Leveraging Genetic Variants for Rapid and Robust Upstream Analysis of Massive Sequence Data

by

Fan Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Hyun Min Kang, Chair
Professor Gonçalo Abecasis
Professor Margit Burmeister
Associate Professor Ryan Mills
Professor Kerby Shedden

Fan Zhang

fanzhang@umich.edu

ORCID iD: 0000-0002-6802-4514

ACKNOWLEDGMENTS

It has been a great journey since August 21st, 2013. I feel honored to have the chance to work with a lot of great minds at Michigan. First and foremost, I would like to express my sincere gratitude to Prof. Hyun Min Kang for his confidence in me from the beginning and the guidance he has offered all the way along. His wisdom and diligence towards research have set an excellent example for me. His patient and kind personality towards students have supported my graduate study. I also must express my appreciation to Prof. Gonçalo Abecasis for giving me the chance to work with him and generously supporting me for the first two years, which played an essential part in making all this journey happen. In addition, I am also very grateful to the help from Prof. Ryan Mills since the first day of my rotation in his lab and to his valuable suggestions from time to time. As an international student, I also wish to thank Prof. Margit Burmeister for her kindness and hospitality, not only because of the Chinese Festival event she hosted but more importantly the welcome I have felt at DCMB since the first day I arrived. Additionally, I would like to say thanks to all my dissertation committee for the essential guidance and feedback on my academic development during the last several years: Prof. Gonçalo Abecasis, Prof. Margit Burmeister, Prof. Hyun Min Kang, Prof. Ryan Mills, and Prof. Kerby Shedden.

I also would like to thank my friends in Ann Arbor for experiencing such an incredible journey with me. I will remember all the joy and laughter we created together.

What's more, I'd like to thank my friends that I knew since BGI. They witnessed the ignition of my pursuit in research, they went through bright and dark hours with me, and they supported me throughout my entire 20s.

Finally, my gratitude should go to my family and my parents, who have been inspiring and supportive through all these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

The rapidly increasing throughput of sequencing technologies allows us to sequence genomes, transcriptomes, and epigenomes at an unprecedented scale. Robust, efficient, and accurate computational methods to analyze sequence reads are crucial for successful large-scale studies. In this dissertation, I address specific computational and statistical challenges in quality assessment of sequence reads, ancestry-agnostic estimation of DNA sample contamination, and deconvolution of genetically multiplexed scRNA-seq sequence data by leveraging genetic variants.

In Chapter 2, I describe rapid and accurate algorithms to produce comprehensive quality metrics directly from raw sequence reads without the requirement of full sequence alignment. To produce a comprehensive set of quality metrics such as GC bias metrics, insert size distribution, contamination rates, and genetic ancestry, existing quality assessment methods usually require full sequence alignment which is the most time-consuming step. My methods offer orders of magnitude faster turnaround time by eliminating this requirement when compared to the widely used 1000 Genomes QC pipeline. The results show that the quality metrics estimated from my methods are highly concordant to full-alignment based methods.

In Chapter 3, I present a robust statistical method that accurately estimates DNA contamination agnostic to genetic ancestry of the intended or contaminating samples. Through experiments with *in-silico* contaminated and real sequence datasets, I demonstrate that existing methods may fail to screen highly contaminated samples at a stringent contamination threshold

due to the bias when the genetic ancestry is misspecified. Meanwhile, in the presence of contamination, genetic ancestry estimates can be substantially biased if contamination is ignored. My method integrates genetic ancestry and DNA contamination into a mixture model by leveraging individual-specific allele-frequencies projected from reference genotypes onto principal component coordinates. I show that my method robustly corrects for the bias in both estimates of contamination rate and genetic ancestry under various scenarios of contamination.

In Chapter 4, I enable genetic multiplexing of single-cell RNA-seq (scRNA-seq) experiment without requiring external genotyping by developing genotyping-free scRNA-seq deconvolution method, *freemuxlet*. Genetic multiplexing of scRNA-seq (mux-seq) allows us to cost-effectively sequence single cell transcriptomes across multiple samples in a single library preparation by harnessing natural genetic variations while dramatically reducing the batch effect. However, the existing statistical method, *demuxlet*, which enables mux-seq, requires external genotypes to be collected *a priori*, limiting its applications when it is difficult to obtain high-quality genotypes such as in model organisms or cancer cells. Furthermore, the additional steps to obtain, process, and impute the external genotypes become a substantial bottleneck to analyze the data within rapid turnaround time. *Freemuxlet* defines the distances between a pair of cell barcodes as Bayes Factors (BF) to determine statistical confidence between possible hypotheses of genetic identities of each cell barcodes. The iterative procedure of multi-class clustering guided by BF distances simultaneously estimates the consensus genotypes of each individual while detecting multiplets and deconvoluting the sample provenances of singlets. I apply *freemuxlet* to real datasets and demonstrate high concordance of estimated droplets identities with other methods (cell hashing, *demuxlet*). I further demonstrate that *freemuxlet* can enable mux-seq on cancer cell line mixtures, where *demuxlet* could not due to the difficulty of

accurately genotyping. My results suggest that *freemuxlet* can deconvolute mux-seq experiment as accurate as methods that utilize external information, facilitating a broader range of applications of population-scale single-cell sequencing.

Chapter I.

Introduction

Overview

Massively parallel sequencing, also known as Next Generation Sequencing, has been one of the most successful biological assay methods in the past ten years. Sequencing technologies provide us with digital snapshots of molecular profiles of cells, such as nuclear DNA sequences, mRNA sequences, or chromatin-accessible DNA sequences. These sequence reads help us estimate various quantities or qualitative states, such as genotypes, gene expression levels, or interaction status between DNA/RNA and other molecules. Various techniques that prepare sequence libraries have further helped enable successful applications of DNA-seq, RNA-seq, ChIP-seq, ATAC-seq, single-cell sequencing, non-invasive prenatal test (NIPT) and more[1,2,3].

More important than the versatility of massively parallel sequencing is the unprecedented high-throughput and ever-decreasing cost of these technologies. Large-scale genomic, transcriptomic, and epigenomic studies are becoming feasible, and have contributed many valuable databases and important findings to the scientific community[4,5]. It is very important to keep increasing the scale of omics studies so that we can not only gain more statistical power to

uncover specific findings but also inspire new questions and new methods to deepen our understanding of life science.

However, unprecedented high-throughput of sequencing technologies also brings us computational and statistical challenges of omics studies at an unprecedented scale. Examples include the 1000 genome project (~2,500 genomes), TOPMed (~150,000 genomes), or UK biobank (~500,000 exomes). These petabyte-scale genomic data challenge us to develop computationally efficient methods to handle upstream data processing, such as quality assessment and alignment of raw sequence reads, as well as downstream analyses focused on answering specific scientific questions.

In this dissertation, I will address various analytic challenges in large-scale omics studies with computationally efficient solutions. While each method is intended to tackle different challenges in analyzing high-throughput sequence reads, their common feature is that they all leverage genetic variants information to enable more rapid, robust, and scalable analysis in several different contexts, including data quality control (QC), contamination rate estimation under heterogeneous genetic ancestry background, as well as genotyping-free de-multiplexing of single-cell RNA-seq.

Background

High-Throughput Sequencing Technologies

High-throughput sequencing (or Next Generation Sequencing) is a collective name used to describe several different technologies that all share the same general massively parallel

sequencing concept. Widely used technologies include Sequencing by Synthesis (Illumina)[6], Sequencing by Ligation (SOLiD sequencing)[7], Ion Semiconductor (Ion Torrent sequencing)[8], and Nanopore Sequencing[9]. Sequencing by Synthesis is by far the most popular mainly due to its cost-efficiency, accuracy, and throughput. Since the release of Illumina's Genome Analyzer I in 2006, sequencing throughput has increased from 0.1 gigabases per run to 6,000 gigabases per run of NovaSeq in 2017, while the cost of sequencing a human genome has dropped faster than Moore's Law. In the scope of this dissertation, I will focus on Sequencing by Synthesis (Illumina HiSeq technologies) in future contents, but most of the principles described in the dissertation should be able to extend to other sequencing technologies.

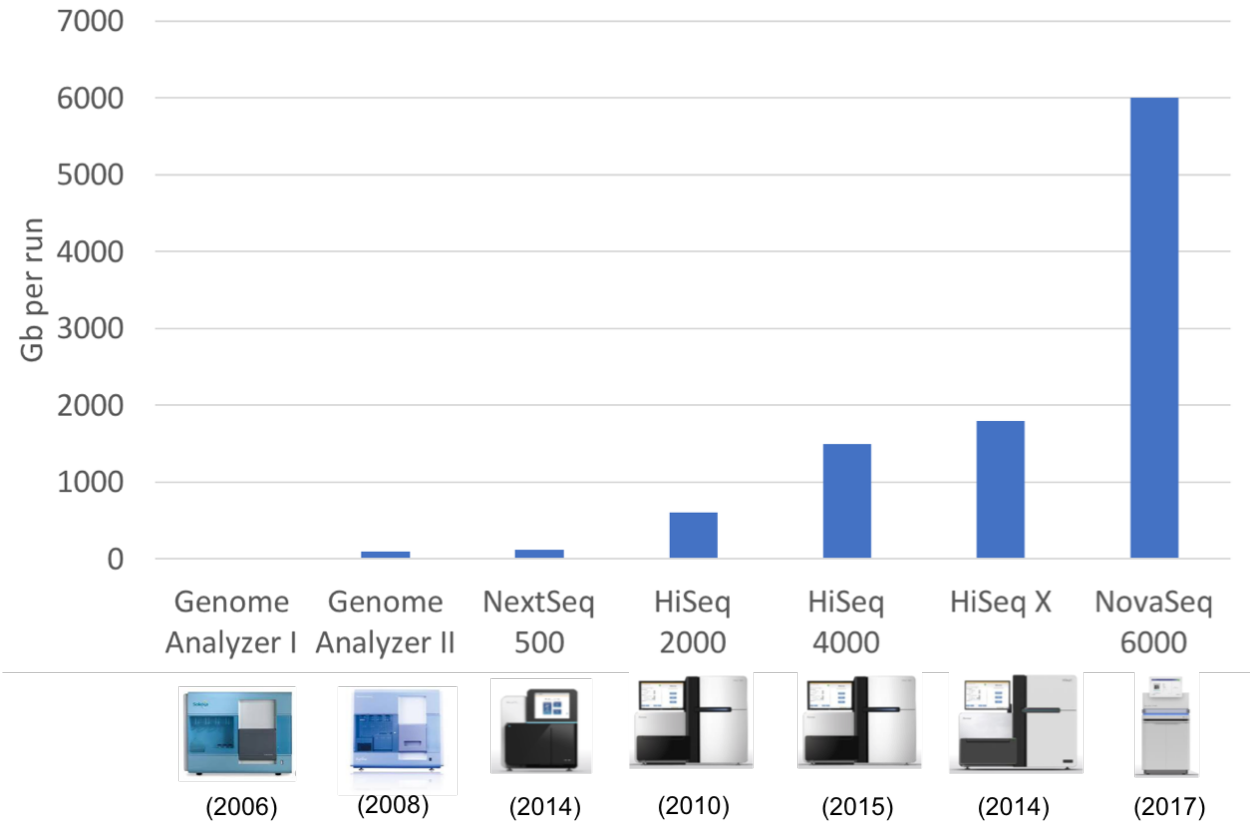

Figure 1.1 Exponential growth in sequencing throughput. This figure reflects the exponential growth of sequencing throughput of sequencers released by Illumina. The X-axis shows typical sequencers released from 2006 to 2017, and Y-axis is sequencing throughput with gigabase per run, in 2006, GA has 0.1Gb/run, in 2017 NovaSeq has 6000Gb/run.

A typical Sequencing by Synthesis procedure starts from a specimen that undergoes sample preparation steps. The procedure continues to library preparation steps that target DNA, mRNA, epigenomic marks, or other molecular features using various molecular technologies (See https://liorpachter.wordpress.com/seq/ for a relatively comprehensive list of examples) to convert them into sequencing libraries. Next, these libraries are physically attached to a flow cell, generating millions of clusters of DNA fragments. The nucleotides of these individual clusters are determined in a massively parallel manner at each cycle, and billions of nucleotide sequence reads are generated simultaneously by typically repeating hundreds of cycles.

Quality Control of Sequence Reads

Ideally, sequence reads should represent unbiased and random samples of actual sequences of target molecules (e.g. DNA or mRNA). However, this assumption may not hold due to technical reasons in practice (e.g. sample degradation or sequencing errors). As a result, the readouts from a sequencing instrument may substantially differ from the actual distribution of sequences of the intended target molecules. Quality control (QC) of sequence reads is a series of steps to evaluate various summary statistics from sequence reads and to identify anomalies that may affect the downstream analysis. Carefully assessing the quality of sequencing data through accurate and comprehensive QCs is the crucial first step to ensure the success of sequencing studies.

From preparing samples, sequencing libraries, running sequencing instruments, to analyzing the digital output from the instrument, there are multiple sophisticated procedures that could be jeopardized by small mistakes or technical errors if not detected by QC. For example, in the sample and library preparation steps, PCR misconfiguration can result in highly biased depth

distribution by GC contents (i.e. GC bias); degraded DNA samples can lead to some portion of the genome to be heavily underrepresented; sample contamination may introduce an excess of heterozygosity in the sequence data. During the sequencing steps, PCR (Polymerase Chain Reaction) enrichment step may result in excessive multi-clonal clusters of identical fragments; Phasing/pre-phasing errors (asynchronized nucleotide binding and cleaving within each cluster) in the sequencing apparatus may result in increased sequencing errors; Failure to ligate indexing tags in multiplexing a single sequence lane may result in apparent contamination due to inaccurate demultiplexing. Ultimately, these errors and biases may result in detection of false variant sites, increased genotyping errors, and false association signals affected by shared technical artifacts within cases and/or controls.

To minimize potential problems in downstream analysis, many tools, such as *FASTQC*[10], *Picard*[11], *QPLOT*[12], *verifyBamID*[13], have been developed to provide quality metrics of the sequence reads. Those metrics include the distribution of base qualities across cycles, depth distribution, GC bias, PCR duplication rate, insert size distribution, contamination rate, and genetic ancestry. These QC metrics aim to detect potential problems that may occur in different sequencing steps. For example, an early decay or sudden drop in base quality in a sequencing cycle may indicate a systematic phasing/pre-phasing error or unexpected event during the run; an unusual depth distribution may indicate coverage bias across the genomic region of interest; strong GC bias may indicate that the excessive technical artifacts may have been introduced during the PCR step; an abnormal insert size distribution may indicate problems in gel electrophoresis, and the sequenced fragments are not concentrated at the expected length.

Quality assessment tools generally can be categorized into pre-alignment and post-alignment tools based on whether all the sequencing reads are required to be fully aligned to the

reference genome, which is the key step that dictates almost the entire computational cost and time of the sequence data processing pipeline. QC metrics such as depth distribution, insert size distribution, and contamination rate requires full alignment while some other metrics, such as base quality distribution, nucleotide compositions do not. Post-alignment methods, such as *QPLOT* and *verifyBamID*, can generate comprehensive QC metrics at the expense of much slower turnaround time, for example, a typical full alignment based QC procedure of a 30x whole genome sequencing dataset takes more than 7 CPU days, whereas pre-alignment methods, such as *FASTQC*, run much faster at the expense of the limited amount of information contained in the QC metrics.

Detection and Estimation of DNA Sample Contamination

DNA sample contamination is one of the most frequently identified problems in the quality control of high-throughput sequencing data. Depending on the contamination source, contamination events can be categorized into cross-species or within-species. Within-species contamination, human to human contamination specifically, is much harder to detect and more common (e.g. due to sample swaps, spillage of the specimen, contamination from other human DNAs during the experimental procedure) because sequence alignment cannot effectively filter out contaminating reads. As a result, within-species contamination will substantially affect genotyping accuracy even for deeply sequenced genomes under modest levels of contamination. For example, genotype discordance increases by 14-fold (0.3% to 4.2%) if the 5-10% of sequence reads are contaminated by another individual compared to uncontaminated sequence reads in a whole exome sequencing study[14].

Methods for detecting and estimating contamination, as well as methods for correcting genotype calls accounting for contamination have been developed previously[13–18]. For example, *verifyBamID*[13] is one of the most popularly used software tools to detect and estimate DNA sample contamination and has been adopted as a part of standard analysis pipeline in most large-scale sequencing centers in the US, and *cleanCall*[14] can correct for DNA contamination in genotype calling.

Estimation of Genetic Ancestry from Sequence Reads

Genetic ancestry plays an important role in many statistical models of genetic analysis such as adjustment for population stratification in large scale genetic association studies. Failing to specify the correct genetic ancestry could result in false association signals or misleading population genetic inferences. Genetic ancestry also plays an important role in the quality control of sequence reads. For example, accurate estimation of DNA contamination requires genetic ancestry information to obtain the correct population allele frequencies[13]. Per-individual variant count or heterozygosity substantially varies by ancestries[4], and quality control steps need to account for ancestries when using such metrics.

In large-scale genetics studies consisting of individuals from a diverse genetic background, it is not uncommon that self-reported ancestry information is incorrect or even unavailable. Inferring genetic ancestry early on in the sequencing analysis can facilitate timely quality control to identify potential sample swaps and also allow QC pipelines to use correct parameters of genetic ancestry when needed.

Many methods, such as *EIGENSTRAT*[19] or *TRACE*[20], have been proposed for estimation of genetic ancestry from array-based genotypes. However, only few methods are currently

available for estimating genetic ancestry from sequence reads. For example, *LASER*[21] enables direct estimation of genetic ancestry of the sequenced sample from sequence reads. The method simulates sequence reads from genotyped reference samples and estimates the principal components of the reference panel and the target sample from the simulated and actual sequence reads. Finally, it projects the PC coordinates of the target sample onto those of the reference panel using *Procrustes* method[22].

Single-cell RNA Sequencing Technologies

The advent of single-cell RNA sequencing(scRNA-seq) technologies allows us to study the impact of environmental and/or genetic perturbations on transcriptomic profiles at a single-cell resolution. Compared to bulk sequencing of mRNAs, scRNA-seq allows us to examine transcriptomic profiles of each cell type, to unravel heterogeneity of cells within the same cell types, and to trace the lineage of cells under developmental or mutational trajectories. For example, in a colorectal cancer single cell study[23], heterogeneous subgroups of cells were identified from a known cell-type identified from bulk sequencing, expanding our understanding intra-tumoral heterogeneity in colorectal cancer and its impact on patient survival and identifying novel differentially expressed genes between tumor and normal cells. Many other examples suggest that scRNA-seq can help us to address various scientific questions that bulk RNA sequencing alone could not address.

The scale of scRNA-seq in studies has been exponentially increasing over the past decade. For example, in 2009, the less than a hundred single cells can be examined in one study, but it has been up to millions[2,24] in 2019. The exponential growth in the number of cells studied in scRNA-seq researches resulted from both sequencing capacity and innovative library

preparation strategies. Unlike early generation of scRNA-seq technologies required separate library preparation of each single cell, recent droplet-barcoding scRNA-seq emerged as a scalable solution to assay thousands of single-cell transcriptomes in a single library preparation[24,25]. This technique mainly relies on a stochastic procedure of placing a cell and a labeled bead within a droplet through a microfluidic device. The droplets that contain both cells and beads will undergo lysis within the droplet then be further processed to extract cDNA sequence that can be demultiplexed based on the bead barcode.

In droplet-barcoding scRNA-seq, controlling cell flow rate is a key factor to determine the tradeoff between throughput and the rates of multiplets. Increasing cell flow rates will increase the chance that a droplet contains more than one cell (i.e. multiplet), whereas reducing the flow rate may result in excessive empty droplets that do not contain cells. As a result, the throughput of droplet-barcoding scRNA-seq will initially increase as we increase cell flow rate but soon reach the ceiling, as further increment of cell flow rate can result in non-identifiable multiplets, which violates the "one-droplet-one-cell" assumption.

Population-scale Single-cell RNA Sequencing with Genetic Multiplexing

The scRNA-seq technologies provide us with promises to understand the regulatory mechanisms relevant to complex traits at a single cell resolution. To identify significant associations between genetic or environmental factors and expression levels of genes in a specific cell type with sufficient statistical power, it is important to scale single-cell sequencing experiments to as many samples as possible. However, due to the limited throughput of the scRNA-seq experiment, the per-sample cost of scRNA-seq experiment is orders of magnitude

more expensive than bulk RNA-seq, so it is currently prohibitive to scale scRNA-seq to a large number of individuals.

Recently, a cost-effective scRNA-seq experiment strategy, *mux-seq*[26], has emerged to address this challenge. Instead of processing one sample per library preparation, *mux-seq* protocol pools many genetically diverse samples together in a single library preparation with much higher cell flow rate. Due to the increased cell flow rate, the number of cells per run increases by orders of magnitude, at the expense of a high fraction of multiplets. Because most of these multiplets contain cells from different individuals, they can be identified statistically by leveraging "genetic barcodes" that are encoded in the scRNA-seq reads overlapping with genetic variants. The singlets can also be demultiplexed into the originating samples using the genetic variants. Similar strategies can be applied for samples with the same genetic background if each sample is uniquely barcoded with additional molecular assays (such as Cell Hashing[27], or MULTI-seq[28]). These multiplexing-based scRNA-seq techniques are rapidly becoming popular due to the ability to substantially reduce the cost while dramatically eliminating batch effects by harnessing natural genetic variations or additional molecular tags unique to each sample.

