[52] B. Li and S. M. Leal. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet*, 5(5):e1000481, 2009.

[53] H. Li. Improving snp discovery by base alignment quality. *Bioinformatics*, 27(8):1157–1158, 2011.

[54] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[55] H. Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, page btu356, 2014.

[56] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

[57] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, 11(5):473–483, 2010.

[58] H. Li, J. Ruan, and R. M. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858, 2008.

[59] R. Li, Y. Li, K. Kristiansen, and J. Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[60] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[61] Y. Li, C. Sidore, H. Kang, M. Boehnke, and G. Abecasis. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.*, 21:940–951, 2011.

[62] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, 34(8):816–834, 2010.

[63] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012, 2012.

[64] X. Liu, S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang. Variant callers for next-generation sequencing data: A comparison study. *PloS one*, 8(9):e75619, 2013.

[65] Y. Lo, H. M. Kang, M. R. Nelson, M. I. Othman, S. L. Chissoe, M. G. Ehm, G. R. Abecasis, and S. Zöllner. Comparing variant calling algorithms for target-exon sequencing in a large sample. *BMC Bioinformatics*, 16(1):75, 2015.

[66] D. MacLean, J. D. Jones, and D. J. Studholme. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4):287–296, 2009.

[67] J. Majewski, J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado. What can exome sequencing do for you? *J. Med. Genet.*, 48:580–589, 2011.

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

[69] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[70] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.*, 11:499–511, 2010.

[71] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

[72] E. R. Mardis. Next-generation dna sequencing methods. *Annu.Rev.Genomics Hum.Genet.*, 9:387–402, 2008.

[73] E. R. Mardis. Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303, 2013.

[74] C. F. Marrs, L. Zhang, and B. Foxman. *Escherichia coli* mediated urinary tract infections: are there distinct uropathogenic *E. coli* (UPEC) pathotypes? *FEMS Microbiology Letters*, 252(2):183–190, 2005.

[75] G. T. Marth, F. Yu, A. R. Indap, K. Garimella, S. Gravel, W. F. Leong, C. Tyler-Smith, M. Bainbridge, T. Blackwell, X. Zheng-Bradley, Y. Chen, D. Challis, L. Clarke, E. V. Ball, K. Cibulskis, D. N. Cooper, B. Fulton, C. Hartl, D. Koboldt, D. Muzny, R. Smith, C. Sougnez, C. Stewart, A. Ward, J. Yu, Y. Xue, D. Altshuler, C. D. Bustamante, A. G. Clark, M. Daly, M. DePristo, P. Flicek, S. Gabriel, E. Mardis, A. Palotie, R. Gibbs, and . G. Project. The functional spectrum of low-frequency coding variation. *Genome Biol.*, 12(9):R84–2011–12–9–r84, 2011.

[76] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.*, 20(9):1297–1303, 2010.

[77] M. Metzker. Sequencing technologies - the next generation. 11:31–46, 2010.

[78] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[79] M. A. Mulvey. Adhesion and entry of uropathogenic *Escherichia coli*. *Cell. Microbiol.*, 4(5):257–271, 2002.

[80] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, page gkr344, 2011.

[81] J. P. Nataro and J. B. Kaper. Diarrheagenic escherichia coli. *Clinical microbiology reviews*, 11(1):142–201, 1998.

[82] M. Nelson, M. Ehm, D. Wegmann, P. S. Jean, C. Verzili, J. Shen, Z. Tang, D. Kessner, S. Bacanu, D. Fraser, L. Warren, J. Aponte, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woolard, S. Topp, M. Hall, K. Nangle, G. Abecasis, J. Wang, S. Zöllner, L. Cardon, J. Novembre, J. Whittaker, S. Chissoe, and V. Mooser. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337:100–104, 2012.

[83] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, and D. A. Nickerson. Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, 42(1):30–35, 2010.

[84] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461:272–276, 2009.

[85] M. H. Nicolas-Chanoine, X. Bertrand, and J. Y. Madec. Escherichia coli st131, an intriguing clonal group. *Clinical microbiology reviews*, 27(3):543–574, 2014.

[86] R. Nielsen, J. Paul, A. Albrechtsen, and Y. Song. Genotype and snp calling from next-generation sequencing data. *Nat. Rev. Genet.*, 12:443–451, 2011.

[87] T. M. W. Nye, P. Lió, and W. R. Gilks. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22:117–119, 2005.