Challenges

There are numerous computational and statistical challenges remaining for achieving accurate and efficient analysis high-throughput sequence data in various types of scientific studies. Compared to downstream analyses that take relatively a small snapshot of the massive amount of data to answer specific biological questions, relatively fewer methods for upstream analysis, which requires efficient processing of massive amount of sequence reads, have been

developed so far. In this dissertation, I address the three specific challenges related to the rapid and robust analysis of the massive amount of sequence data in various contexts of population-scale genomic and transcriptomic studies.

Rapid, Comprehensive, and Accurate Quality Control of Ultra-High-Throughput Sequence Reads

With the advent of ultra-high-throughput sequencing technologies, a rapid turnaround time of quality control is becoming increasingly important. Delay or failure in detecting potential problems in the sequencing or library preparation protocol can result in massive financial loss. For example, the NHLBI TOPMed project has sequenced 50,000 genomes a year, which costs $1M/week in sequencing at $1,000/genome. Delayed quality control more than a week may cause loss of multimillion dollars, not to mention the waste of time and, more importantly, the waste of valuable samples.

The challenge with existing QC methods is the tradeoff between the turnaround time and the capabilities to generate comprehensive QC metrics. The pre-alignment methods are fast but unable to generate important QC metrics such as insert size distribution or contamination rates; on the other hand, the post-alignment methods can generate comprehensive metrics, but the turnaround time is orders of magnitude longer. To obtain QC metrics within rapid turnaround time, we need to avoid the full alignment of sequence reads to the reference genome. However, estimation of many QC metrics require aligned positions of reads or their genomic context, so at least certain informative reads need to be aligned to the genome to obtain comprehensive QC metrics.

In my dissertation, I address these challenges through a rapid algorithm that focuses only on selected regions of the genome that allows us to estimate comprehensive QC metrics, while filtering most of the reads outside the focused region to reduce the turnaround time by orders of magnitudes compared to the existing post-alignment QC methods.

Robust Estimation of DNA Contamination and Genetic Ancestries

One of the most widely used methods to estimate DNA contamination from sequence genome is *verifyBamID*[13]. In practice, it works well in various scenarios of DNA contamination, but it requires users to correctly specify population allele frequencies for each variant. When incorrect population allele frequencies are specified, it is known to give a biased estimation of contamination rates[13]. When a large number of samples from diverse genetic ancestries are sequenced, such as in major sequencing centers, accurately estimating DNA contamination regardless of genetic ancestry is not impossible, but practically challenging.

First, while it is possible in principle to obtain the genetic ancestry information of each sequenced sample, prepare the population allele frequencies of each sequenced sample, and run *verifyBamID* with sample-specific parameters, it is very cumbersome or even infeasible to implement such a pipeline in practice for large-scale sequencing centers or studies. Second, self-reported ancestry could be incorrect or not available *a priori*, and it is unclear how to best prepare the population allele frequencies for the sample in such cases. Third, due to the practical difficulties to implement the best practice (i.e. specifying correct population allele frequencies for each sample), many investigators run *verifyBamID* with the default setting, using pooled allele frequencies across diverse populations in 1000G. However, this common practice

introduces population-specific biases in the contamination estimates, making certain populations or ethnic groups more prone to an incomplete screening of DNA contamination.

Moreover, estimating genetic ancestry from sequence reads can inform many important details about the sequenced samples in the quality control perspectives. The current methods to estimate genetic ancestry from sequence reads, such as *LASER*[21], usually assume the sequenced genome is free of contamination. The genetic ancestry estimates can be biased if this assumption does not hold. To address these issues in estimating DNA contamination and genetic ancestry, I develop a more robust approach that jointly accounts for contamination and genetic ancestry together.

Population-scale Single-cell Sequencing Experiments without External Genotyping

While genetic multiplexing (*mux-seq*) workflow offers a population-scale solution for cost-effective scRNA-seq experiment with reduced batch effects, the existing statistical method, *demuxlet*[26] requires external genotypes to be collected by array genotyping or DNA sequencing. If it is possible to implement a genotyping-free version of *demuxlet*, it will enable much broader applications of *mux-seq* workflow in cases like model organisms or cancer cells, where genotype information is not easy to obtain. Moreover, the genotyping-free *mux-seq* workflow will simplify *mux-seq* experiment and analysis workflow by removing the bottleneck of processing and imputing genotype data, which requires substantial time and effort in addition to the cost of external genotyping.

To overcome this limitation, in my dissertation I propose a genotyping-free de-multiplexing method based on the fact that a large fraction of RNA sequence read itself already contains allelic information of known variants. Unlike the scenario with external genotypes,

where demultiplexing is essentially a supervised classification problem; de-multiplexing without external genotypes becomes an unsupervised clustering problem, where sparsity of available reads overlapping with genetic variants makes it a much more difficult problem than its supervised version.

Chapter Overview

My dissertation topics focus on accurate, robust, and efficient methods for upstream analyses of massive amount of sequenced genomes and single-cell transcriptomes. The goals of these topics range from enabling rapid and accurate quality control of sequence data to enabling more cost-effective and seamless population-scale single-cell sequencing experiments. The relationships among these chapters can be summarized in Figure 1.2.

In Chapter 2, I will introduce methods for ultra-fast quality control of ultra-high throughput sequence reads. By focusing on a subset of a reference genome around a specific set of common genetic variants, I integrate the advantages of rapidly filtering out negative hits from a spaced k-mer hash table and detailed information from a BWT based local alignment[29]. My methods first extract flanking sequences (250-1,000bp) around known genetic variant sites to construct a reduced reference genome and rapidly filters out >94% of unalignable reads using the spaced hashing technique. The filtered reads are aligned to the reduced reference genome using a computationally optimized version of *BWA*[29]. Compared to the conventionally used 1000 Genomes alignment pipeline, my method can reduce the computational time to generate thorough quality metrics on a 38x genome from 160 hours to 1hr. The results show that the quality metrics estimated from my methods are highly concordant with the quality metrics

generated from the full-alignment pipeline-based methods. Because my method collects information about genetic variation between individuals, it also provides us with pileup information and performs model-based estimation using genotype likelihoods, including estimation of DNA contamination and genetic ancestry.

In Chapter 3, I will show that the estimation of DNA contamination and genetic ancestry can intertwine and cause bias if not handled properly. Specifically, I propose a robust statistical method, *verifybamID2*[30], that jointly account for DNA contamination and genetic ancestry. Because *verifybamID2* also estimate genetic ancestry, it can accurately estimate DNA contamination without requiring genetic ancestry of the intended or contaminating genomes. My method integrates the estimation of genetic ancestry and DNA contamination in a mixture model based likelihood framework by leveraging individual-specific allele frequencies[31,32] projected from reference genotypes onto principal component coordinates. Based on the evaluation of my method on real datasets, I show that *verifyBamID2* robustly corrects for the bias in both contamination rate estimates and genetic ancestry estimates under various scenarios of DNA contamination.

In Chapter 4, I propose a novel method, *freemuxlet*, to multiplex/demultiplex large scale single cell sequencing samples, which increases multiplexing throughput without using external demultiplexing barcode other than genetic variants information within droplet itself. The key idea is to aggregate droplets that belong to the same sample into clusters, and to calculate the genotype likelihood of each cluster. With genotype information inferred from scRNA-seq reads, similar to *demuxlet*, it is possible to detect and remove doublets, to assign membership to each droplet. *Freemuxlet* initiates this clustering process by first assigning most likely singlets to clusters, then continue to iteratively update cluster genotype likelihood and droplet membership.

Freemuxlet cluster droplets based on the pairwise distance described by the Bayes Factor (BF) that is defined by the hypotheses whether droplets have the same sample origin or not. Clusters are further refined by detecting and removing doublets using mixture model assuming equal proportions. The results show that *freemuxlet* produces highly consistent results to the independent cell-hashing method[27] based on additional molecular tags and to the *demuxlet* based on external genotypes.

In Chapter 5, as a conclusion, I summarize the three main chapters, discuss the strength and weakness of these methods with possible future directions of these methods and the related fields.

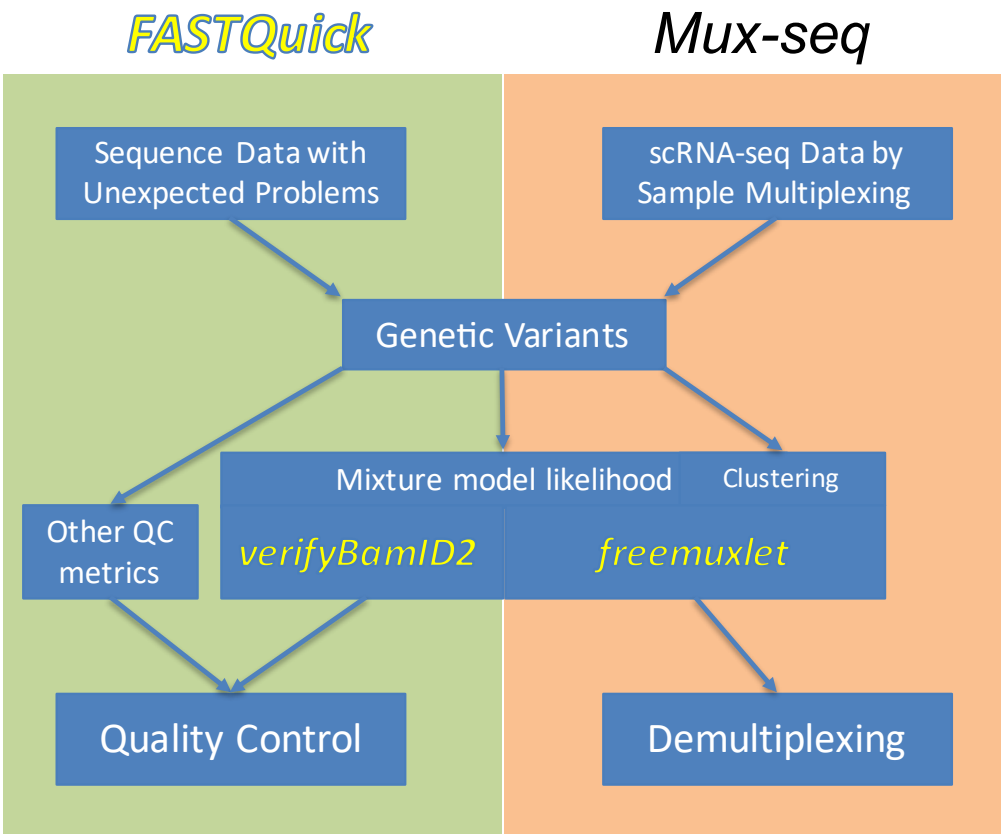

Figure 1.2 Illustration of the relationship among chapters. Dissertation chapters address challenges that arise from different scenarios of sequence analyses but share common key factors. 1. Both *FASTQuick* and *Mux-seq* leverage genetic variants to collect allelic information on sequences to calculate QC metrics or genotype likelihood. 2. Both *verifyBamID2* and *freemuxlet* applied mixture model based likelihood to deconvolute mixed components in sequencing dataset. (contamination for verifyBamID2 or multiplexed sequence for *freemuxlet*.)

Chapter II.

FASTQuick: Rapid and Comprehensive Quality Assessment Tool from Raw Sequence Reads

Introduction

Efficient and thorough quality assessment from deeply sequenced genomes in ultra-high-throughput scale is crucial for successful large-scale sequencing studies. Delay or failure in detecting contamination, sample swaps, quality degradation, or other unexpected problems in the sequencing or library preparation protocol can result in enormous loss of time, money, and invaluable specimens if, for example, hundreds or thousands of samples are found to be contaminated weeks or months later. Ensuring comprehensive quality control of sequence data at real-time speed will assure generation of high-quality sequence reads, and subsequently successful outcomes in the downstream analyses.

Existing quality assessment tools mainly fall into two categories – pre-alignment and post-alignment methods – based on whether they require full alignment of the genome prior to the quality assessment. Pre-alignment methods, such as *FASTQC*[10], *PIQA*[33], and *HTQC*[34], produce read-level summary statistics that can be obtained from sequence reads, such as base

compositions, k-mer distributions, base qualities, and GC bias levels. However, these pre-alignment methods do not estimate many key quality metrics required for comprehensive quality assessment. These missing metrics include mapping rate, depth distribution, fraction of genome covered, sample contamination, or genetic ancestry information. Other post-alignment methods, such as *QPLOT*[12], *Picard*[35], *GotCloud*[36], and *verifyBamID*[13], provide a subset of these key quality metrics but require full alignment of sequence reads, which typically takes hundreds of CPU hours for deep (e.g. >30x) sequence genome.

We describe *FASTQuick,* a rapid and accurate set of algorithms and software tools, to combine the merits of QC tools from both categories. By focusing on a variant-centric subset of reference genome(reduced reference genome), our methods offer up to >50-fold faster turnaround time than existing post-alignment methods for deeply sequenced genome, while providing a comprehensive set of quality metrics comparable with *QPLOT* and *verifyBamID* with the help of statistical adjustments to account for the reduced reference genome.

Results

FASTQuick Overview

The key algorithms and procedures of *FASTQuick* are illustrated in Figure 2.1, and further details can be found in the Methods section. Briefly, *FASTQuick* constructs a reduced reference genome and indices from a set of flanking sequences surrounding known SNPs, and rapidly filters out unalignable reads and align filtered sequence reads. Three types of generic QC summary statistics – per-base, per-read, and per-variant summary statistics - are generated from

the aligned reads to be translated into an interpretable and user-friendly quality metrics described in Table 2.1.

Table 2.1 Quality assessment metrics provided by different tools

Metrics	*FASTQC*	*PIQA*	*HTQC*	*QPLOT*	*Picard*	*verifyBamID2*	*FASTQuick*
Base Quality Per Cycle	✓	✓	✓	✓	✓		✓
GC Bias				✓	✓		✓
PCR Duplication Rate				✓	✓		✓
Insert Size Distribution				✓	✓		✓
Contamination Estimate					✓	✓	✓
Genetic Ancestry						✓	✓
% Mapped Reads				✓	✓		
Depth Distribution				✓	✓		✓
Total Number of Reads	✓			✓	✓		✓
K-mer Distribution	✓						
Read Length Distribution	✓		✓	✓	✓		✓
Full-Alignment not Required	✓	✓	✓				✓

Computational Efficiency

The primary goal of *FASTQuick* is to achieve comprehensive QC with much less computational cost than full-alignment-based QC procedures. A large fraction of the computational gains come from the usage of the reduced reference genome and filtering of unalignable reads through mismatch-tolerant spaced k-mer hashing(Figure 2.1A). Compared to alignment to the full human reference genome, aligning a 3x genome on the reduced reference genome reduced the run time by 34.9-fold (94,020 vs. 2,697 seconds) using the same algorithm. Using mismatch-tolerant spaced k-mer filtered out >90% of unalignment reads with no loss of alignable reads when 3 or more hits are required (default parameter) to be considered as alignable reads, saving additional 65% of computational time (Figure 2.1B). Putting them

together, the alignment step of *FASTQuick* (with default parameters) was 100-fold faster (94,020 vs. 939 seconds) than the full genome alignment. We observed that >99% of unalignable reads could be filtered out with a more stringent threshold at the expense of 0.01% loss of alignable reads. However, the additional computational gain was only 14% (939 vs. 811 seconds).

We also evaluated the overall computational efficiency between *FASTQuick* and the *GotCloud*-based QC pipeline on high-coverage genome (38x) and low-coverage (3x) genomes from the 1000 Genomes Project (Table 2.2). The results demonstrate that *FASTQuick* produces a comparable set of QC metrics to *GotCloud* with 30~100-fold faster turnaround time.

Table 2.2 Running time comparison(in hours)

# of Thread	*FASTQuick* Time		*Gotcloud* QC Time(*BWA*)	
	HG00553(3X)	NA12878(38X)	HG00553(3X)	NA12878(38X)
1	1.03	5.48	30.95	369.56
2	0.53	2.46	21.53	230.85
4	0.33	1.76	15.83	154.91
8	0.24	1.75	12.74	131.85

Running time is evaluated as wall clock elapsed time on a machine with Intel(R) Xeon(R) CPU (X7560 @ 2.27GHz)

Accuracy of QC Metrics

We compared the distribution of QC metrics generated from *FASTQuick* with those from *GotCloud* on multiple sequenced genomes. The QC metrics shared between *FASTQuick* and that GotCloud[36] are listed in Table 2.S1. The visualization QC metrics such as base quality recalibration (Figure 2.1E), normalized mean depth by GC content (Figure 2.1F), and depth distribution, were very close between our methods and *GotCloud*. Quantitatively, the QC metrics

were very similar too. For example, the two-sample Kolmogorov-Smirnov (KS) test statistics, which quantifies the maximum differences between two empirical cumulative distribution functions were D = 0.040.

Insert Size Correction with Kaplan-Meier Estimator

One challenge arising from QC based on partial alignment of sequence reads to the reduced reference genome is the estimation of insert size distribution. The insert size distribution is typically estimated from distances between the aligned pairs of reads from the fully aligned reads. When using a reduced reference, a large proportion of paired reads may not be fully mapped, and the read pairs that have shorter insert size are more likely to be mapped in both ends. As a result, estimating insert size distribution based only on the reads where both ends are mapped will result in biased estimates of insert sizes, as empirically demonstrated using the 38x genome in Figure 2.2.

We resolved this challenge first by extending 10% of the variant-centric contigs to be sufficiently long (2000bp), and by estimating insert size only from the reads mapped to longer contigs. This way, we prevent the reduced reference genome from becoming too large to achieve computational efficiency, while substantially reducing the bias of insert size estimation. The observed insert size distribution from the longer 2,000bp contig was closer to that from a full alignment (Figure 2.2). However, bias still exists at a smaller level.

To systematically correct for biased estimation of insert sizes, we statistically integrated the observed insert sizes across all contigs inverse probability weighting based on Kaplan-Meier curve[37] (See Materials and Methods for details). Applying our correction produces estimated insert size distribution much closer to that from the full alignment (Figure 2.1G). The KS-test

statistics were reduced from 0.60 (using 500bp contigs), to 0.18 (adding 2,000bp contigs), and to 0.017 (with Kaplan-Meier adjustment) compared to the estimates from the full alignment.

Estimation of Contamination Rate and Genetic Ancestry

To evaluate the estimation accuracy of contamination rate and genetic ancestry, we prepared artificially *in-silico* mixed 1000g samples as described in the Materials and Methods section. Then we compare the estimated contamination rate and genetic ancestry from *FASTQuick* with the estimation from the full-alignment QC pipeline-based result (Figure 2.1H). Results show that *FASTQuick* can estimate contamination rate and genetic ancestry as accurate as the full-alignment pipeline-based methods.

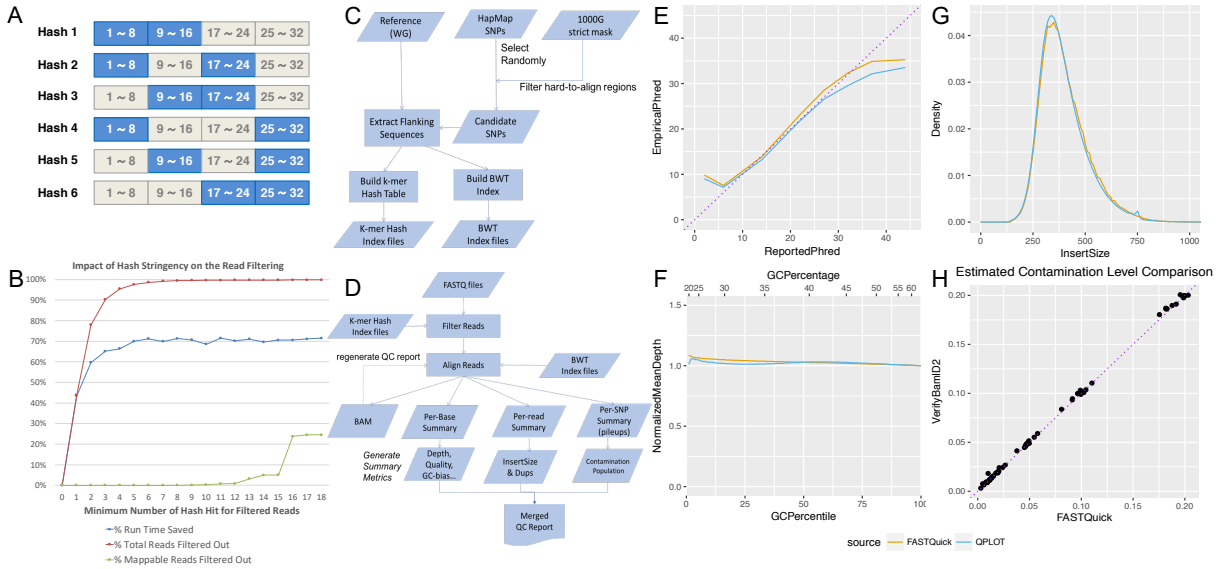

Figure 2.1 Illustration of *FASTQuick*. A) Spaced K-mer Hash filter design with the tolerance of mismatches. **B)** Hash filter efficiency and time saving conditional on different threshold of hash hits. **C)** Reduced reference genome indexing process. We randomly select a subset of variant markers in the HapMap database after filtering out variants in hard-to-align regions. Flanking region sequences of these variants are constructed to form the reduced reference genome. Spaced K-mer Hash and BWT indices are constructed based on the reference genome. **D)** Spaced K-mer Hash filter based sequence alignment process. Spaced K-mer Hash rapidly filter out negative hits of reads while maintaining tolerance of mismatches. Only a small fraction of reads will be aligned using bwa algorithms. Alignments will be analyzed using our methods to recover and report the QC status. **E)** Empirical Phred Score versus Reported Phred score. **F)** Normalized Mean Depth versus GC content. **G)** Insert Size Distribution recovered with Kaplan-Meier estimator. H) Contamination Level Estimation Comparison. Each point represents an artificially mixed 1000g sample with mixing rate ranging from 0.01 to 0.2.

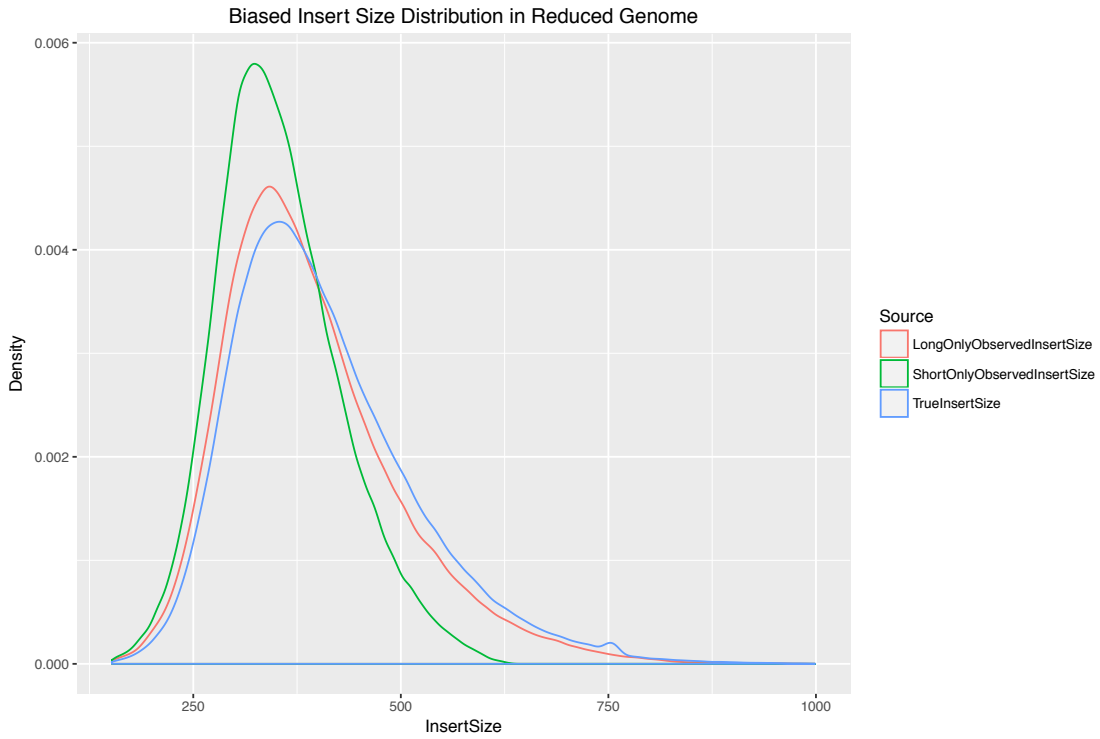

Figure 2.2 Biased insert size distribution in reduced genome under 250bp(short) or 1000bp(long) flanking length configuration. Each color represents one scenario of insert size estimation without correction. "LongOnlyObservedInsertSize" (red) is when insert size distribution estimated only using reads mapped to the long flanking region; "ShortOnlyObservedInsertSize"(green) is when only using reads mapped to the short flanking region; "TrueInsertSize" (blue) is insert size distribution estimated under full genome alignment.

Discussion

Rapidly generating comprehensive QC metrics for ultra-high-throughput sequencing experiments is crucial for all the omics studies. Timely feedback during the early stage of data production can help us detect problems early and avoid further loss. However, rapid QC of ultra-high-throughput sequencing can often be impeded by computational challenges. *FASTQuick* addresses these computational challenges by analyzing sequence reads mappable to an informative subset of the reference genome and by extrapolating observed QC metrics to genome wide-scale with tailored statistical methods. Our results demonstrate that the QC metrics

produced by *FASTQuick* are comprehensive and highly concordant to methods that require full sequence alignment.

Compared to previous quality assessment methods that do not align sequence reads at all, *FASTQuick* provides more comprehensive QC metrics such as depth distribution, insert size distribution, contamination, and genetic ancestry. The key idea behind the rapid alignment includes customized data structures, which combine the speed of the spaced-kmer hash table and the detailed alignment of BWT-based methods. *FASTQuick* rapidly filters out negative hits while generating detailed local alignment for positive hits. Tailored statistical estimation correct biases of the metrics from the reduced reference genome to make the QC metrics more accurate, as demonstrated in the estimation of insert size distribution using the Kaplan-Meier estimator.

The speed of *FASTQuick* is about 50-fold faster than conventional full-alignment-based methods. Interestingly, our computational time is much faster than the time required to convert Illumina's BCL formatted files into FASTQ files (~5 hours). Therefore, our methods can work as a UNIX pipe between the conversion procedures, so that it does not increase the end-to-end wall-clock time if additional CPU cores are available in the machine. This ultra-fast speed can inspire new applications that also require real-time feedback while providing detailed alignment information in target regions.

We acknowledge that *FASTQuick* also has certain limitations. The current version is only suitable for short sequence data. To make it compatible with long sequences, we need to further improve alignment algorithms to absorb features of other methods like minimap[38]. Another issue is that it still relies on a linear reference genome. Compared to linear reference genome-based alignment methods, the variant graph-based method [39] has advantages such as more sensitivity to non-diallelic variants and more balanced reads depth around indel or structural variants, which

characterizes genome-wide diversity better. Extending *FASTQuick* to other types of sequence data, such as exome sequencing, RNA-seq, ChiP-seq, and ATAC-seq should also be possible, if the technology-specific target regions are properly considered and accounted for.

Materials and Methods

Overview of *FASTQuick*

FASTQuick first constructs a reduced reference genome from a set of flanking sequences surrounding known SNPs and build a BWT index[29] and mismatch tolerant k-mer hash table(Figure 2.1C). Once the indices are built, *FASTQuick* rapidly filters out unalignable reads whose first 96-bp have less than 3 hits (out of 18 potential hits, among which 6 hits per 32-mer) against the spaced k-mer hash indices, and align filtered sequence reads to the reduced reference genome using the BWT index (Figure 2.1D). The small fraction of filtered aligned reads will be stored in binary Sequence Alignment/Map format (BAM) [40]. Next, three types of generic QC summary statistics – per-base, per-read, and per-variant summary statistics – are generated from the aligned reads. Per-base summary statistics informs about mapping rate, depth distribution, GC-bias assessment, and base quality assessment. Per-read summary statistics allows us to estimate insert size distribution, with adjustment using inverse probability weighting based on Kaplan Meier curve[37] to account for pair-end alignment bias due to the reduced reference genome, and duplication rate. Per-variant summary statistics allows us to estimate DNA contamination rate and genetic ancestry. Finally, these summary statistics are combined, jointly analyzed, and translated into an interpretable and user-friendly quality report.

Construction of Reduced Reference Genome using Flanking Sequences of SNPs

FASTQuick constructs reduced reference genome based on flanking sequences around known common SNPs to enrich the reads that are informative both for genotype likelihood based inference (e.g. contamination and ancestry) and other quality metrics that require reads alignment. Starting from an arbitrary set of known SNPs, *FASTQuick* randomly selects a designated number of SNPs from known common (MAF>5%) SNP set, such as HapMap3[41], while excluding SNPs near hard-to-align regions (e.g. 1000 genome project strict mask region). *FASTQuick* then constructs reduced reference genome using short flanking sequences of the majority of SNPs(e.g. 90%) and long flanking sequences of the remained SNPs.

Filtering Unalignable Reads with Mismatch-tolerant Hash

Because the reduced reference genome is a small subset of the whole genome sequence, we expect that only a small fraction of reads will be alignable. However, attempting to align all the reads is still computationally expensive. *FASTQuick* builds a hash-based index to rapidly filter out the reads that are unlikely to be aligned to the reduced reference genome. To make the hash robust against sequencing errors, *FASTQuick* builds six locally sensitive hash tables of 16-mers for each 32-mer(Figure 2.1A), so that 32-mers with 2 or less mismatches can still be guaranteed to match to at least one of the hash tables.

FASTQuick partitions each sequence read into multiple 32-mers and performs hash lookups for each possible 16-mers. For example, for a 100-bp read, eighteen 16-mers (6 per 32-mer) across three 32-mer will be matched to the hash table. For reads longer than 96-bp reads, only the first 96-bp reads are used. *FASTQuick* will decide to filter out a read or not based on whether the number of matching 16-mers is less than a certain threshold k. For example, if k is 3,

reads with less than 7 mismatches are guaranteed to pass the filter, and many other reads with more mismatches will pass the filter. If k is 10, reads with less than 3 mismatches are guaranteed to pass the filter. We chose k=3 based on our experiment based on empirical observations described in the Results section.

Generating Base-level and Read-level QC Metrics

Using the reads aligned to the reduced reference genome, *FASTQuick* generates both base-level and read-level QC metrics. Base-level metrics, such as depth distribution and the number of mismatches, are recorded and summarized by GC content, reported base quality, and sequencing cycle. Because the reads spanning the end of flanking sequences may be poorly aligned, *FASTQuick* produces base-level metrics only on the fully alignable portion of flanking sequences. Let the length of flanking sequence be *w* and the read length be *r*. Then, only *2*(w-r) +1* bases spanning the variant site will be considered when collecting base-level summary statistics. Read-level QC metrics, such as the fraction of mapped reads, the fraction of duplicated reads and insert size distribution are estimated and reported based on reads alignment result.

Bias-Corrected Estimation of Insert Size Distribution

Due to the limited length of flanking sequences in the reduced reference, the observed distribution of insert sizes obtained from the reads that both ends are mapped will be biased towards smaller values. In order to recover the full distribution of insert sizes adjusting for the "censored" reads due to large insert sizes beyond the flanking sequences, we adopted the Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average[37] as described below. We define a tuple (t_o, t_l, t_r) for each mapped DNA segment (or read pair), where t_o is

the observed insert size, t_l is the maximal insert size of *read 1*, and t_r is the maximal insert size of *read 2*. The maximal insert size is defined as the distance between the leftmost/rightmost base of *read 1*/*read 2* and the rightmost/leftmost base of the flanking region sequence. This tuple is fully specified only when a read pair is properly aligned, otherwise, for single-end mapped read pair(including partially mapped pair) only one of the two maximal insert sizes (t_l or t_r) is available and unobserved value is set to missing, the rest of the read pairs, such as read pairs that are mapped to different chromosome, with low mapping quality, or in abnormal orientation, are discarded in insert size estimation. Empirically, among all N properly aligned read pairs, we can estimate insert size by counting the frequency of different observed insert sizes, t_o, and the cumulative distribution of insert size hence becomes:

$$F(t) = \frac{1}{N}\sum_{i=1}^{N} I\left[t_{o,i} \leq t\right]$$

However, this direct estimation will be severely biased because of reads mapped only in a single end is more likely to have larger insert sizes.

To correct this bias, analogous to estimation of survival function, which satisfy $S(t) = 1 - F(t)$, we can view the leftmost/rightmost base on each flanking region as the start time point, the exact insert size t_o as the time when data point is observed to fail, and the maximal insert size, t_l and t_r, as the time when data point is censored. Let the ordered observed time points t_o and censored time points t_l (or t_r) be τ. Denote o_t as the number of observed failure cases, i.e. the number of read pairs have observed insert size less than or equal to t, and also denote c_t as the number of censored cases at time *t*, i.e. the number of single-end mapped read pairs have maximal insert size less than or equal to t, then let $I\left[\tau_j \geq t\right]$ be indicator function if j-th time point larger than certain time t(j-th insert size larger or equal to t). Then the risk set is:

$$Y(t) = \sum_{j=1}^{J} (o_j + c_j) I[\tau_j \geq t]$$

Then the Kaplan-Meier estimator $\widehat{S_{km}}$ of $S(t)$:

$$\widehat{S_{km}}(t) = \prod_{\{j|\tau_j \leq t\}} \left(1 - \frac{n_j}{Y(\tau_j)}\right)$$

Satten *et al.*(2001)[37] proposed a simplified algorithm to iteratively estimate survival function for failure times and survival functions for censoring times, by which we conveniently estimate $F(t)$.

Estimation of Contamination Rates and Genetic Ancestry

We also implemented the likelihood model based methods to estimate genetic ancestry and contamination rate in *FASTQuick*. The details of these methods will be fully described in Chapter 3. In *FASTQuick*, to seamlessly integrated these methods into our ultra-fast QC procedure, we designed compatible variant-centric data structures and input/output interfaces that can directly deliver sequence information and estimated statistics from *FASTQuick* to modules that estimate contamination and genetic ancestry.

Experimental Data

We selected a deeply sequenced genome of a publicly available sample (NA12878) from the Trans-Omics Precision Medicine (TOPMed) project for most evaluations. To evaluate computational efficiency for low-pass sequence genome, we also evaluated another sample

(HG00553) from the 1000 Genomes Project (ERR013170, ERR015764, and ERR018525). To evaluate the accuracy of contamination estimates we constructed 10 genomes with *in-silico* contamination by randomly sampling aligned sequence reads from samples in 1000 Genomes phase 3 project and then mixing reads from different samples proportional to the intended contamination rates $\alpha \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, as described in Chapter 3.

Chapter III.

Ancestry-agnostic Estimation of DNA Sample Contamination from Sequence Reads*

Introduction

Sample contamination is a common problem in DNA sequencing studies. Contamination may occur during sample shipment (due to spillage across wells, pipetting errors or insufficient dry ice), library preparation (due to gel cut-through in fragment size selection or unexpected switch between barcoded adaptors *in-vitro*), *in-silico* demultiplexing from a sequenced lane into barcoded samples, or on many other unexpected occasions. Even modest levels of contamination (e.g. 2-5%) within a species substantially increase genotyping error, even for deeply sequenced genomes[42]. Accurate estimation of DNA contamination rates allow us to identify and exclude contaminated samples from downstream analysis, and genotypes of moderately contaminated samples (e.g. <10%) can be improved by accounting for contamination in genotype calling[42].

---

* This chapter has been accepted by *Genome Research* as Zhang, F., Flickinger, M., Abecasis, G., Boehnke, M. & Kang, H. M. Ancestry-agnostic estimation of DNA sample contamination from sequence reads.

Previously we developed methods and a software tool, *verifyBamID*[43], to estimate DNA contamination from sequence reads given known population allele frequencies of common variants. Many investigators and most major sequencing centers use *verifyBamID* as a part of their standard sequence processing pipeline. However, we have shown that *verifyBamID* can underestimate DNA contamination rates if the assumed population allele frequencies are inaccurate[43]. Such an underestimation can be avoided if correct population allele frequencies are provided in ideal circumstances. However, in early stages of sequence analysis, performing a tailored customization of quality control (QC) steps for each sequenced genome based on their ancestry is not always feasible or is sometimes impossible. Such a tailored customization requires planned coordination between sequencing centers and study investigators prior to sequencing to share the self-reported ancestry (which is not always accurate) or estimated ancestry from external genotypes (which is not always available). Modifying the QC pipeline to accommodate study-specific or sample-specific parameters may not be an option for large sequencing centers. Even if such a tailored customization of QC pipeline is possible, preparing per-sample ancestry prior to QC may delay time-sensitive issues in the sequencing procedure. If contamination rates can be accurately estimated without having to know the ancestry or allele frequencies a priori this will simplify the sequence analysis pipeline and expedite the QC.

Here we describe a novel method to robustly detect and estimate DNA contamination by modelling the probability of observed sequence reads as a function of "individual-specific allele frequencies" that account for genetic ancestry of a sample. Instead of assuming that the population allele frequencies are known, we represent individual-specific allele frequencies as a function of genetic ancestry using principal component coordinates and the reference genotypes from a diverse population, e.g. Human Genome Diversity Project (HGDP)[44] or 1000 Genomes[45].

We then jointly estimate genetic ancestry and contamination rates of a sequenced individual based on a mixture model, without requiring the assumption that population allele frequencies are known. As a result, our method enables robust estimation of DNA sample contamination without relying on externally provided genetic ancestry information. Instead, our method simultaneously estimates the genetic ancestry accurately from sequence reads through a unified likelihood framework.

Results

Our previous method (*verifyBamID*) can estimate sample contamination rate with external genotypes or with population allele frequencies only. Because both methods accurately estimate contamination rates, the latter approach, which only requires allele frequencies, has dominated its practical use (Figure 3.1A). However, if allele frequencies are misspecified or unknown, the estimated contamination rates can be severely biased.

Our new method (*verifyBamID2*) avoids such a bias due to misspecified allele frequencies by modelling individual-specific allele frequencies as a function of genetic ancestry, and by jointly estimating genetic ancestry and contamination rates to maximize the likelihood of sequence reads. The genetic ancestry can be represented as coordinates of principal components from cosmopolitan reference panel, such as 1000 Genomes or HGDP (Figure 3.1B). In addition, Our new method can also be used for genetic ancestry estimation, similar to *TRACE/* LASER[20,21], but accounting for potential sequence contamination together. We show that our method provides (1) comparable or more accurate estimates of genetic ancestry than existing methods such as *TRACE*/LASER even in the absence of contamination and (2) reduced bias in contamination rate estimates compared to our previous method requiring known population

allele frequencies using *in silico* contaminated datasets and sequenced genomes from the InPSYght psychiatric genetics sequencing study.

We assessed our new methods in the following steps. First, in the absence of contamination, we demonstrate that our estimation of genetic ancestry provides comparably accurate estimates of genetic ancestry as other state-of-art methods. Second, in the presence of contamination, we demonstrate that joint estimation of genetic ancestry and contamination substantially improves the estimation accuracy of both parameters. Third, using *in-silico* contaminated samples, we demonstrate that our methods robustly provide more accurate estimates than previous methods across various combinations of genetic ancestries and contamination rates. Fourth, from the analysis of deeply sequenced genomes in the InPSYght study, we demonstrate that our new methods deliver more accurate contamination estimates than the previous methods.

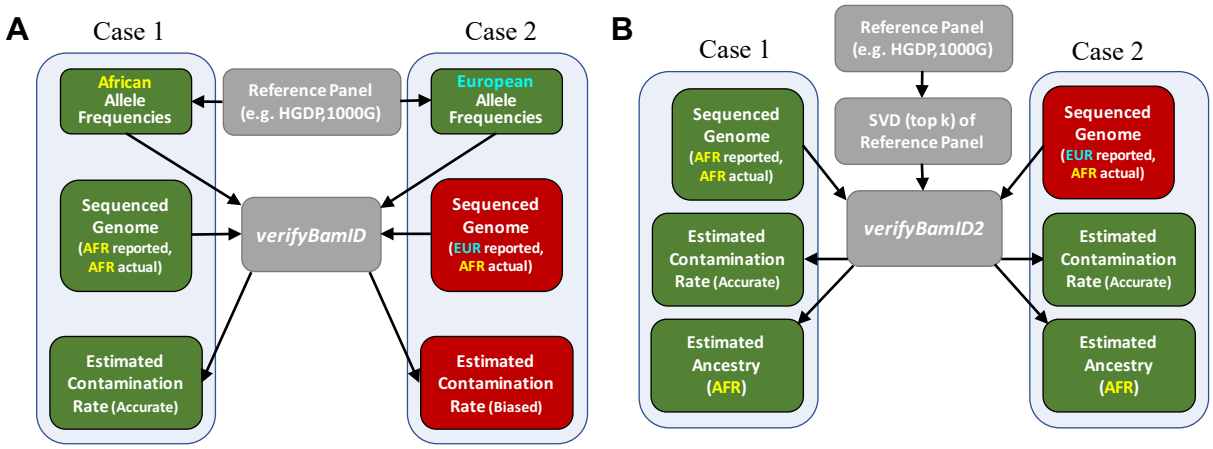

Figure 3.1 Overview of *verifyBamID* and *verifyBamID2* software tools. (A) verifyBamID takes aligned sequence reads (in BAM format) and known variant sites annotated with population allele frequencies (in VCF format) to estimate DNA contamination rates. When allele frequencies are correctly specified, the estimated DNA contamination rates are expected to be accurate (green boxes). However, when the allele frequencies are misspecified (e.g. due to incorrect self-reported ancestry), the estimates of DNA contamination rates may be biased (red boxes). **(B)** verifyBamID2 takes aligned sequence reads (in BAM/CRAM format) and top k singular value decomposition (i.e. PCs and SNP loadings) to estimate the genetic ancestries and contamination rates together. Because verifyBamID2 does not rely on self-reported ancestry, even if ancestry of sample is misspecified or unknown (red box), the estimated contamination rates will be unbiased (green box). In addition, genetic ancestries are also estimated in PC coordinates, adjusting for potential contamination

New Model-based Methods Accurately Estimate Genetic Ancestry

In the absence of contamination, widely used methods such as LASER and *TRACE* are known to estimate genetic ancestry accurately. Because we propose using a new model-based approach to estimate the genetic ancestry (jointly with contamination rates), we first compared the accuracy of our new method, in the absence of contamination, with LASER and *TRACE*. We randomly chose 500 ethnically diverse samples from the 1000 Genomes Project low-coverage (4×) genomes, and 500 African American samples from the deeply sequenced (32×) genomes from the InPSYght project. We estimated their genetic ancestries using 100,000 SNPs from the HGDP reference panel (see Methods for details) and compared their genetic ancestry estimates obtained by LASER (using the same sequence data), and *TRACE* (using the hard-call genotypes). As illustrated in Figure 3.2A, 3.2C, 3.2E, the estimated PC coordinates of the 1000 Genomes individuals are located close to their corresponding HGDP populations across all three methods. Compared to *TRACE* and LASER, we observed that the estimated genetic coordinates from *verifyBamID2* were the closest to the centroid of corresponding HGDP population (Table 3.1) in 4 of the 5 populations (all except TSI). These results suggest that our method provides estimates at least as precise compared to those for other state-of-the-art methods.