[88] H. Ochman, S. Elwyn, and N. A. Moran. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12638–12643, 1999.

[89] T. A. Oelschlaeger, U. Dobrindt, and J. Hacker. Virulence factors of uropathogens. *Curr. Opin. Urol.*, 12(1):33–38, 2002.

[90] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*, 5(3):28, 2013.

[91] R. K. Patel and M. Jain. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*, 7(2):e30619, 2012.

[92] E. Pennisi. The human genome. *Science*, 291(5507):1177, 2001.

[93] M. Pérez-Enciso and L. Ferretti. Massive parallel sequencing in animal genetics: wherefroms and wheretos. *Animal Genetics*, 41(6):561–569, 2010.

[94] B. A. Peters, B. G. Kermani, A. B. Sparks, O. Alferov, P. Hong, A. Alexeev, Y. Jiang, F. Dahl, Y. T. Tang, J. Haas, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406):190–195, 2012.

[95] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137, 2001.

[96] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, and C. de Filippo. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

[97] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. D. Bakker, and M. J. Daly. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, 2007.

[98] A. Ratan, W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One*, 8(2):e55089, 2013.

[99] M. Reppell, M. Boehnke, and S. Zöllner. The impact of accelerating faster than exponential population growth on genetic variation. *Genetics*, 196(3):819–828, 2014.

[100] K. Robasky, N. E. Lewis, and G. M. Church. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62, 2014.

[101] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[102] S. J. Salipante, D. J. Roach, J. O. Kitzman, M. W. Snyder, B. Stackhouse, S. M. Butler-Wu, C. Lee, B. T. Cookson, and J. Shendure. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome research*, pages gr–180190, 2014.

[103] L. Salmela and J. Schroder. Correcting errors in short reads by multiple alignments. *Bioinformatics (Oxford, England)*, 27(11):1455–1461, 2011.

[104] A. Sboner, X. J. Mu, D. Greenbaum, R. K. Auerbach, M. B. Gerstein, et al. The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125, 2011.

[105] V. M. Schaibley, M. Zawistowski, D. Wegmann, M. G. Ehm, M. R. Nelson, P. L. S. Jean, G. R. Abecasis, J. Novembre, S. Zöllner, and J. Z. Li. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.*, 23(12):1974–1984, 2013.

[106] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nature*, 200(8):16–18, 2007.

[107] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[108] R. R. Spurbeck, P. C. D. Jr, S. T. Walk, A. E. Stapleton, T. M. Hooton, L. K. Nolan, K. S. Kim, J. R. Johnson, and H. L. Mobley. *Escherichia coli* isolates that carry *vat, fyuA, chuA, and yfcV* efficiently colonize the urinary tract. *Infection and immunity*, 80(12):4115–4122, 2012.

[109] R. R. Spurbeck and H. L. Mobley. Uropathogenic escherichia coli. *Escherichia coli: Pathotypes and Principles of Pathogenesis*, page 275, 2013.

[110] I. Surakka, M. Horikoshi, R. Mägi, A.-P. Sarin, A. Mahajan, V. Lagou, L. Marullo, T. Ferreira, B. Miraglio, S. Timonen, et al. The impact of low-frequency and rare variants on lipid levels. *Nature genetics*, 2015.

[111] M. Tarchouna, A. Ferjani, W. Ben-Selma, and J. Boukadida. Distribution of uropathogenic virulence genes in *Escherichia coli* isolated from patients with urinary tract infection. *International Journal of Infectious Diseases*, 17(6):e450–e453, 2013.

[112] M. A. Taub, H. C. Bravo, and R. A. Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med*, 2(12):87, 2010.

[113] P. N. Taylor, E. Porcu, S. Chew, P. J. Campbell, M. Traglia, S. J. Brown, B. H. Mullin, H. A. Shihab, J. Min, K. Walter, et al. Whole-genome sequence-based analysis of thyroid function. *Nature communications*, 6, 2015.

[114] J. A. Tennessen, A. W. Bigham, T. D. O?Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, B. GO, S. GO, and on behalf of the NHLBI Exome Sequencing Proj. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337:64–69, 2012.

[115] J. Terr and J. Mullikin. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.*, 19 (R2):R145–151, 2010.

[116] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[117] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[118] M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. L. Bouguénec, M. Lescat, S. Mangenot, V. Martinez-Jéhanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Médigue, E. P. C. Rocha, and E. Denamur. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.*, 5(1):e1000344, 2009.