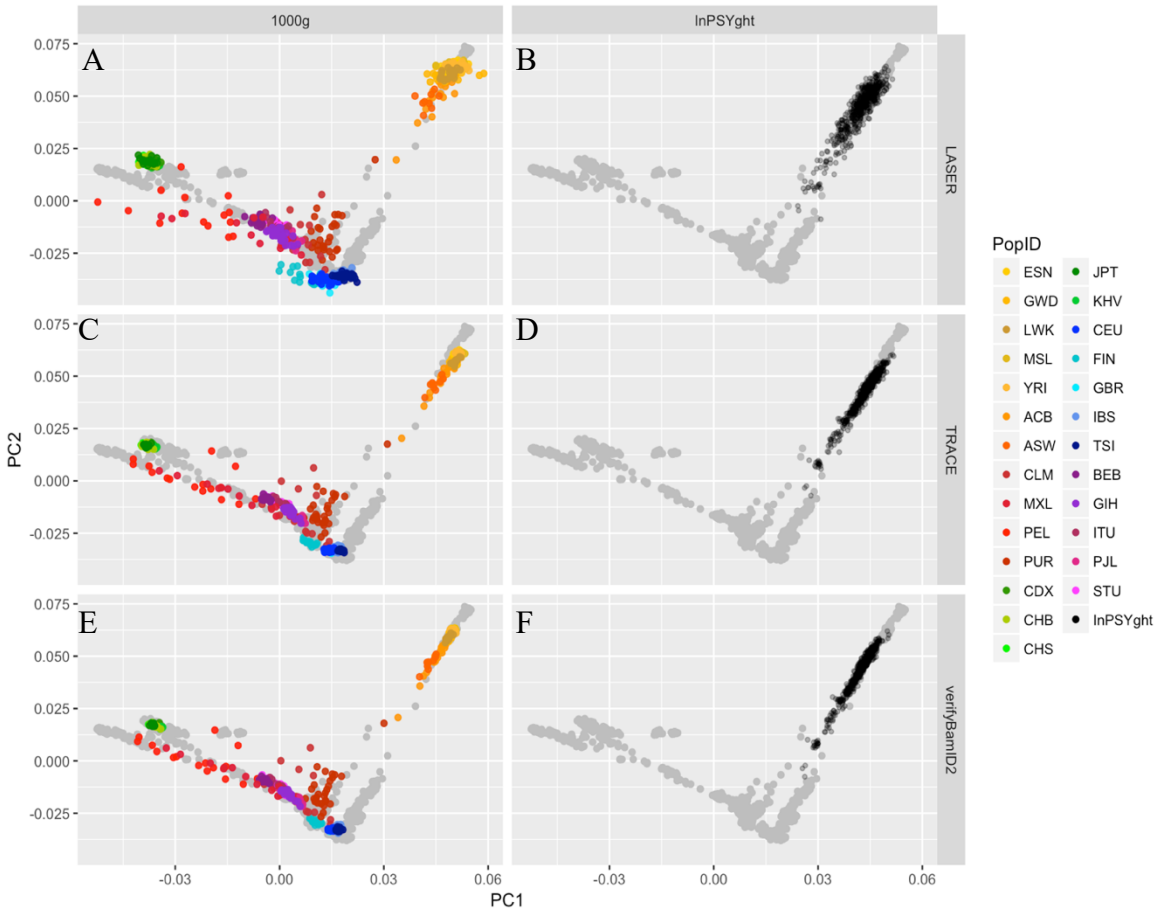

Figure 3.2 Evaluation of estimated genetic ancestry coordinates, in the absence of contamination, between TRACE, LASER, and verifyBamID2 on samples from the 1000 Genomes low coverage genome (n=500, diverse ancestry) sequence data (A,C,E) and from the InPSYght deep genome (n=500, African Americans) sequence data (B,D,F). Panel A and B show results from TRACE, C and D from LASER, and E and F from verifyBamID2 (assuming no contamination). Each point represents a sample, each color represents a population ancestry with the exception that grey point represents PCA coordinates of reference (HGDP) samples.

Table 3.1 Distance between estimated PCA coordinates of HGDP and 1000G populations*

Population Label		*TRACE*	LASER	*verifyBamID2*
1000G	HGDP			
CHB	Han-NChina	1.89	3.01	**0.82**
CHS	Han	1.76	1.81	**1.25**
TSI	Tuscan	**1.62**	2.78	1.86
YRI	Yoruba	2.35	2.62	**0.59**
JPT	Japanese	1.66	1.99	**1.29**

*Mean distances were measured between the PCA coordinates across the population in HGDP (estimated from the array data of Wang et al.[46] and the PCA coordinates estimated from 1000 Genomes low coverage sequence data of the corresponding population, projected onto the same PCA coordinates using TRACE, LASER, or verifyBamID2 (assuming no contamination). Bold face represents the smallest distance among the three methods for each population.

Genetic Ancestry Estimates may be Confounded by DNA Contamination

Next, we constructed *in-silico* contaminated sequenced data from the 1000 Genomes Project and estimated contamination parameters and genetic ancestries jointly. We observed that when sequences are contaminated between different continental populations, the genetic ancestry estimates in PC coordinates drift towards the contaminating population when contamination is ignored (Figure 3.3A) or when assuming that intended and contaminating samples originated from the same population (Figure 3.3C). As the contamination rate increases, drift increases (Figure 3.3A, 3.3C, 3.3E).

However, when we accounted for possible differences in genetic ancestries between the two intended and contaminating samples using our new methods, PC coordinates remained similar to those for uncontaminated samples (Figure 3.3E), and contaminated samples constructed from individuals that belong to the same population (Figure 3.3B, 3.3D, 3.3F).

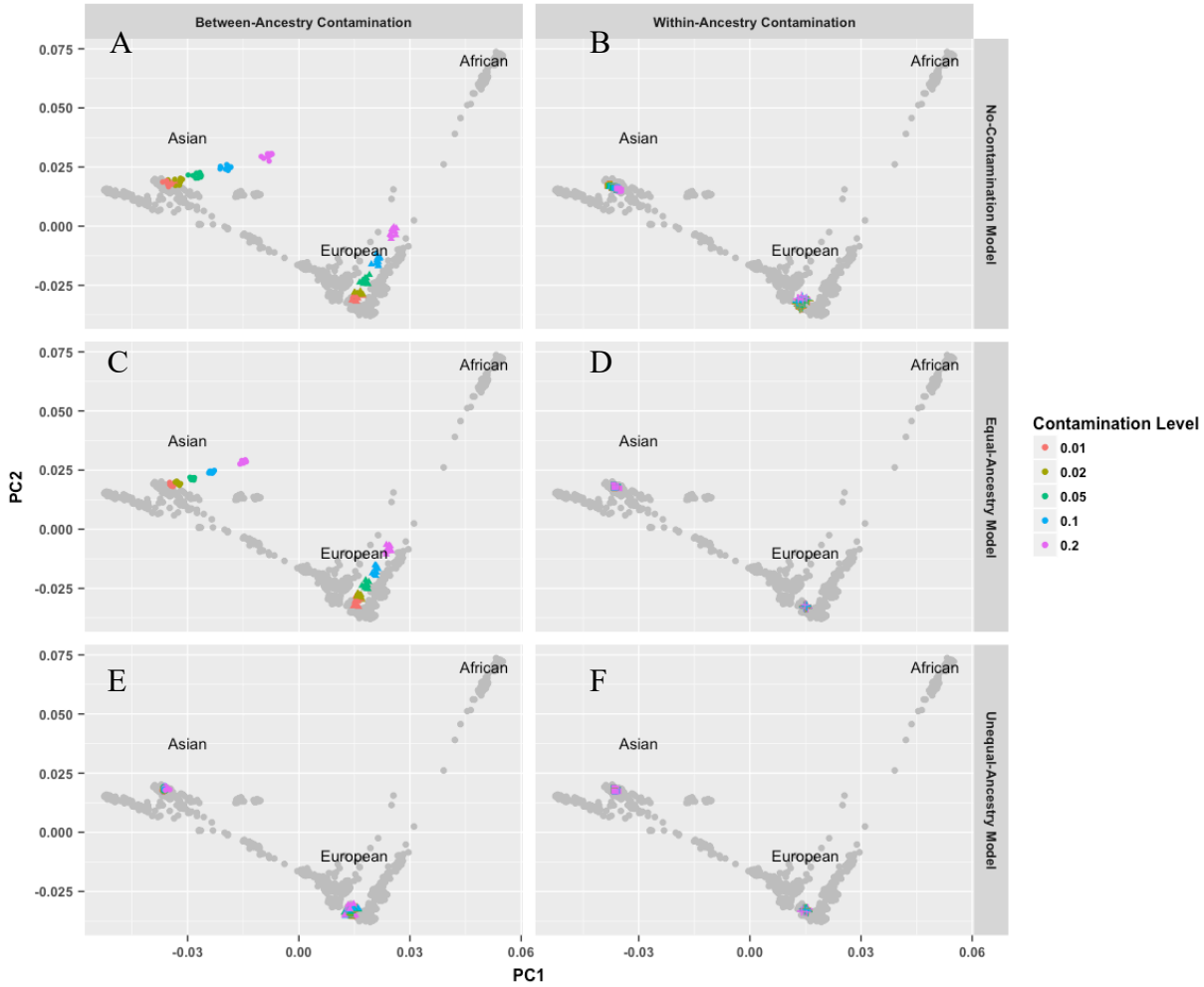

Figure 3.3 Impact of DNA sample contamination on the estimation of genetic ancestry. Each point represents a sample. Grey point represents reference (HGDP) sample and its PCA coordinates, similar to Figure 3.2. Each colored point represents in-silico contaminated samples across various contamination rates and populations. In panel A, C, E, European (GBR) and East Asian (CHS) samples are contaminated with African (YRI) samples at different contamination rates (i.e. between-ancestry contamination). In panel B, D, F, European (GBR) and East Asian (CHS) samples are contamination with another sample in the same population (i.e. within-ancestry contamination). Different colors represent different contamination rate ranging from 1% to 20%. Upper panels (A, B) show verifyBamID2 estimates without modelling contamination. Middle panels (C, D) show verifyBamID2 estimates under the assumption that intended and contaminating populations are identical (i.e. equal-ancestry model). Lower panels (E, F) show verifyBamID2 estimates under the assumption that intended and contaminating populations can be different (i.e. unequal-ancestry model).

Robust, Accurate, Ancestry-agnostic Estimation of DNA Contamination

Next, we evaluated the effect of genetic ancestry misspecification in estimating DNA contamination rates. We constructed contaminated samples between various combinations of

populations and compared the accuracy of estimated contamination rates using both the original methods which assume known allele frequencies and the new methods which estimate contamination rate and genetic ancestry jointly.

When contamination happens within the same population, running original methods with correct continental population allele frequencies specified provided accurate contamination estimates (Figure 3.4A, 3.4E, 3.4I). However, using pooled allele frequencies, which would be a default option when it is infeasible to specify population information *a priori* before sequencing, consistently underestimated contamination rates. Bias was particularly large when intended individuals were of African ancestry.

Specifying incorrect population allele frequencies results in even larger contamination estimation bias. For example, using African allele frequencies on East Asian samples resulted in an average estimate of 2.9% contamination for samples with contamination 10% (Table 3.S1), implying that a large fraction of 10% contaminated samples within East Asian ancestry would not have been flagged for contamination-based exclusion at the contamination-exclusion threshold of 1-3% used by many studies e.g. the Trans-Omics Precision Medicine (TOPMed) study[47].

Our results consistently demonstrated that the ancestry-agnostic method provides as accurate estimates as the original methods specified with correct population labels (Figure 3.4A, 3.4E, 3.4I, Table 3.S1), and the estimates are substantially better than those from pooled allele frequencies or incorrectly specified allele frequencies (Table 3.2).

When the intended and contaminating populations are different, we observed that contamination is sometimes overestimated due to increased fraction of heterozygous genotypes than expected by a given contamination rate under single population model. Our method based on unequal-ancestry model outperforms all the other methods in terms of overall bias and Mean Squared Error (MSE) (Figure 3.4, Table 3.S4), correcting for both upward and downward biases in various ancestry combinations. For example, the relative deviation of estimated to intended contamination rate (i.e. $|\hat{\alpha}/\alpha - 1|$) is reduced by 80% (73-88%) compared to the original *verifyBamID* with various population allele frequencies, suggesting reduced bias. MSE is also reduced by 92% (86-97%). This robustness reflects the ability to incorporate differences in population allele frequencies between intended and contaminating individuals (Figure 3.4B, 3.4C, 3.4D, 3.4F, 3.4G, 3.4H, Table 3.S1).

We also examined the accuracy of our methods for admixed populations by performing a similar experiment using the Mexican population (MXL) and obtained consistent results (Supplementary Table 3.S2).

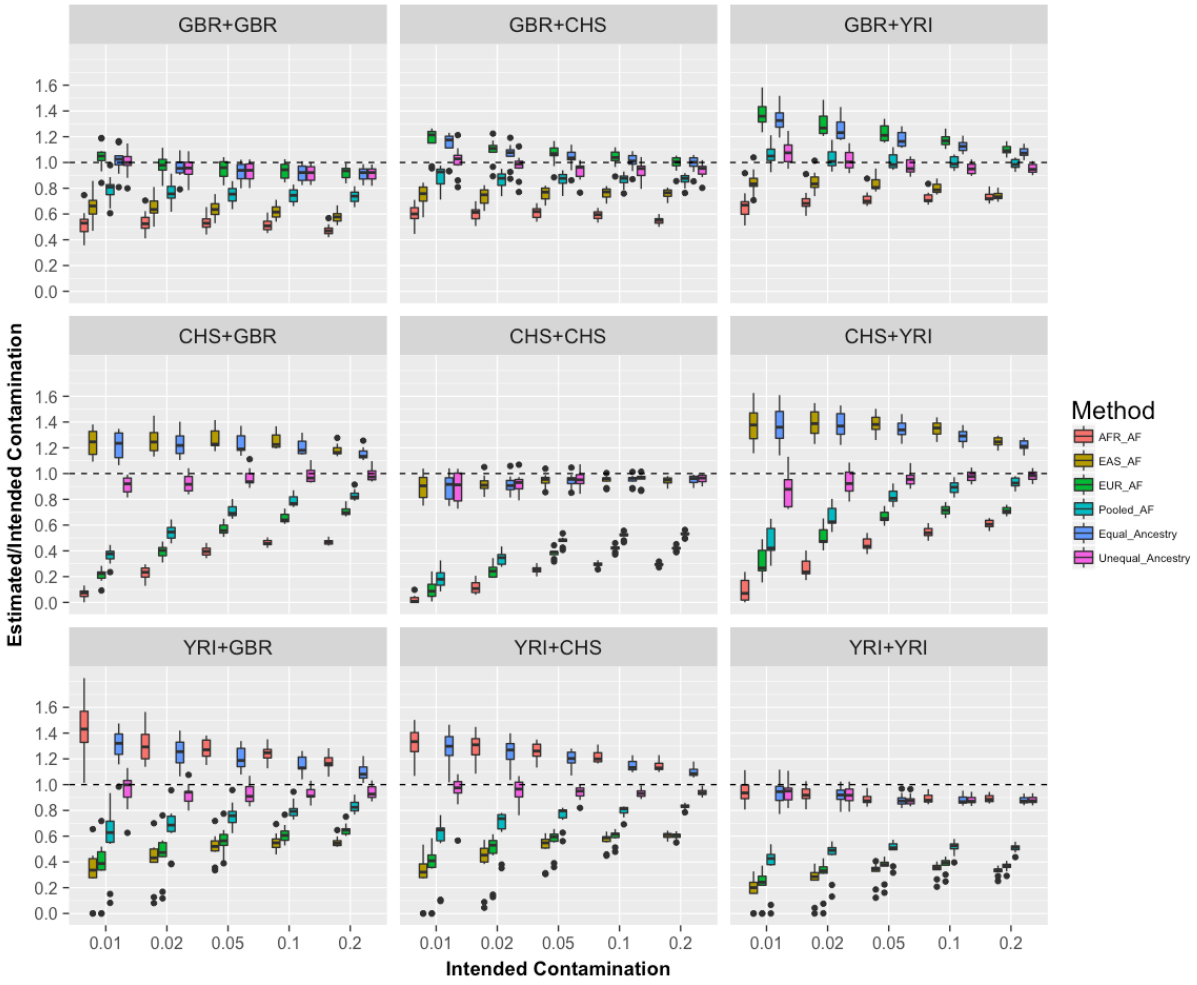

Figure 3.4 Comparison of different models to estimate contamination rates. Horizontal (x) axis shows intended Contamination rate, vertical (y) axis shows the ratio of estimated to intended contamination rates. Each color represents different models to estimate contamination rates. EUR_AF, EAS_AF, AFR_AF represents old verifyBamID using European, East Asian, and African allele frequencies across the continental population using the 1000 Genomes data. Pooled_AF represents the old verifyBamID using aggregated allele frequencies across all 2,504 individuals in the 1000 Genomes Project. "Equal_Ancestry" represents the verifyBamID2 assuming that intended and contaminating samples belong to the same population. "Unequal_Ancestry" represents verifyBamID2 allowing different genetic ancestries between intended and contaminating samples (recommended setting). Each panel represents different combinations of intended (row) and contaminating (column) populations, in the order of GBR, CHS, and YRI.

Table 3.2. Average contamination estimates for 5% contaminated samples (size n=10).

Sample Population		Original Model (Fixed Allele Frequencies)				Equal-Ancestry Model	Unequal-Ancestry Model
Intended	Contaminating	European	East Asian	African	Pooled		
GBR	GBR	**4.73%**	3.19%	2.67%	3.76%	4.63%	4.63%
CHS	CHS	1.90%	4.73%	1.25%	2.38%	4.73%	**4.76%**
YRI	YRI	1.78%	1.58%	**4.44%**	2.45%	4.40%	4.40%
CHS	YRI	3.33%	6.91%	2.27%	4.10%	6.71%	**4.81%**
YRI	CHS	2.79%	2.55%	6.29%	3.76%	5.99%	**4.67%**
GBR	YRI	6.13%	4.16%	3.60%	5.04%	5.90%	**4.83%**
YRI	GBR	2.81%	2.57%	6.38%	3.80%	6.01%	**4.63%**
CHS	GBR	2.87%	6.33%	1.98%	3.55%	6.13%	**4.83%**
GBR	CHS	5.32%	3.78%	3.05%	4.32%	**5.16%**	4.67%

Average contamination estimates of in-silico contaminated samples when the true contamination rate is 5%. Each mixing configuration (e.g. GBR+CHS) contains 10 samples that are constructed with 95% reads coming from the intended sample and 5% reads from the contaminating sample. The estimated contamination rates are obtained using the original version verifyBamID by specifying prior allele frequencies as European, East Asian, African, and Pooled, respectively. Bold represents the closest estimate to the true value of 5%.

Results with Deep Whole Genome Sequence Data from the InPSYght Study

Next, we applied our methods to 500 African American samples from the InPSYght study (see Methods). Consistent with the results from our *in-silico* contamination studies, we observed that the average contamination rate was 1.1-fold higher with newer method (0.36% for unequal-ancestry, 0.37% for equal-ancestry) compared to the original method with pooled allele frequency (0.33%) (Figure 3.5). The number of samples with estimated contamination rate >1% increased from 16 (original method with pooled allele frequency) to 21 (unequal-ancestry method) or 23 (unequal-ancestry method), suggesting our new method more rigorously screens for contaminated samples.

All 500 deeply sequenced genomes in InPSYght study are reported to be African Americans, and indeed the estimated PC coordinates for all 500 individuals under all three methods lie between European and African samples. Compared to other methods to estimate genetic ancestry, our estimates resulted in tighter clustering along the European-African segment than LASER, and similarly tight clustering to *TRACE* (Figure 3.2B, 3.2D, 3.2F). For example, the correlation coefficient between the PC1 and PC2 coordinates were 0.927 for LASER, 0.981 for *TRACE*, and 0.985 for *verifyBamID2*, corroborating that *verifyBamID2* results in more precise estimate of African ancestry along the European-African segment in PC coordinates.

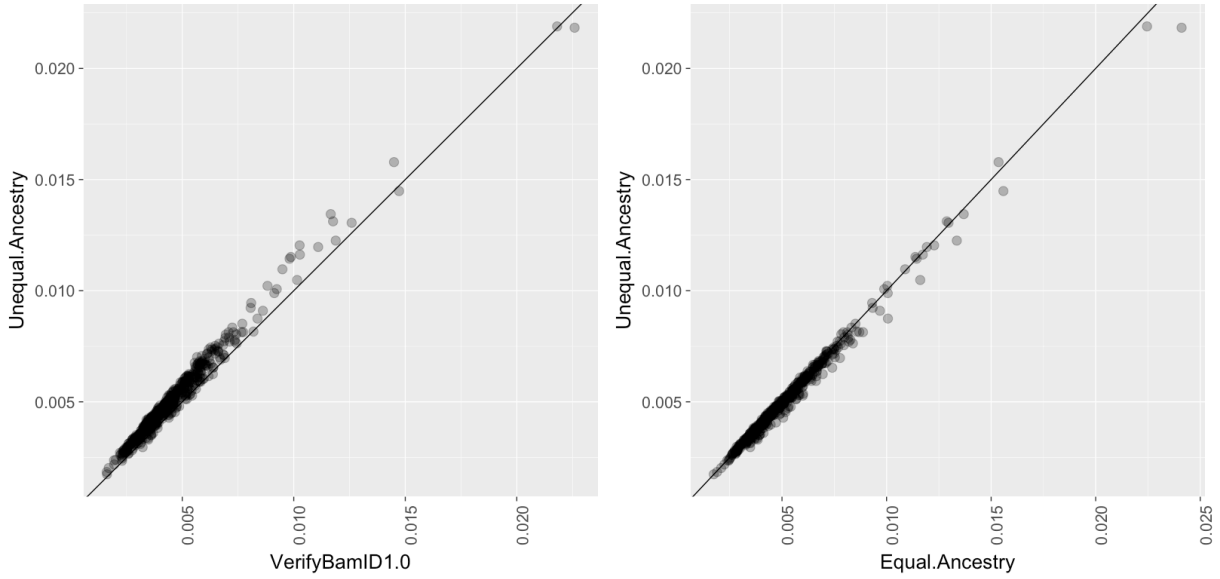

Figure 3.5 Comparison of contamination estimation between using *verifyBamID* and *verifyBamID2* on 500 InPSYght samples. All subjects are African Americans. Each dot represents the pair of contamination rate estimates using different methods. The left panel shows the estimated contamination rates of the original verifyBamID with pooled allele frequencies, which is the default setting of verifyBamID in x-axis. Y-axis shows verifyBamID2 with unequal-ancestry model (y-axis). Each point represents a sequenced subject. The right panel compares the estimated contamination rates between two models (unequal-ancestry vs. equal-ancestry) of verifyBamID2 on the same dataset.

Impact of Number of Markers on Accuracy, Computational Cost, and Memory Requirements

As we have shown previously[43], there are trade-offs between computation cost and accuracy of contamination estimates. Using as many as 100,000 variants results in accurately estimated intended contamination rate. For example, MSE of relative deviation (i.e. $|\hat{\alpha}/\alpha - 1|$) was 0.02, 0.01, 0.01 when the intended contamination was 1%, 2%, and 5%, respectively. When we use 10,000 variants, the MSEs modestly increased to 0.11, 0.04, and 0.01, respectively. When we use only 1,000 variants, MSEs further increased to 0.69, 0.25, 0.11, suggesting that the estimates may not be precise for low contamination rate when using only 1,000 variants. (Supplementary Table 3.S3).

We also evaluated the computational cost and memory consumption of *verifyBamID2* on whole genome sequence data with various coverages. For the BAM files from the 1000 Genomes whole genome sequence data (4.3-5.1× coverage), the average wall-clock running time was 5.5 minutes with a single thread and peak memory consumption was 505 MB when using 10,000 markers in a server with Xeon 2.27GHz processor. When using 100,000 markers, the average wall-clock running time was 20.5 minutes with a single thread and 8.0 minutes with four threads, and peak memory consumption was 528 MB.

For deep genome data from the InPSYght study (31× coverage) stored in CRAM format, the average wall-clock time was 17.3 minutes and peak memory consumption was 514 MB when using 10,000 markers. For 100,000 markers the average wall-clock time was 155.6 minutes (single thread) or 96 minutes (four threads) and peak memory consumption was 548 MB.

Discussion

Contamination detection is an essential step in the sequence analysis process that has important effects on following downstream analyses. Early and accurate estimation of DNA contamination can prevent wasted effort, time, and money by identifying the problems early on before too many samples are sequenced using contamination-prone protocols. Our previous method enabled such a timely contamination detection from sequence data and population allele frequencies at known variant sites, without requiring independent SNP genotype data. Our new method maintains these advantages, and in addition provide three more. First, because our joint analysis method is agnostic to genetic ancestry, it eliminates sample-to-sample variation in the parameter settings for the contamination checking procedure, simplifying the sequence analysis pipeline. Second, it provides more robust contamination estimates against potentially misspecified population allele frequency of the intended (or contaminating) samples when relying on the reported ancestry information. Third, it provides accurate estimates of genetic ancestries for both intended and contaminating samples. This enables additional sanity checking of the sequence data, such as determining whether a sequenced sample matches its expected (participant-reported) ancestry. It also facilitates incorporating ancestry information in the variant calling and downstream analysis and allows us to track the source of contamination more precisely when contamination occurs.

Our method can be used not only to detect and estimate contamination, but also to estimate genetic ancestry from sequence data. Relatively few methods, such as LASER and *bammds*[48], exist for estimating genetic ancestry from sequence data while several methods have been developed for array-based genotypes, such as EIGENSOFT[49], FRAPPE[50], ADMIXTURE[51], and *TRACE*[46]. We have demonstrated that our method provides ancestry estimates as or more

accurate than LASER, particularly when the sequenced samples are contaminated between different ancestries.

By jointly estimating genetic ancestry and contamination, we are able to accurately estimate contamination without requiring ancestry information *a priori*. Since obtaining population allele frequency information may be infeasible or even impossible at the time of sequencing, it is important to highlight that our ancestry-agnostic approach provides more timely and accurate feedback to the sequencing facilities. Our ancestry-agnostic approach also simplifies the sequence analysis pipeline, because the same input arguments can be applied across all samples regardless of their genetic ancestry. In the case where self-reported ancestries are available, our method can identify errors in the self-reported ancestries while estimating contamination.

The key idea of using individual-specific allele frequencies (ISAF) to account for population structure in genetic analysis has been suggested previously in the context of characterizing population structure or identifying highly differentiated variants across populations[31,32]. To the best our knowledge, our method describes the first likelihood-based model utilizing ISAF to represent high throughput sequence reads under population structure and/or contamination. While previous studies proposed logistic models as alternative to linear model[31,32], we used linear models (bounded by minimum and maximum value) between allele frequencies and population structure represented by Singular Value Decomposition (SVD) on the genotype matrix. We made this choice because the logistic model is computationally more intensive, and the linear model is accurate for the common variants we use, as demonstrated by the previous studies[32].

Even though our method substantially improves the accuracy of contamination estimates compared to the original *verifyBamID*, we do see slightly underestimation of contamination rates, especially when intended contamination rate is high. Our method overestimates contamination if there are more heterozygous genotypes than expected by allele frequencies under HWE, and underestimate contamination if there are less heterozygous genotypes than expected. We believe that slightly inaccurate allele frequency estimate (even with ISAF) and violation of HWE (due to population structure or copy number variants) are contributing to the slight underestimation of contamination rates but have not validated the conjecture experimentally yet.

Because we use Nelder-Mead optimization for maximum likelihood estimation, it is possible that the estimates do not converge to the global maximum, especially when many principal components are used. We observed that estimating the full unequal-ancestry model parameters sometimes does fail to converge especially when there is little or no contamination, due to the limited identifiability of the genetic ancestry of contaminating samples in this situation. Starting by estimating contamination rate and shared genetic ancestry parameters using the equal-ancestry model, and using those estimates as start values for the unequal-ancestry model to allow different ancestries between the intended and contaminating samples dramatically improved convergence; in fact, the method converged to consistent estimates across multiple starting points within 1,000 iterations in all our benchmark cases, in both real and *in-silico* contaminated data. When the contamination rate is extremely small (e.g. <0.1%), estimation of genetic ancestry of contaminating samples can still be challenging, but the its impact on genotyping accuracy is likely small as demonstrated previously [43]. We allow unequal ancestries between intended and contaminating samples only when the likelihood substantially

improves beyond AIC threshold between equal ancestry and unequal ancestry models. This procedure effectively removed all outlier estimates of genetic ancestries of contaminating samples in our experiments.

There are other possible useful extensions to our joint contamination and estimation method. We are extending these methods to detect and estimate contamination for RNA-seq and other epigenomic sequence data. The method can also be extended to handle contamination in cancer genomic data. The same model has potential utilities in other areas, such as single cell transcriptomics[52]. As our method leverages excess heterozygosity to estimate contamination rates, it is important that the sequence reads have many variant sites with read depth 2 or greater to have sufficient power to estimate contamination in the extended models.

We expect that our new *verifyBamID2* software will facilitate more accurate, convenient, and timely quality control of sequence genomes. Our software tool is publicly available at http://github.com/Griffan/verifyBamID. Our GitHub repository provides reference files that can be used as test input for our methods. These files contain key input files required for *verifyBamID2*, including variant loadings, supporting various genome builds (GRCh37 and GRCh38), and various numbers of variants.

Methods

Overview

We aim to jointly estimate sample contamination rates and genetic ancestry from sequence reads without specifying population allele frequencies. First, we describe our previous mixture model to estimate contamination rates assuming population allele frequencies are

known. Second, we introduce a model for sequence reads using population allele frequencies as a function of genetic ancestry represented in principal component coordinates. Third, we extend the model to enable joint estimation of contamination rates and genetic ancestry. Fourth, we evaluate our methods using *in silico* contaminated samples and whole genome sequence data from the InPSYght study.

Likelihood-based Mixture Model for DNA Sequence Contamination

In our previous contamination detection methods[43], we assumed that the DNA sequence reads from an intended sample are contaminated by sequence reads from at most one contaminating sample from the same population, and that the population allele frequencies of all analyzed genetic variants are known. For each bi-allelic variant i $(1 \leq i \leq m)$, let $b_{ij} \in \{R, A, O\}$ $(1 \leq j \leq D_i)$ be the observed base call representing the reference allele (R), alternate allele (A), or other allele (O) for the j-th read that overlaps the variant; D_i is the observed sequence depth at variant i. Let $e_{ij} \in \{0,1\}$ be a random variable indicating whether a sequencing error did (1) or did not (0) occur for observed base b_{ij}; we assume e_{ij} follows a Bernoulli distribution with success probability $10^{-\frac{Q_{ij}}{10}}$ where Q_{ij} is a phred-scale base quality score of b_{ij}. In the absence of contamination, if the true genotype $g_i^s \in \{0,1,2\}$ represents the count of alternate alleles of the sequenced sample $s \in \{1,2\}$, then $\Pr\left(b_{ij} \middle| g_i^s, e_{ij}\right)$ can be easily represented as in Table 3.3**,** making the simplifying assumption of equally likely errors across four possible nucleotides. We assume that the observed sequence reads are a $(1-\alpha) : \alpha$ mixture of intended and contaminating reads given a contamination rate $0 \leq \alpha \leq 1$. Let g_i^1 and g_i^2 represent the true genotypes of the intended and contaminating samples at variant i, respectively. Then the mixture model likelihood of each observed base becomes

$$\Pr\left(b_{ij}\middle|g_i^1, g_i^2, e_{ij}; \alpha\right) = (1-\alpha)\Pr\left(b_{ij}\middle|g_i^1, e_{ij}\right) + \alpha\Pr(b_{ij}|g_i^2, e_{ij}) \quad (1)$$

Assuming a homogenous population with known population allele frequency f_i and Hardy-Weinberg Equilibrium (HWE), Pr $(g_i^2; f_i)$ follows a $\text{Binomial}(2, f_i)$ distribution. Under the simplifying assumption of independent variants, the likelihood of the contamination rate becomes

$$L(\alpha) = \prod_{i=1}^{m} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{D_i} \sum_{e_{ij}} \Pr\left(b_{ij}\middle|g_i^1, g_i^2, e_{ij}; \alpha\right)\Pr\left(e_{ij}\right) \right\} \Pr(g_i^2; f_i)\Pr(g_i^1; f_i) \quad (2)$$

The maximum likelihood estimate (MLE) of contamination rate $\hat{\alpha}$ can be obtained using Brent's algorithm[53].

As we previously reported[43], this model assumes correctly specified population allele frequencies f_i.

Table 3.3. Conditional probability $P(b_{ij}| g_i, e_{ij})$ of read b_{ij} given true genotype g_i and the variable representing the event of base calling error e_{ij}

True Genotype g_i	**Base Calling Error Event e_{ij}**	**Pr(b_{ij} = R)**	**Pr(b_{ij} = A)**	**Pr(b_{ij} = O)[b]**
g_i = RR[a]	e_{ij} = 0	1	0	0
	e_{ij} = 1	0	1/3	2/3
g_i = RA[a]	e_{ij} = 0	1/2	1/2	0
	e_{ij} = 1	1/6	1/6	2/3
g_i = AA[a]	e_{ij} = 0	0	1	0
	e_{ij} = 1	1/3	0	2/3

[a] RR, RA, AA: homozygous reference, heterozygous, and homozygous non-reference genotypes

[b] O: alleles other than R or A; assumes four possible alleles (bases)

Likelihood-based Estimation of Genetic Ancestry (in the absence of contamination)

We extend this model to incorporate genetic ancestry. The key idea of this extension is to use the individual-specific allele frequency (ISAF)[31,32] to model the likelihood of the sequence reads. Several methods, including Spatial Ancestry Analysis (SPA)[54] and logistic factor analysis (LFA)[32], previously proposed modelling allele frequency as a function of genetic ancestry via principal component (PC) coordinates.

Let G be an $m \times n$ genotype matrix (where G_{ir} = 0, 1, or 2 is the number of non-reference alleles at variant *i* in individual *r*) of a genetically diverse reference panel of size n, such as 1000 Genomes or HGDP. We define ISAF f_i $(0 \leq f_i \leq 1)$ for variant *i* as a weighted average of genotypes from the reference panel ($f_i = \sum_{r=1}^{n} w_r G_{ir}$), where $0 \leq w_r \leq 1$ and $G_{ir} \in \{0,1,2\}$ for individual *r*. For a homogenous population, $w_r = \frac{1}{2n}$ results in a *pooled allele frequency* across all individuals in the reference panel. If each individual can be categorically represented as a one of *k* mutually exclusive subpopulations, the *population-specific allele frequency* for the subpopulation $s \in \{1,2,\cdots,k\}$ can be represented as $w_r = \frac{I(s_r=s)}{2n_s}$, where and $s_r \in \{1,2,\cdots,k\}$ represents the subpopulation that individual $\underline{r}$ belongs to, and n_s represents the size of subpopulation s . More generally, if individual's genetic ancestry is represented as continuous variables (such as PCs, SPAs, or LFAs), the individual-specific allele frequency (ISAF) can be represented as a function of the continuously represented genetic ancestry[32,55]. The estimated ISAF can be viewed as one half times the genotype dosages approximated from a fixed number(=*K*) of factors, such as PCs, SPAs, or LFAs. In our method, we used a linear model to estimate ISAF from PCs, similar to previous studies[31,32]. Given the reference panel

genotype matrix G, let $\frac{1}{2}\hat{G}$ be the *ISAF matrix* as a function of top *K* factors. ISAF matrix $\frac{1}{2}\hat{G}$ should well approximate $\frac{1}{2}G$. For example, under a linear model, typical principal component analysis takes the singular value decomposition (SVD) of the mean-centered genotype matrix $\overline{G} = G - 2\boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}} = UDV^T$, where $\boldsymbol{\mu} = \frac{1}{2n}G\mathbf{1}_{\boldsymbol{n}}$ is the pooled allele frequencies and $\mathbf{1}_{\boldsymbol{n}}$ is the column-vector of ones. Using the top *K* eigenvalues and corresponding eigenvectors $U^{(K)}, D^{(K)}, V^{(K)}$ from the SVD, it is known that $\hat{G} = \frac{1}{2}U^{(K)}D^{(K)}\left[V^{(K)}\right]^T + \boldsymbol{\mu}\mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}$ minimizes $\left\|G - \hat{G}\right\|_2 = \sum_{i,r}(G_{ir} - \hat{G}_{ir})^2$ among all possible rank *K* matrices[56], making it a good proxy for the ISAF matrix.

For a new individual *s* with genetic ancestry represented as $\boldsymbol{x_s} \in \mathbb{R}^K$ in the PC (eigenvector) space of the reference panel, the ISAF for *i*-th variant can be modelled as $f_i(\boldsymbol{x_s}) = \frac{1}{2}\boldsymbol{u_i^{(K)}}D^{(K)}\boldsymbol{x_s^T} + \mu_i$, where $\boldsymbol{u_i^{(K)}}$ is *i*-th row of $U^{(K)}$ and μ_i is the *i*-th element of $\boldsymbol{\mu}$. To avoid boundary condition, we constrain $\frac{\varepsilon}{2n} \leq f_i(\boldsymbol{x_s}) \leq 1 - \frac{\varepsilon}{2n}$ for a fixed ε (we used $\varepsilon = 0.5$ in our experiments). Then the overall likelihood of an individual's genetic ancestry $\boldsymbol{x}$ is

$$L(\boldsymbol{x_s}) = \prod_{i=1}^{m}\sum_{g_i}\left\{\prod_{j=1}^{D_i}\sum_{e_{ij}}\Pr\left(b_{ij}\middle|g_i, e_{ij}\right)\Pr\left(e_{ij}\right)\right\}\Pr\left(g_i; f_i(\boldsymbol{x_s})\right) \qquad (3)$$

where g_i represents the unobserved genotype of the sequenced sample at variant *i*. The maximum-likelihood genetic ancestry coordinates can be estimated as $\widehat{\boldsymbol{x}}_{\boldsymbol{s}} = \text{argmax}_{\boldsymbol{x_s}\in\mathbb{R}^k}L(\boldsymbol{x_s})$ using the Nelder-Mead[57] algorithm, starting with PC coordinates of a randomly selected individual from the reference panel. In all our experiments, we always obtained consistent estimates of $\widehat{\boldsymbol{x}}_{\boldsymbol{s}}$ regardless of start values with *K*=4, which is the default parameter of our implementation. Using *K*=4 gave us noticeably more precise estimates of contamination rates and genetic ancestry than smaller K (data not shown). Using larger values of *K* (e.g. *K*=8)

substantially increased the computational time of Nelder-Mead algorithm and failed to converge occasionally.

Joint Estimation of Genetic Ancestry and DNA Contamination

Because our goal is to obtain unbiased estimates of the DNA contamination rate α agonistic to prior knowledge of the genetic ancestry, we propose to jointly estimate α and ancestry by combining the models described in the previous sections. Let $\boldsymbol{x_1}, \boldsymbol{x_2} \in R^K$ be the genetic ancestries of the intended and contaminating samples. Then the likelihood under the combined model is

$$L(\alpha, \boldsymbol{x_1}, \boldsymbol{x_2})$$

$$= \prod_{i=1}^{m} \sum_{g_i^1} \sum_{g_i^2} \left\{ \prod_{j=1}^{D_i} \sum_{e_{ij}} \Pr\left(b_{ij} \middle| g_i^1, g_i^2, e_{ij}; \alpha\right) \Pr\left(e_{ij}\right) \right\} \Pr\left(g_i^1; f_i(\boldsymbol{x_1})\right) \Pr\left(g_i^2; f_i(\boldsymbol{x_2})\right)$$

When the contamination rate $\alpha \approx 0$, the parameters corresponding to $\boldsymbol{x_2}$ do not contribute (much) to the likelihood and the estimates of $\boldsymbol{x_2}$ may be unstable. To address this problem, we initially assume that the intended and contaminating samples are from the same population $\boldsymbol{x_1} = \boldsymbol{x_2}$ ('equal-ancestry' model) and then repeat the analysis allowing for $\boldsymbol{x_1} \neq \boldsymbol{x_2}$ ('unequal-ancestry' model). The dimension of parameter space for the unequal-ancestry model is $2k + 1$. We choose final parameter estimates between the two models based on Akaike Information Criterion (AIC)[58].

Evaluation on *in-silico* Contaminated Data Based on 1000 Genomes Project Samples

We constructed *in-silico* contaminated DNA sequence reads using aligned low-coverage whole genome sequence reads from the 1000 Genomes phase 3 project[45]. We filtered out unmapped and mark-duplicated reads and then randomly sampled aligned sequence reads proportional to the intended contamination rates $\alpha \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. To match the mixing proportion of sequence reads originated from intended and contaminating to be $(1-\alpha):\alpha$, each read was sampled with probability $(1-\alpha)$ and $\frac{B_1}{B_2}\alpha$ from each sample, where B_1 and B_2 are number of aligned bases from unique reads from intended and contaminating samples. We selected four populations, CHS (Han Chinese South), GBR (British in England and Scotland), MXL (Mexican Ancestry from Los Angeles USA), YRI (Yoruba in Ibadan, Nigeria), and arbitrarily selected 10 pairs of individuals with similar sequencing depths within the same population and across populations. To estimate genetic ancestry and/or contamination rate for these *in-silico* contaminated sequence reads, we used a reference panel of 938 HGDP[44] individuals across 1,000, 10,000 and 100,000 randomly chosen SNPs (pooled MAF > 0.5%), avoiding variants masked by the 1000 Genomes Project[45]. When we compared estimated genetic ancestry with LASER, we used the same set of selected SNPs and sequence reads as input. For *TRACE*, we used genotypes from the phase 3 release (for 1000 Genomes) or an interim callset from the *GotCloud* software tool[36] (for InPSYght, see next section for details) on the same SNP set.