[119] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

[120] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[121] S. I. Vrieze, S. M. Malone, U. Vaidyanathan, A. Kwong, H. M. Kang, X. Zhan, M. Flickinger, D. Irons, G. Jun, A. E. Locke, et al. In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes. *Psychophysiology*, 51(12):1309–1320, 2014.

[122] R. Wan, V. N. Anh, and K. Asai. Transformations for the compression of FASTQ quality scores of next-generation sequencing data. *Bioinformatics*, 28(5):628–635, 2012.

[123] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164–e164, 2010.

[124] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani. Estimation of sequencing error rates in short reads. *BMC bioinformatics*, 13(1):185, 2012.

[125] Y. Wang, J. Lu, J. Yu, R. A. Gibbs, and F. Yu. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population ngs data. *Genome Res.*, 23(5):833–842, 2013.

[126] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7(2):256–276, 1975.

[127] S. Wielgoss, J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Medigue, R. E. Lenski, and D. Schneider. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3 (Bethesda, Md.)*, 1(3):183–186, 2011.

[128] T. J. Wiles, R. R. Kulesus, and M. A. Mulvey. Origins and virulence mechanisms of uropathogenic *Escherichia coli*. *Exp. Mol. Pathol.*, 85(1):11–19, 2008.

[129] T. Wirth, D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. J. Maiden, H. Ochman, and M. Achtman. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.*, 60(5):1136–1151, 2006.

[130] X. Yang, S. P. Chockalingam, and S. Aluru. A survey of error-correction methods for next-generation sequencing. *Briefings in bioinformatics*, 14(1):56–66, 2013.

[131] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17(1):32–43, 2000.

[132] B. C. Young, T. Golubchik, E. M. Batty, R. Fung, H. Larner-Svensson, A. A. Votintseva, R. R. Miller, H. Godwin, K. Knox, R. G. Everitt, Z. Iqbal, A. J. Rimmer, M. Cule, C. L. Ip, X. Didelot, R. M. Harding, P. Donnelly, T. E. Peto, D. W. Crook, R. Bowden, and D. J. Wilson. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl. Acad. Sci. U. S. A.*, 109(12):4550–4555, 2012.

[133] X. Yu and S. Sun. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, 14:274–2105–14–274, 2013.

[134] Y. W. Yu, D. Yorukoglu, J. Peng, and B. Berger. Quality score compression improves genotyping accuracy. *Nature biotechnology*, 33(3):240–243, 2015.

[135] K. W. Yun, H. Y. Kim, H. K. Park, W. Kim, and I. S. Lim. Virulence factors of uropathogenic *Escherichia coli* of urinary tract infections and asymptomatic bacteriuria in children. *Journal of Microbiology, Immunology and Infection*, 47(6):455–461, 2014.

[136] X. Zhan, D. E. Larson, C. Wang, D. C. Koboldt, Y. V. Sergeev, R. S. Fulton, L. L. Fulton, C. C. Fronick, K. E. Branham, J. Bragg-Gresham, G. Jun, Y. Hu, H. M. Kang, D. Liu, M. Othman, M. Brooks, R. Ratnapriya, A. Boleda, F. Grassmann, C. von Strachwitz, L. M. Olson, G. H. S. Buitendijk, A. Hofman, C. M. van Duijn, V. Cipriani, A. T. Moore, H. Shahid,

Y. Jiang, Y. P. Conley, D. J. Morgan, I. K. Kim, M. P. Johnson, S. Cantsilieris, A. J. Richardson, R. H. Guymer, H. Luo, H. Ouyang, C. Licht, F. G. Pluthero, M. M. Zhang, K. Zhang, P. N. Baird, J. Blangero, M. L. Klein, L. A. Farrer, M. M. DeAngelis, D. E. Weeks, M. B. Gorin, J. R. W. Yates, C. C. W. Klaver, M. A. Pericak-Vance, J. L. Haines, B. H. F. Weber, R. K. Wilson, J. R. Heckenlively, E. Y. Chew, D. Stambolian, E. R. Mardis, A. Swaroop, and G. Abecasis. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat. Genet.*, 45:1375–1379, 2013.

[137] L. Zhang and B. Foxman. Molecular epidemiology of *Eschichia coli* mediated urinary tract infections. *Front. Biosci.*, 8:e235–244, 2003.

[138] L. Zhang, B. Foxman, and C. Marrs. Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group b2. *J. Clin. Microbiol.*, 40:3951–3955, 2002.

[139] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.