Experiment with Real Sequence Data from the InPSYght Study

Next, we applied our method to 500 deeply sequenced (mean depth 32x) genomes from the first two batches of the InPSYght study. For each sample, we evaluated the results from the six models: (1) the original *verifyBamID* using pooled allele frequencies; the original *verifyBamID* using (2) African, (3) East Asian, and (4) European allele frequencies; (5) the new *verifyBamID2* under the equal-ancestry model; and (6) *verifyBamID2* under the unequal-ancestry model. To calculate pooled, population-specific, and individual-specific allele frequencies, we used the 1000 Genomes phase 3 reference panel (n=2,504), randomly selecting 100,000 SNPs among the sites also polymorphic in Illumina Human Omni 2.5 array, with the same filtering criteria (MAF > 5% and 1000 Genomes mask) as above.

Software Availability

The software is published under the MIT license. The source code of *verifyBamID2* is available in the Supplemental Material as well as at https://github.com/Griffan/VerifyBamID.

Chapter IV.

Genotyping-free Deconvolution of Multiplexed Single Cell Experiment over Multiple Individuals

Introduction

The advent of massively parallel single-cell RNA sequencing (scRNA-seq) technologies dramatically enhances the resolution to understand how genetic and/or environmental factors alter transcriptomic profiles of individual cells and affect interactions between them. First-generation scRNA-seq technologies require preparation of sequencing libraries of individual cells separately, limiting the throughput to assay a large number of cells and creating cell-to-cell batch effects within individual samples. Barcoding-based scRNA-seq technologies have emerged to overcome these limitations by enabling massive digital barcoding of individual cells in parallel, allowing us to profile thousands of single-cell transcriptomes with a single library preparation.

Digital barcoding of individual cells is performed in either of the two ways, using STAMP (single-cell transcriptome attached to microparticle) or combinatorial indexing. In STAMP-based methods, such as Drop-seq[59], InDrops[60], 10x Chromium[61], or Seq-well[62], a pair of single cell and barcoded microparticle are contained within a droplet or a microwell, often aided

by microfluidic devices, to perform cell lysis and RNA hybridization. These STAMPs undergo reverse-transcription into cDNAs and amplification together to be sequenced as a single library. In combinatorial indexing methods, such as sci-RNA-seq or SPLiT-seq[63,64], cells are randomly divided into distinct subsets and barcoded iteratively for multiple rounds using split-pool barcoding. Similar technologies are applied to assay additional molecular profiles at single-cell resolution. Single nucleus RNA-seq technologies, such as DropNc-seq, enable massively parallel transcriptomic profiling not only from fresh tissue, but also from frozen or lightly fixed tissues, facilitating scRNA-seq studies for broader tissue types[65]. Other single cell epigenomic profiling technologies, such as scATAC-seq[66] and sci-CAR[67] are also being actively developed for broader use.

Both of the single cell barcoding methods - STAMP-based and combinatorial indexing - aim to associate a barcode of nucleotide sequences to a specific cell, so that the scRNA-seq reads matching a specific barcode sequence is assumed to be originated from the same cell. However, a barcoded sequence may represent multiple cells instead of a single cell if the ideal assumption does not hold. In STAMP-based methods, a droplet or a microwell may contain multiple cells (i.e. multiplets) by chance. The chance of having multiplets increases when loading a large number of cells to maximize the utility of the barcoded microparticles, so there is an inverse correlation between the multiplet rate, and the number of cells assayed per library. Similarly, in combinatorial indexing methods, a single combinatorial index of nucleotide sequences may represent multiple cells. The chance of such a barcode collision should be very low in an ideal situation where the split-pool barcoding procedure fully randomizes the possible barcodes across the cells. However, a higher collision rate may be observed when the barcodes are non-randomly

distributed or when the split-pool procedure does not completely randomize the cells. Barcode collision may also happen in STAMP-based methods, too.

Recently, a new cost-effective experimental strategy, *mux-seq*, enabled scRNA-seq across tens of samples in a single library preparation[68]. *Mux-seq* harnesses natural genetic variation as "genetic barcodes" to deconvolute the sample origin of individually barcoded cells. More importantly, by leveraging scRNA-seq overlapping with genetically polymorphic variants, *mux-seq* allows us to detect multiplets (i.e. cell barcodes representing two or more cells) that originated from two or more individuals, reducing the undetected multiplet rate by ~*N*-fold when *N* samples are multiplexed. This increased sensitivity in multiplet detection in return allows us to load ~*N*-fold more cells per run than the standard workflow at a fixed multiplet rate, substantially saving per-sample and per-cell cost for library preparation and dramatically reducing sample-to-sample batch effects. Due to several-fold lower per-sample cost, *mux-seq* workflow enables cost-effective scRNA-seq studies across diverse samples at a population scale.

The statistical method behind the *mux-seq* workflow, *demuxlet*, uses a mixture model to model the likelihood of scRNA-seq reads overlapping with genetic variants to evaluate the likelihood of possible singlets and doublets to determine which configuration of sample origin best explains the observed reads of a cell barcode based on a likelihood-based model selection criterion. To evaluate the likelihood, *demuxlet* requires that the genotypes of each multiplexed sample are available from an external source via array-based genotyping or DNA sequencing. However, requiring external genotyping can often become a bottleneck in the analysis of scRNA-seq data, due to the additional time and efforts required to prepare and perform external genotyping experiments and to process the genotype data, including quality control, strand matching, imputation, and sample identity matching between different types of data. As

investigators who design and perform multiplexed experiments do not necessarily have sufficient expertise in handling genetic data, the bottleneck in external genotyping steps often becomes more serious than it seems in practice.

On the other hand, in principle, *demuxlet* does not have to require external genotypes to enable sample demultiplexing and doublet detection. Suppose an ideal case where we can perfectly assign each droplet into its originating sample except for one droplet. In such a case, scRNA-seq reads that belong to each individual can be merged to calculate genotype likelihood at each variant site, and *demuxlet* can be used to infer the source of one remaining target droplet. In practice, a similar procedure may be possible (1) by clustering each droplet into the most-likely originating sample probabilistically, (2) by evaluating the genotype likelihood of each cluster using the assigned reads, and (3) by re-evaluating the likelihood that a droplet being originated from each cluster (and each pair of clusters to account for doublets) to re-assign each droplet to most likely cluster or a pair of clusters. Following on this idea, in this Chapter, we propose a new method, *freemuxlet*, to perform sample demultiplexing and doublet detection without requiring external genotyping. In *freemuxlet*, we utilize Bayes Factors to evaluate the likelihoods of a barcoded droplet being a doublet and estimate the pairwise genetic distance between a pair of droplets to determine whether they should belong to the same individual or not.

Results

Overview of *freemuxlet* Algorithm

Briefly, a *mux-seq* workflow[68] pools thousands of cells from many unrelated individuals together to prepare a scRNA-seq sequencing library. Each barcoded droplet (or combinatorial index) may contain a cell from a single individual (singlet) or two or more cells from multiple individuals (multiplet). If a barcoded droplet contains multiple cells from the same individual, the current implementation of *freemuxlet* considers it as a singlet. *Freemuxlet* takes a list of variant sites with known population allele frequencies and examines scRNA-seq reads overlapping with the variant sites to cluster each barcoded droplet into their samples of origins if they are singlets while detecting multiplets (Figure 4.1).

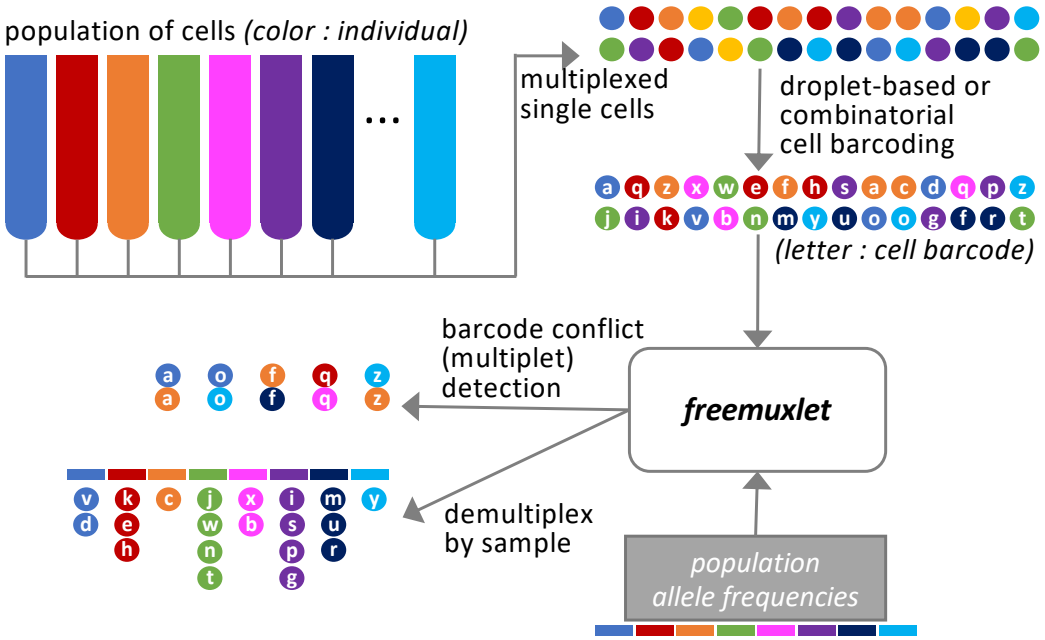

Figure 4.1. Overview of the *mux-seq* workflow based on *freemuxlet*. Different colors represent different samples of origin. Each circle represents an individual cell, and the letter represents the barcode of each cell. If the cell barcode is unique, it is considered as singlet will be clustered based on the origin of samples. When a barcode represents two or more cells from different individuals, *freemuxlet* detects them as conflicts

For each barcoded droplet, *freemuxlet* evaluates a metric called "singlet score", a Bayes Factor quantifying whether scRNA-seq reads from the droplet are more likely originated from a single individual (i.e. diploid) or two individuals (i.e. tetraploid). For each pair of droplets, it evaluates another Bayes Factor metric called "genetic distance" to evaluate whether scRNA-seq reads from the pair of droplets were originated from the same individual or not (Figure 4.S1). Based on these two metrics, *freemuxlet* iteratively cluster putative singlets into the samples of origin using a greedy algorithm guided by the two metrics, while identifying doublets based on the likelihood model similar to *demuxlet* (See Methods for details).

Evaluation on 7-way *Mux-seq* Data with Cell-Hashing

We first applied our *freemuxlet* method on a scRNA-seq dataset that contains 23,111 barcoded droplets sequenced and multiplexed using 7-way *mux-seq* workflow. Especially, this dataset was multiplexed using two independent methods, (1) "genetic multiplexing" (*mux-seq*) with external (array-based and imputed) genotyping data available and (2) "cell hashing"[69] that leverages additional antibody tags on the cell surfaces to identify the source of samples. Therefore, by evaluating the concordance between three different approaches – *demuxlet*, cell hashing, and *freemuxlet* – we can better understand how each method behaves in comparison to other two methods, even though no single method can be considered as the "gold standard".

We first evaluated the marginal number of droplets classified as singlets and doublets by each method. Assuming the multiplet rate is 1% when loading 1,000 cells per run, the expected number of detectable doublets among 23,111 cells is $\frac{6}{7}(1 - 0.99^{23.111}) = 17.8\%$ (4,106 cells), if 7 samples are uniformed multiplexed together. The estimated fractions of doublets using cell hashing, *demuxlet*, and *freemuxlet* were 28.9%, 19.3%, and 18.8%, respectively, excluding the

ambiguously assigned barcoded droplets. The estimated fraction of doublets from *demuxlet* and *freemuxlet* were much more concordant to the theoretical expectation (Figure 4.2A).

Next, we evaluated the concordance between different methods in classifying each barcoded droplet into a singlet from one of the 7 individuals, a doublet, or ambiguous classification. Because there is no established "gold standard", when evaluating a specific method, we only considered the barcoded droplets that were consistently classified by other two methods and evaluated how many droplets agreed with the consensus classification from the other two methods, excluding ambiguously classified droplets. We observed that while 94.9% of droplets consistently classified as multiplets by both *freemuxlet* and *demuxlet* were also classified as multiplets as cell hashing methods, only 89.1% of droplets classified as singlets by both *freemuxlet* and *demuxlet* were classified as singlets in cell hashing, suggesting that cell hashing is probably overcalling multiplets (Figure 4.2B). When we evaluated *demuxlet* using the consensus call between *freemuxlet* and cell hashing in a similar manner, the concordance for multiplets and singlets were 98.7% and 99.1%, respectively, suggesting that the inference from *demuxlet* is highly reliable. Finally, when we evaluated *freemuxlet* compared to other two methods, we observed that singlets have comparable accuracy with *demuxlet* (99.2%), but the accuracy of multiplets were 96.4%, suggesting that *freemuxlet* may misclassify a fraction of multiplets as singlets, even though the accuracy was better than cell hashing.

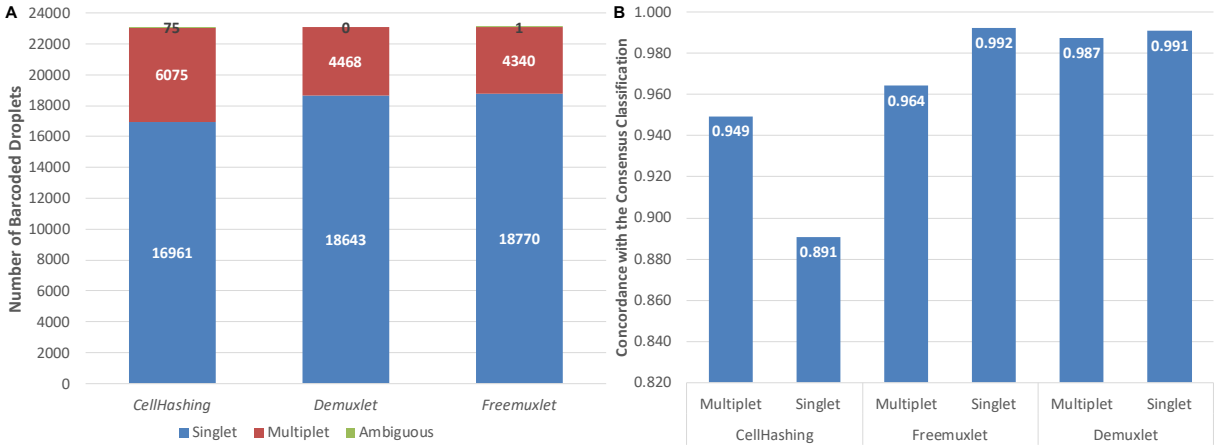

Figure 4.2. Comparison between Cell Hashing, *freemuxlet*, and *demuxlet* in 7-way mixture experiments. (A) The marginal number of barcoded droplets classified as singlets, multiplets or ambiguous droplets by cell hashing, *freemuxlet*, and *demuxlet* methods among 23,111 barcoded droplets. The number represents the number of barcoded droplets that belongs to the colored category. **(B)** Comparisons between each of the method vs. consensus classification between the other two methods. The x-axis represents different methods, and the consensus classification of droplets from the other two methods, and the y-axis represents the fraction of droplets that were classified concordantly with the consensus classification. Only the droplets that were consistently classified by the other two methods was used as the denominator, and the numerator is the number of droplets that were consistently classified all three methods.

Pairwise Genetic Distance between Droplets Based on Bayes Factor

One of the key metrics *freemuxlet* uses for clustering each barcoded droplet into their originating samples is the pairwise genetic distance between individuals defined as a Bayes Factor (See Methods). To evaluate how informative the Bayes Factors are, we used the pairwise genetic distance matrix between every pair of droplets as an input to generate 2-dimensional manifold using the UMAP[70,71] and tSNE[72] methods, and colored each droplet with the best-guess classification from each of the method (Figure 4.3).

Overall, we observed that droplets that are classified as singletons clearly belong to individual clusters while doublets tend to be located at the boundary of clusters (Figure 4.S2 and Figure 4.S3). The ambiguous assignment of droplets in cell hashing methods appeared to be singlets in most cases. Even though *freemuxlet* does not use sophisticated clustering algorithm or manifold algorithms, these results clearly show that the clusters generated by *freemuxlet* is consistent to the manifold generated by more sophisticated algorithms, and also that Bayes

Factors are useful input metrics to recapitulate the cluster assignments when embedded onto a low-dimensional manifold such as UMAP or tSNE.

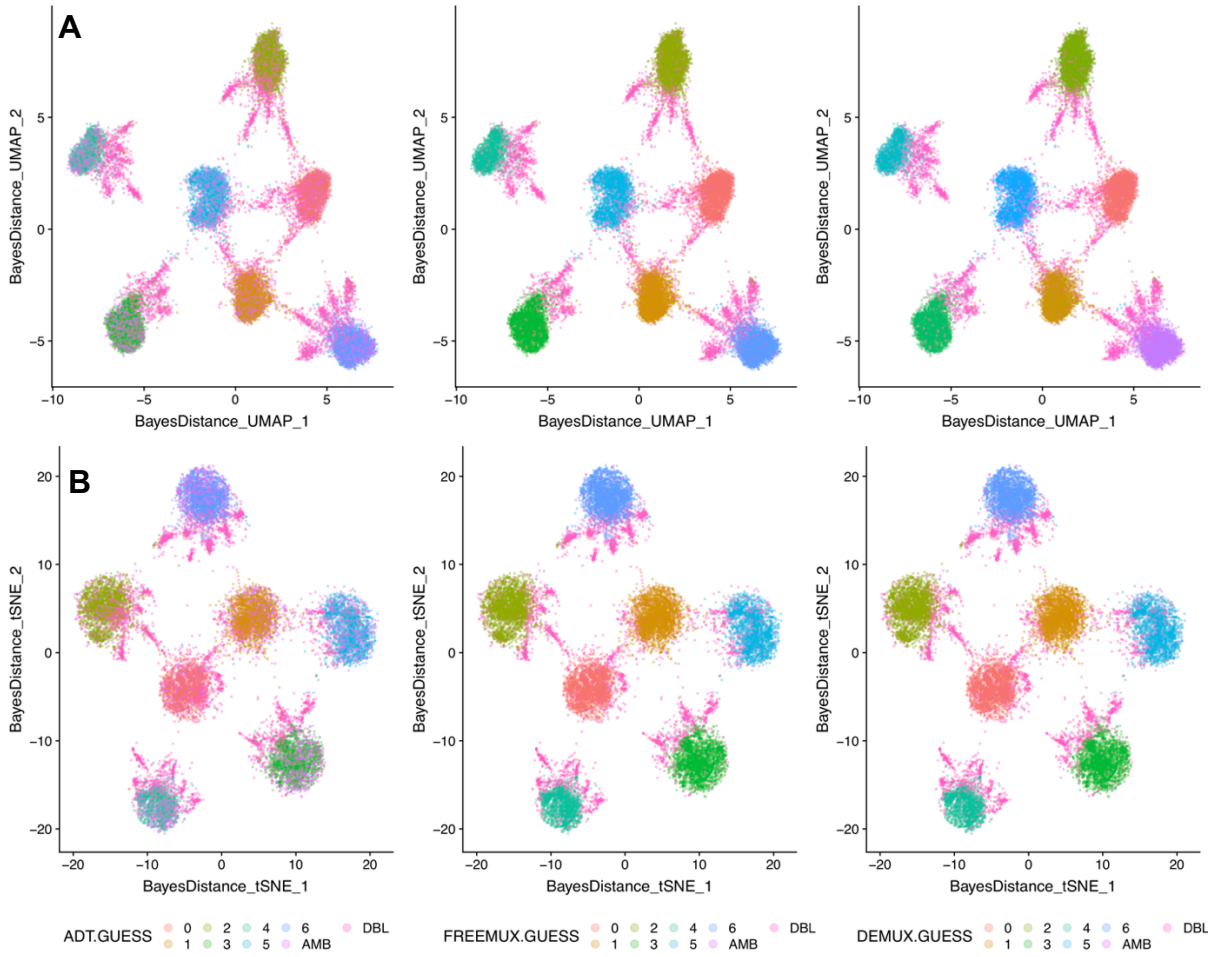

Figure 4.3. 2-dimensional manifold plots of 23,111 droplets using (A) UMAP and (B) tSNE based on the pairwise genetic distance between droplet(defined as the Bayes Factor). Droplets assignment based on cell-hashing (ADT.GUESS), *freemuxlet* (FREEMUX.GUESS), and *demuxlet*(DEMUX.GUESS) are shown in each panel. The UMAP coordinates generated based on the Bayes Factor genetic distance. Each color represents an individual, pink dots represent doublets (DBL) or ambiguous droplet assignment (AMB).

Accuracy of Genotypes Inferred from scRNA-seq Reads

To further evaluate clustering performance of *freemuxlet*, we compared the genotype called from *freemuxlet* and that from chip-based genotyping of the same individuals. (Figure 4.4) We observe that most of the genotype calls fall into a high-confidence area and the empirical distribution of genotype accuracy is consistent with genotype probability, suggesting *freemuxlet* can accurately call genotypes of each individual and in return supports our claim that *freemuxlet* can generate high-quality droplets clustering result based on Bayes Factor distance. For example, when using posterior probability threshold 0.9, 0.95, 0.99, and 0.999, the empirical accuracies are 0.89, 0.93, 0.97, and 0.94. Because a fairly large number of confident genotypes demonstrate inaccuracies, we introduced additional parameter $\epsilon = 0.1$ to account for genotype error.

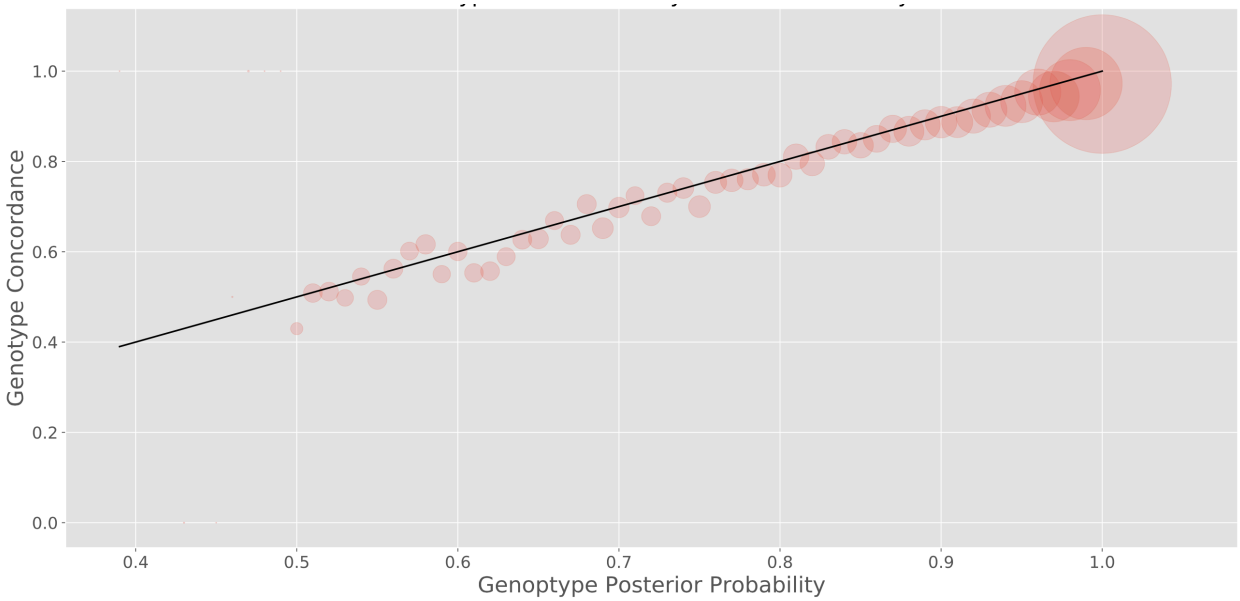

Fig. 4.4. Genotype accuracy of each cluster evaluated based on array-based genotyping of Cell Hashing dataset. The X-axis is genotype posterior probability calculated from *freemuxlet*, and Y-axis is genotype concordance by comparing most likely genotype with array-based genotype. The size of each circle represents the number of variants that has the particular genotype probability. The black line is y=x diagonal.

Application to Cancer Cell-line Mixtures

We applied our preliminary implementation of *freemuxlet* algorithm to the 12,557 droplets sequenced using 3-way mux-seq of colon cancer cell lines, generated using Drop-seq. Our original intention was to demultiplex these cell lines using *demuxlet*, using the publicly available genotypes available at COSMIC[73] or CCLE[74] database. However, we realized that the quality of genotypes of these cancer cell lines was not high due to multiploidy and the lack of accurate genotype calling algorithm designed for cancer cell line data. We therefore applied our preliminary *freemuxlet* implementation and identified 11,361 (90.5%) singlets and 1,1196 (9.5%) doublets with no ambiguous assignment. When we clustered cells, purely based on the expression levels (with *Seurat* software tool)[75] of these 11,361 singlets showed 99.97% concordance with those identified from *freemuxlet.* (Figure 4.5)

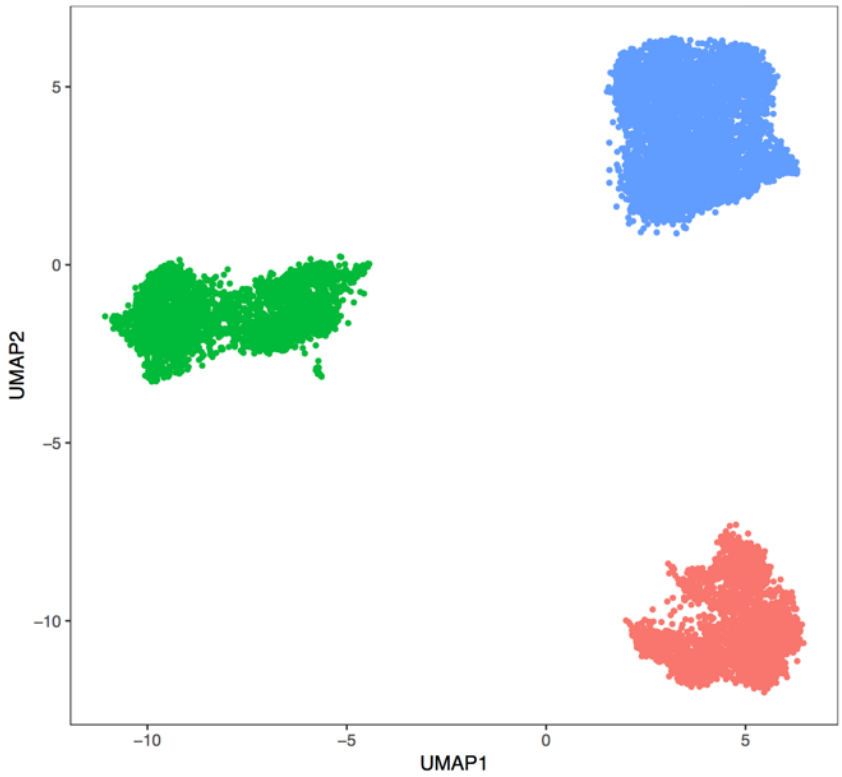

Figure 4.5. UMAP visualization (based on expression levels) of 11,361 cells sequenced with Drop-seq across 3 colon cancer cell lines, after removing 1,196 doublets inferred by *freemuxlet***.** Different colors represent different individual estimated from freemuxlet, which is >99.9% concordant to the identity of cell lines inferred from expression level.

Discussion

In this Chapter, we presented *freemuxlet,* a novel genotyping-free method to deconvolute genetically multiplexed single-cell RNA-seq reads while detecting and removing multiplets. The *freemuxlet* makes use of two different types of Bayes Factors, the single score for each barcoded droplet, and the genetic distance between a pair of droplets to determine how likely a barcoded droplet contains a single cell, or originated from the same sample, respectively. The accuracy of our greedy clustering algorithm guided by these Bayes Factors, combined with iterative refinement in the subsequent steps produces highly accurate multi-class clusters of barcoded droplets based on their sample identities.

The main advantage of *freemuxlet* is that it does not require external genotyping of each multiplexed individual. This additional benefit will enable mux-seq to be further applied in research areas where genotype is difficult or infeasible to obtain, including model organisms, cancer cell lines. The examples of cancer cell line mixture clearly demonstrate such an advantage. It will also eliminate time-consuming steps to collect, process, impute, and match external genotypes, and make the overall *mux-seq* workflow much more seamless.

Although *freemuxlet* demonstrates much higher estimated accuracy than the cell hashing method, its accuracy is slightly lower than *demuxlet*. This may be partly due to the unavailability of external genotypes, but the accuracy can still be improved by further improving the model. We observed that the estimated fraction of multiplets is sensitive to input parameter settings such as assumed genotype error rates in each cluster. Moreover, there are many potentially important factors not currently modeled in *freemuxlet* and *demuxlet* methods, such as allelic-specific expression, burst effect, or cell-type-specific expressions to further improve the accuracy of both

methods. The clustering algorithm currently implemented using a greedy method guided by Bayes Factors can also be improved. To the best of our knowledge, there is no robust clustering algorithm currently developed when each object has different priorities (e.g. singlet scores) or when multiplets exist, and this may open a new door to developing a more general method for the clustering problems that share these features.

Even though the likelihood model of *freemuxlet* is not perfect, it provides us with successful results in deconvolution of genetically multiplexed scRNA-seq reads in practice. It should be possible to extend *freemuxlet* or *demuxlet* beyond scRNA-seq, such as single nucleus RNA-seq (snRNA-seq) or single cell ATAC-seq (scATAC-seq). The main challenges of these new types of data will be increased sparsity. It will particularly be more challenging to make *freemuxlet* as accurate as in scRNA-seq when relatively a small number of sequence reads share reads at the same variant sites between barcoded droplets. Further improvements of the method and evaluations on simulated and real dataset are needed to ensure that these methods can be extended to additional types of single cell sequence reads.

Materials and Methods

Summary of *freemuxlet* Algorithm

We developed a new method, *freemuxlet*, to enable genetic demultiplexing and doublet detection of multiplexed scRNA-seq reads without requiring external genotypes. When external genotypes exist, by evaluating the conditional genotype likelihood of each droplet given external genotype of candidate sample/cluster, the deconvolution problem is an instance of a multiclass classification problem. However, without external genotypes, the true genotype of each candidate

cluster follows probability distribution governed by cluster constitution. The deconvolution problem becomes a more complicated unsupervised clustering problem. The key idea of *freemuxlet* is to cluster barcoded droplets based on the pairwise genetic distance between each pair of droplets, where each cluster represents a sample rather than an individual cell type. By aggregating droplets in the same cluster, we can in return evaluate the genotype likelihoods representing the clustered individual. Similar to *demuxlet*, we then assign each barcoded droplet to the closest cluster as a singlet or classify them as a multiplet.

Detailed Algorithms

Specifically, we first sort each droplet based on "Singlet Score" in descending order. Droplets with higher "Singlet Score" will be preferably assigned to a cluster (essentially used to define clusters). Each Droplet will be voted by previously assigned droplets to include or exclude from the defined clusters based on genetic distance between the current droplet and the previously assigned droplet (Algorithm 1). Next, *freemuxlet* randomly initialize droplet order, and vote each droplet by all the other droplets iteratively (Algorithm 2). Then, in cluster refinement steps, droplets belong to the same cluster will be merged. Genotype likelihood of each cluster will be recalculated to identity droplet membership and to detect multiplets. We iteratively repeat these cluster refinement steps to evaluate the genotype likelihoods and infer the sample identity of the droplets for a fixed number of iterations or the inferred identities of each barcoded droplet no longer changes over iteration (Algorithm 3). (Figure 4.S1)

Algorithm 1: InitCluster Initialize the cluster membership of each droplet based on Singlet Score and Genetic Distance

Input: Number of Cluster n;
An array of droplet index, $A = \{a_1, a_2, \ldots, a_m\}$, that is descending sorted based on Singlet Score;
A 2-D Genetic Distance matrix $D_{m \times m}$
Output: Cluster membership of each droplet

```
Clust = {0, 0, ..., 0} where |Clust| = m
for i ← 1 to m do
    Vote = {0, 0, ..., 0} where |Vote| = n
    for j ← 1 to i − 1 do
        if D(a_i, a_j) < threshold then
            Vote[Clust[a_j]] = Vote[Clust[a_j]] + 1
        else
            Vote[Clust[a_j]] = Vote[Clust[a_j]] − 1
    maxVote = 1
    for i ← 1 to n do
        if Vote[i] > maxVote then
            maxVote ← i
    Clust[a_i] = maxVote
return Clust
```

Algorithm 2: Clustering Refine initial clustering based on Genetic Distance

Input: Number of Cluster n;
A array $Clust$ indicates initial clustering membership of each droplet;
A 2-D Genetic Distance matrix $D_{m \times m}$
Output: Refined clustering membership of each droplet

```
for iter ← 1 to 10 do
    for i ← 1 to m do
        Vote = {0, 0, ..., 0} where |Vote| = n
        for j ← 1 to m do
            if D(a_i, a_j) < threshold then
                Vote[Clust[a_j]] = Vote[Clust[a_j]] + 1
            else
                Vote[Clust[a_j]] = Vote[Clust[a_j]] − 1
        maxVote = 1
        for i ← 1 to n do
            if Vote[i] > maxVote then
                maxVote ← i
        Clust[a_i] = maxVote
return Clust
```

Algorithm 3: CLASSIFY Further refine cluster membership and detect doublet

Input: Number of Cluster n;
A array $Clust$ indicates initial clustering membership of each droplet;
A array $C = \{c_1, c_2, \ldots, c_m\}$ contains sequence data of each droplet
Output: Refined clustering membership of each droplet;
Cluster-wise genotype probability

```
for iter ← 1 to 10 do
    for s ← 1 to n do
        MergeDropletsInCluster(s);
        CalculateGenotypeProbabilityOfCuster(s);
    for i ← 1 to m do
        llk = 0_{n×n};
        for j ← 1 to n do
            for k ← 1 to j do
                if j == k then
                    llk(j, k) = L(c_i|j)
                else
                    llk(j, k) = L(c_i|j, k; α = 0.5)
        (s_1, s_2) = argmax_{j,k} llk;
        if s_1 == s_2 then
            Clust[i] = s_1;
        else
            RemoveDoublet(i);
return Clust
```

Likelihoods of Singlets and Doublets

Let $s \in \{1,2,\cdots,S\}$ be a sample index, $c \in \{1,2,\cdots,C\}$ be the index of barcoded droplets, and $v \in \{1,2,\cdots,V\}$ be the index of genetic variants considered. Let $d_{cv} \in \{0,1,\cdots\}$ be the depth of sequence reads from droplet c overlapping with variant v, and let $b_{cvi} \in \{0,1,2\}$ be the allele of i-th sequence read ($i \in \{1,2,\cdots,d_{cv}\}$) consistent to the reference allele (0), the alternate allele (1), or other alleles (2), and $q_{cvi} \in \{0,1,\cdots\}$ be the phred-scale base quality score[76]. If $g_{sv} \in \{0,1,2\}$ and $e_{cvi} \in \{0,1\}$ are latent variables representing true underlying genotypes, the event of a sequencing error, $\Pr(b_{cvi}|g_{sv},e_{cvi})$ is assumed to follow the distribution widely used in other studies (Table 4.S1), and $\Pr(e_{cvi})$ follows $\text{Bernoulli}\left(10^{-\frac{q_{cvi}}{10}}\right)$.[13,30] Then, the probability of allelic read given genotype is $\Pr(b_{cvi}|g_{sv}) = \sum_{e_{cvi}=0}^{1} \Pr(b_{cvi}|g_{sv},e_{cvi}) \Pr(e_{cvi})$.

Under the assumption that all the reads in barcoded droplet c were originated from sample s (i.e. c is a singlet from s), the probability of sequence reads can be modeled as

$$L_1(c) = \prod_{v=1}^{V}\left[\sum_{g_v=0}^{2}\left\{\Pr(g_v)\prod_{i=1}^{d_{cv}}\Pr(b_{cvi}|g_v)\right\}\right]$$

where $\Pr(g_v)$ is the probability of the unobserved genotype g_v. In *demuxlet*, we assumed that $\Pr(g_v)$ is given based on the external genotypes (based on the posterior probability of imputed genotypes or best-guess genotypes with predefined error rates). In the initial steps of *freemuxlet*, we model $\Pr(g_v)$ assuming $g_v \sim \text{Binomial}(2, f_v)$ where f_v is allele frequency of variant v. Under the assumption that the reads in B_c were originated from two samples with mixing proportion $(1-\alpha):\alpha$ is

$$L_2(c,\alpha) = \prod_{v=1}^{V}\sum_{g_{v1},g_{v2}}\left\{\Pr(g_{v1})\Pr(g_{v2})\prod_{i=1}^{d_{cv}}\Pr(b_{cvi}|g_{v1},g_{v1};\alpha)\right\}$$

where $\Pr(b_{cvi}|g_{v1},g_{v1};\alpha) = (1-\alpha)\Pr(b_{cvi}|g_{v1}) + \alpha\Pr(b_{cvi}|g_{v2})$.

Singlet Score of Each Barcoded Droplet

We define "singlet score" (SS) of each barcoded droplet as a log Bayes Factor between the singlet and doublet likelihood as follows.

$$SS(c) = \log\left[\frac{L_1(c)}{L_2(c, \alpha = 0.5)}\right]$$

The singlet score informs whether the scRNA-seq reads from a barcoded droplet is likely singlets or doublets without requiring external genotypes. The likelihood models used to obtain single score is identical to the models used for detecting sample contamination from DNA sequence reads[13], except that α is fixed to 0.5.

For droplets with relatively few reads, SS_c may not be much informative and the value will be close to zero. For droplets with larger read counts, SS_c will becomes more informative. When performing clustering, we sort barcoded droplets based on decreasing orders of SS_c so that putative singlets are clustered first, so that doublets will have less chance to confound the clustering results.

Genetic Distance between Droplets

For each pair of the droplets (c_1, c_2), we evaluate the probability that both barcoded droplets are originated from the same individual in the following model.

$$L_1(c_1, c_2) = \prod_{v=1}^{V}\left[\sum_{g_v=0}^{2}\left\{\Pr(g_v)\prod_{i=1}^{d_{c_1 v}}\Pr\left(b_{c_1 vi}|g_v\right)\prod_{i=1}^{d_{c_2 v}}\Pr\left(b_{c_2 vi}|g_v\right)\right\}\right]$$

Similarly, the probability that the pair of droplets are originated from different individuals can be modeled as follows.

$$L_2(c_1, c_2) = \prod_{v=1}^{V}\sum_{g_{v1}, g_{v2}}\left\{\Pr(g_{v1})\Pr(g_{v2})\prod_{i=1}^{d_{c_1 v}}\Pr\left(b_{c_1 vi}|g_{v1}\right)\prod_{i=1}^{d_{c_2 v}}\Pr\left(b_{c_2 vi}|g_{v2}\right)\right\}$$

Note that neither of these models accounts for doublets. These models assume that both droplets consist of singlets and aim to determine whether they share the sample identity samples or not. We define genetic distance between pair of cells using the following log Bayes Factor.

$$D(c_1, c_2) = \log\left[\frac{L_2(c_1, c_2)}{L_1(c_1, c_2)}\right]$$

Positive $D(c_1, c_2)$ values suggest that the two droplets are likely originated from different individuals and negative $D(c_1, c_2)$ values suggest that the two droplets are likely originated from the same individual, under the assumption that both of them are singletons.

Initial Clustering

A good initial clustering assignment will provide accurate genotype probability of clusters, reducing the time to converge to the optimal solution within a limited number of iterations. While several clustering algorithms such as Smart Local Moving(SLM) algorithm based on Louvain's method[77] is one of the widely used methods, there are important limitations that these algorithms may not robustly cluster barcoded droplets by their sample identities. First, we expect that a large proportion of barcoded droplets are doublets, but existing algorithms do not assume that there will be objects that belong to multiple samples. Second, the single scores $SS(c)$ should be helpful to inform whether c should belong to a cluster or considered as doublets, but existing algorithms do not have room to incorporate such information.

We found that applying a greedy clustering algorithm, ordering each droplet from the highest $SS(c)$ to the lowest highest, and sequentially assigning clusters based on majority vote based on $D\left(c_{current}, c_{prev}\right)$ where c_{prev} is the droplets whose clusters have already been assigned. We used a specific threshold, corresponding Bayes Factor p-value of 10^{-3} in either direction, to account for the majority vote procedure. If all of the possible clusters result in

negative total votes, the droplet may be assigned to a new cluster, as long as the total number of clusters did not reach to the number of multiplexed samples in the scRNA-seq data.

Iterative Refinement of Clusters

The initial clustering procedure does not classify a droplet as doublets, and some of the singlet assignments may be incorrect. We refine initial clustering results with an iterative procedure of estimating the consensus genotype probability of each cluster and updating the best-guess of the cluster assignment of each barcoded droplet based on the consensus genotypes.

At each iteration, we aggregate all reads from droplets that are assigned to a particular cluster and calculate genotype probability. The genotype probability is calculated from allele frequencies and the genotype likelihoods from sequence reads. Suppose that, at a certain iteration, cluster k consists of a set of droplets represented as C_k. The genotype likelihood, given genotype g_v, is calculated using the standard form $\prod_{c \in C_k} \prod_{i=1}^{d_{cv}} \Pr(b_{cvi}|g_v)$ assuming independence between reads. The posterior probability can be calculated using Bayes rule, using $\Pr(g_v)$ as the prior.

With updated genotype probability, we have $\widehat{\Pr}(g_{sv})$ for each sample (cluster) s, and the same likelihood model used for *demuxlet* can be applied to determine whether a specific droplet is a singlet or a doublet. Specifically, we use the following model for singlets (L_1) and doublets (L_2).

$$L_1(c|s) = \prod_{v=1}^{V} \left[\sum_{g_{sv}=0}^{2} \left\{ \widehat{\Pr}(g_{sv}) \prod_{i=1}^{d_{cv}} \Pr(b_{cvi}|g_{sv}) \right\} \right]$$

$$L_2(c|s_1, s_2; \alpha) = \prod_{v=1}^{V} \sum_{g_{s_1 v}, g_{s_2 v}} \left\{ \widehat{\Pr}(g_{s_1 v}) \widehat{\Pr}(g_{s_2 v}) \prod_{i=1}^{d_{cv}} \Pr(b_{cvi}|g_{s_1 v}, g_{s_2 v}; \alpha) \right\}$$

Maximum likelihood estimates adjusting for Akaike Information Criterion across all possible L_1 and L_2 is used to determine whether c is a doublet or not, and it is removed from the cluster if determined as doublet.

At the final step to determine the assignment of a barcoded droplet to an individual cluster or a doublet, we allow genotyping errors in calculating Pr (g_{sv}). Specifically, given assumed genotyping error rate ϵ, we use $\widetilde{\Pr}(g_{sv}) = (1-\epsilon)\,\widehat{\Pr}(g_{sv}) + \epsilon \Pr(g_v)$ where Pr (g_v) is calculated by allele frequency only.

Recovering Sample Identities from Individual Clusters

Even though *freemuxlet* can robustly cluster cell barcodes into distinct clusters while detecting and filtering out multiplets, it does not automatically assign each cluster to sample identity. To connect a cluster of *freemuxlet* to individual identity, we need additional steps. *Freemuxlet* can be used in two different circumstances; (1) external genotype data (e.g. SNP array, exome sequencing, or bulk RNA-seq) are available, but *freemuxlet* was used either due to limited quality of external genotype data to apply *demuxlet* or to reduce turnaround time of the analysis; (2) no external source of genotype is available. In each circumstance, we offer solutions to resolve sample identities.

First, when external genotypes are available, our recommended approach to produce genotype calls from external genotype data and compare them with the genotype likelihood of each cluster estimated by *freemuxlet*. *Freemuxlet* produces a VCF file that contains genotype likelihood for each cluster representing a sample. This VCF file can be compared to the VCF file generated from the external genotype data. For example, in the cancer cell line experiment, even though the genotypes of cancer cell line obtained from COSMIC[73] database was not accurate enough to apply *demuxlet*, we were able to successfully match the sample identity between the *freemuxlet* cluster and COSMIC-genotyped samples to clearly distinguish which individual was by calculating Bayes Factor distance between each possible pair of matched samples.

Second, when we do not expect to have external genotypes, we can resolve sample identities by including each sample twice in a mux-seq experiment. For example, in the example shown in Figure 4.S4, each of 6 samples is included twice in 4 batches of 3-way mixtures. Because *freemuxlet* reliably identifies a pair of clusters that have the same genetic identity between batches, the identity of the individual can be resolved by pairs of batches that share an individual. In Figure 4, for example, the samples shared between batch 1 and 2 represent sample ID1, and those shared between batch 2 and 4 represent sample ID4. In general, when there are b possible batches of experiments, the number of samples to be included in the design should is b(b-1)/2, and each batch should have b-1 samples to multiplex together. In this way, 45 samples can be multiplexed across 10 batches, with 9 samples per batch. Up to 120 samples can be multiplexed across 16 batches, with 15 samples per batch. In addition to avoiding the need for collecting external genotype data, this Sudoku-like pairwise design has the benefit to have replicates for each individual sample. These replicates can be used for correcting for technical batch effects, or each replicate can contain a different environmental condition to test for differential expression.

Genetically Multiplexed Cell Hashing Data

To evaluate the accuracy of *freemuxlet* compared to alternative approaches such as cell hashing or *demuxlet*, we generated scRNA-seq reads using 10x Chromium (v2 chemistry) across 23,111 barcoded droplets from peripheral blood mononuclear cells genetically multiplexed across 7 different samples, with additional molecular tagging of sample identity using antibody-based cell hashing[69] method. Their genotypes are imputed from Haplotype Reference Consortium (HRC) panel, and 217,411 exonic SNPs with minor allele frequency (MAF) >1%

were used as the source of external genotypes when running *demuxlet*. The sample identities inference from cell hashing data were inferred using the recommended settings[75]. For *freemuxlet*, we used 241,322 exonic SNPs from the 1000 Genomes Phase 3 genotypes, filtering by European MAF > 1%. We used European MAFs as the source of allele frequencies. The UMAP and tSNE manifolds were created using the default settings. Posterior probabilities (GP field) were used to run *demuxlet*, and the assumed genotype error probability was $\epsilon = 0.1$ for both *demuxlet* and *freemuxlet*.

Mixture of Cancer Cell Lines

We also generated a mixture of three colon cancer cell lines – RKO, HCT116, and SW480 – under various environmental conditions across 10 batches using the Drop-seq[59] technique. We obtained a total of 12,557 droplets with 800 or more unique reads, and demultiplexed sample identities using *freemuxlet* while filtering out 1,196 droplets predicted to be doublets or ambiguous droplet. We used the remaining singlets to cluster the cells based on the gene expressions with UMAP[71] based on top 100 principal components with Seurat[75]. The UMAP clearly identified three clusters of cell types and the identity was unequivocally inferred from the UMAP manifold. The genetic identities of *freemuxlet* samples were determined by comparing the likelihood of COSMIC genotypes with respect to the sequence reads for each cluster. The genetically inferred sample identities were compared with the identities inferred by gene expressions to evaluate the concordance.

Chapter V.

Discussion

Summary of the Chapters

The importance of genetic variants in association studies, population genetic studies, or expression quantitative loci (eQTL) studies cannot be stressed more. However, in the context of upstream analysis of ultra-high-throughput sequencing studies, such as large-scale genome sequencing or single cell genome sequencing, the importance of genetic variants is currently underappreciated, and it deserves further exploration. In my dissertation, I described novel methods that can robustly and rapidly generate QC metrics to aid high-quality data production, methods that can jointly estimate sample contamination and genetic ancestry, as well as a method that enables genotyping-free deconvolution of genetically multiplexed scRNA-seq reads. I will summarize each of these methods and discuss their limitations and remaining opportunities.

Rapidly generating comprehensive quality control (QC) metrics for high-throughput sequencing experiments in the early stage of data production is crucial for all the omics studies because timely feedback of potential problems can help avoid further loss. In Chapter 2, I developed *FASTQuick* to substantially reduce the turnaround time of comprehensive QC of raw sequence reads by focusing on sequence data alignable in genomic regions around known genetic

variants while maintaining highly concordant estimation of QC metrics with methods that require sequence data to align against the full reference genome (post-alignment method).

FASTQuick combined advantages of 1) spaced k-mer hash tables[78], which can rapidly filter out negative hits at O(1) time complexity, 2) BWT-based alignment methods[29], which efficiently align the positive hits, and 3) tailored statistical methods to robustly estimate comprehensive QC metrics from partially aligned reads by extrapolating the metrics to genome-wide. *FASTQuick* not only provide "genomic" QC metrics comparable to widely used post-alignment methods such as QPLOT or Picard but also produces "genetic" QC metrics that require genetic variant information such as DNA contamination of genetic ancestries. Compared to other post-alignment methods, *FASTQuick* not only delivers more comprehensive QC metrics (Table 2.1) but also reduce the QC turnaround time by >50-fold. Experiments on 38x coverage genomes show that *FASTQuick* can reduce the QC turnaround time from ~160 hrs. to ~1 hr.

In Chapter 3, I addressed the known problem that incorrectly specifying allele frequency can introduce substantial bias in the estimation of contamination rate[13]. I developed *verifyBamID2* that can accurately estimate the contamination rate regardless of genetic ancestry. The key idea is to introduce individual-specific allele frequency (ISAF) as a function of genetic ancestry in the PC space and model the likelihood of sequence reads with ISAF and a mixture-model-based genotype likelihood function to jointly estimate the genetic ancestry and contamination together. The ability of *verifyBamID2* to estimate contamination rate in an ancestry-agnostic way allows us 1) to simplify the genome analysis pipeline by eliminating the complication of having to specify sample-specific parameter settings in large-scale studies, 2) to estimate contamination rates robustly against potentially misspecified population allele frequencies of the intended samples, 3) to infer genetic ancestries for both intended and contamination samples to track down potential source of

the contamination, and 4) to more accurately infer genetic ancestries in the presence of contamination. Through experiments with in-silico contaminated samples, I confirm that the estimation of contamination rate will be biased if genetic ancestry is incorrectly specified (Figure 3.4), and the estimation of genetic ancestry will be biased if contamination is ignored (Figure 3.3).

The availability of genetic ancestry and contamination estimates in QC procedure enables many tailored analyses in variant calling steps. For example, in the TOPMed variant calling pipeline that recently called >140,000 genomes jointly, the genetic ancestry and contamination estimates are used to perform more accurate genotyping and produce more informative population-level summary statistics such as Hardy-Weinberg equilibrium test statistics account for genetic ancestries. The *verifyBamID2* software tool can also be used solely as the tool to estimate genetic ancestry in lieu of alternative methods such as LASER[21], regardless of the presence of potential contamination due to its higher accuracy (Figure 3.3).

Individual-specific allele frequency (ISAF) is the bridge that connects genetic ancestry parameters to *verifyBamID2*'s unified genotype likelihood model[31,32]. The idea behind ISAF is that allele frequencies can be represented as a function of PC coordinates. Other methods in studying population structure used a similar idea[31,32], but for the first time, we introduce ISAF into a unified likelihood-based model to model the sequence data under population structure and/or contamination. My results show that *verifyBamID2* robustly estimates the contamination rate more accurately than the previous method *verifyBamID1*[13] across different configuration of contamination scenarios (Figure 3.4).

Contrary to unexpected contamination (or a mixture) occurring in sequence data, a mixture between different samples in sequence data can also be intentionally introduced to facilitate cost-effective single-cell RNA-seq studies via genetic multiplexing (mux-seq). Mux-seq significantly

reduces per-sample and per-cell library preparation cost by increasing single cell loading throughput while maintaining acceptable doublet rate by detecting and removing doublet with the help of *demuxlet* using external sample genotype. In Chapter 4, I developed genotyping-free single-cell RNA-seq deconvolution method, *freemuxlet*, to enable mux-seq experiment without requiring external genotyping. The *freemuxlet* iteratively applies multi-class clustering of sequenced droplets guided by Bayes Factor distances, and simultaneously estimate the consensus genotypes of each cluster representing a genetically distinct individual. Meanwhile, *freemuxlet* also detect and remove multiplets using a similar strategy to *demuxlet*.

In addition to inheriting advantages of the *demuxlet* that enables cost-effective scRNA-seq experiment with reduced batch effects, *freemuxlet* also removes the requirement of external genotyping of each individual. This additional benefit will enable mux-seq to be further applied in research areas where genotype is difficult or infeasible to obtain, including model organisms, cancer cell lines. It will also facilitate seamless analysis of scRNA-seq experiment generated by mux-seq flow, by eliminating the time-consuming steps to collect external genotypes and to process and impute the genotype data. The evaluation shows that *freemuxlet* generates highly concordant estimates of droplet identities inferred by other methods such as cell hashing and *demuxlet*[26]. I further demonstrated that *freemuxlet* could successfully deconvolute multiplexed scRNA-seq of cancer cell lines, where *demuxlet* could not deconvolute the droplet identities due to the difficulty of accurately genotyping cancer cell lines.

Remaining Challenges and Future Directions

In this dissertation, each of the methods presented in Chapter 2, 3, and 4 demonstrated results that are comparable or better than existing tools in the evaluation of simulated and real

data, with clear advantages in at least one of the evaluation criteria such as computational efficiency, robustness, or accuracy. At the same time, these methods still have plenty of rooms for further improvements or extension to additional data types to make the methods even more useful to the scientific community.

In Chapter 2, I demonstrated that the ultra-fast speed of *FASTQuick* could dramatically reduce the turnaround time of quality assessment on raw sequence data stored in the standard FASTQ format. *FASTQuick* can be even more useful if it interfaces with Illumina's software tools process their own proprietary BCL format. For example, the computational time to run *FASTQuick* on a 38x genome is much faster than the time spent to convert BCL files into FASTQ files using the *bcl2fastq* software tool (~5 hours). Therefore, in principle, *FASTQuick* can work as a UNIX pipe during the conversion procedures between file formats required for most sequence analysis pipeline, without increasing the turnaround time at all. The implementation of such a procedure, however, requires additional parameters that allow *FASTQuick* to seamlessly interface with *bcl2fastq* in various parameters settings such as demultiplexing sample indices and parallelization.

I observed that the computational turnaround time of *FASTQuick* does not linearly decrease by the number of threads. In further evaluations, we identified the input/output (I/O) overhead, which includes disk I/O and the computational cost to compress and decompress the data was the major bottleneck that affects the overall turnaround time. For these reasons, using *FASTQuick* as a part of a UNIX pipeline is even more attractive in implementing large-scale sequence analysis pipeline. For even further speedup, it should be possible to run *FASTQuick* on the BCL or FASTQ files that are split into multiple pieces, to collect the sufficient statistics of QC metrics (such as pileups and summary statistics) in parallel, and to merge them to produce

combined QC metrics. The last step to combine QC metrics for a deeply sequenced genome should take only a few minutes.

Currently, *FASTQuick* is focused on processing ultra-high-throughput whole genome sequence data, although the metrics can also be useful in other types of sequencing. The QC metrics will be even more useful if FASTQuick is tailored for various types of sequencing. For example, for exome or targeted sequencing, it would be useful to estimate the fraction of reads alignable to the targeted regions and compute QC metrics focusing on the reads located in the target regions. Similarly, for transcriptomic and epigenomic sequence reads, having QC metrics focused on expressed genes or specified regulatory elements will be useful to understand the quality of the sequence reads with respect to the expectation according to the type of sequence data.

In Chapter 2, we focused on the scenario where sequence reads are relatively short, usually ~100bp to ~250bp. As the sequencing technologies develop, reads length will become longer and longer[79,80]. The algorithm used in *FASTQuick* can be extended to incorporate features from long sequence alignment algorithms, such as *minimap2*[38], to broaden the range of applications. The initial filtering of negative hits can be more robust against potentially high sequencing errors in certain technologies such as PacBio sequencing, by allowing more mismatches per k-mer or by accounting for the local distribution of hits and misses in spaced k-mer hashes across long reads, rather than filtering based only on the overall number of hits or misses. Another limitation is that the reads alignment procedure in *FASTQuick* is still assuming that the reference genome has a linear structure. As more and more polymorphisms in human genomes across different populations are characterized, graph-based alignment algorithm[81] may become the *de facto* standard for alignment algorithms, and *FASTQuick* can still be used in such

context by incorporating the new alignment algorithm in lieu of the current BWT-based algorithm.

In Chapter 3, I demonstrated that *verifyBamID2* substantially reduces the potential bias of contamination estimates compared to the original *verifyBamID* due to its ability to jointly estimate genetic ancestry and robustly account for the ancestry in estimating contamination rates. The presented results only show scenarios where intended contamination rate >1%. This is because the 1000 Genomes sequence reads used <3% as the threshold to exclude contaminated samples, so the sequenced reads of certain samples are possibly contaminated at a low level. In addition, when the contamination rate is extremely small, its impact on genotyping accuracy is likely small as demonstrated previously[14]. Nevertheless, in further exploratory experiments, I observed that *verifyBamID2* could accurately estimate the contamination rate as low as 0.1%. At such a very low level of contamination, the convergence properties may not be as robust. By modifying the numerical optimization algorithm to incorporate the gradient of the likelihood function, I believe that the accuracy of *verifyBamID2* can be improved to be more sensitive to the extremely small level of contamination or in the case where much more parameters (e.g. using 10 PCs) need to be estimated.

Although the performance is evaluated based on DNA contamination scenario, *verifyBamID2* is not limited to DNA sequence data. For example, in our experience, *verifyBamID2* is able to detect and estimate contamination for RNA-seq and epigenomic sequence data, even though the model can be further improved to increase the accuracy of estimation. The same model was also used in single cell transcriptomics[26] as shown in Chapter 4. Another context of great interest is whether *verifyBamID2* can also be extended to handle contamination in cancer genomic data. Even though contamination between tumor-normal

samples will not be detected using *verifyBamID2*, contamination from another individual should be able to be detected. However, the current model does not explicitly model ploidy changes or imbalance between the allelic reads in cancer samples, so modeling these factors may improve the accuracy of the method for cancer genome sequence reads.

Sample swap is also an important topic related to sample contamination. Sample swaps cannot be detected with additional information such as reported genetic ancestry, sex, or external genotypes. While *verifyBamID2* produces the genetic ancestry information to aid sample swap checking, it currently does not check the biological sex of the sequenced genome, and it should be possible to extend to estimate the biological sex. Also, the original *verifyBamID* has an option to check sample swaps against external genotypes. However, the feature is not currently implemented in *verifyBamID2*, but it should be straightforward to incorporate such a feature. Moreover, it should be possible to not only check the sample swaps to unrelated samples but also identify sample swaps to close relatives by incorporating a Hidden Markov Model or by estimating the probability of IBS0 (zero identity-by-state), IBS1, and IBS2 genome-wide[82]. Furthermore, the unified likelihood model used in *verifyBamID2* actually can be further extended to estimate metrics like blood type, HLA type, paternal/maternal haplogroup, polygenic risk scores, or other of highly heritable phenotypes.

In Chapter 4, although *freemuxlet* can accurately assign singlets and detect doublets, its accuracy to correctly distinguish singlets from multiplets can be further improved. Currently, in the 7-sample mixture cell-hashing data, we observed that the inferred identities between *freemuxlet* and *demuxlet* agree 97.8% of droplets. In 18,451 droplets where both methods predicted the droplet as singlets, 18,450 (>99.99%) of them were assigned to the identical sample. However, in the remaining 4,660 droplets where either method classified the droplet as a

multiplet, only 4,146 (89%) of them were consistently assigned as multiplets by both methods. These observations suggest that distinguishing multiplets from singlets is a much more challenging problem than assigning the correct identity of singlets. Moreover, we observed that the estimated fraction of multiplets is sensitive to input parameter settings such as genotype error rates.

There are many potentially important factors not currently modeled in the *freemuxlet* method. For example, allelic-specific expression or burst effect is currently not accounted for in the likelihood model of *freemuxlet*, and it is possible that the likelihood of homozygous genotype is inflated due to the unmodeled factors. In addition, *freemuxlet* currently does not attempt to infer genetic ancestry of each individual or make use of individual-specific allele frequencies as *verifyBamID2* does. Instead, it relies on externally provided allele frequency information as original *verifyBamID* did. Therefore, unmodeled bias may be affecting the distinction between singlets and multiplets due to systematic increase or decrease of observed heterozygosity than expected by the externally provided allele frequencies.

Currently, the likelihood model of *freemuxlet* (and also that of *demuxlet*) ignores the fact that different cells may have different distributions of expression levels due to the difference in cell types, cell cycles, or other conditions. For example, even if the total number of reads in a doublet is equally contributed by each contributing cell (i.e. exact 1:1 mixture), the number of reads per each gene may be substantially imbalanced if the two contributing cells are of different cell types. A more systematic way to incorporate cell-type specific expression levels should incorporate both the total number of unique reads and the allelic read information at each gene together[83]. Even though this requires an additional implementation to collect scRNA-seq information beyond genetic variants, this will be an important extension to enable more

comprehensive inference of multiplexed scRNA-seq reads with respect to both sample identities and cell types.

In summary, this dissertation contributes new methods to large-scale, ultra-high-throughput sequencing studies focusing on DNA sequencing and single-cell RNA-seq. The methods described in this dissertation enables rapid, robust, and scalable upstream analyses of massive sequence data. In particular, I show that genetic variants can play an important role in making useful inferences from upstream analysis of sequence reads or even enable new experimental designs of single-cell studies. I believe that the combination of well-designed statistical models and efficient algorithms is fundamental to enable best practices of large-scale sequencing studies and enhance our understanding of human genetics and genomics.

Appendix A

Summary Statistics Report by *FASTQuick*

Table 2.S1. Summary statistics report by *FASTQuick*

Output File Names	Description
output_prefix.AdjustedInsertSizeDist	Adjusted Insert Size Distribution
output_prefix.DepthDist	Depth distribution
output_prefix.EmpCycleDist	Empirical Base Quality V.S. Sequencing Cycle
output_prefix.EmpRepDist	Empirical Base Quality V.S. Reported Base Quality
output_prefix.GCDist	GC Content Distribution
output_prefix.InsertSizeTable	Insert Size for Each Reads Pair
output_prefix.Likelihood	Genotype Likelihood
output_prefix.Pileup	Pileup format information
output_prefix.RawInsertSizeDist	Not Adjusted Insert Size Distribution
output_prefix.Summary	General Summary Report
output_prefix.bam	Reads Alignment

Appendix B

Detailed Contamination Estimation of *in-silico* Contaminated Samples

Supplementary Table 3.S1: Mean estimated contamination rates of *in-silico* contaminated population across different intended contamination rate, populations of intended and contaminating samples, and the estimation methods.

Population		Intended % Contam.	Equal-Ancestry (VB2)	Unequal-Ancestry (VB2)	Pooled AF (VB1)	EUR AF (VB1)	EAS AF (VB1)	AFR AF (VB1)
Intended	Contam.							
GBR	GBR	1%	1.0%	1.0%	0.8%	1.0%	0.6%	0.5%
		2%	1.9%	1.9%	1.5%	2.0%	1.3%	1.1%
		5%	4.6%	4.6%	3.8%	4.7%	3.2%	2.7%
		10%	9.2%	9.2%	7.4%	9.4%	6.2%	5.2%
		20%	18.3%	18.3%	14.7%	18.5%	11.6%	9.5%
GBR	CHS	1%	1.1%	1.0%	0.9%	1.2%	0.7%	0.6%
		2%	2.1%	1.9%	1.7%	2.2%	1.5%	1.2%
		5%	5.2%	4.7%	4.3%	5.3%	3.8%	3.1%
		10%	10.1%	9.4%	8.6%	10.4%	7.6%	5.9%
		20%	19.8%	18.7%	17.3%	19.9%	15.1%	10.9%
GBR	YRI	1%	1.3%	1.1%	1.1%	1.4%	0.8%	0.7%
		2%	2.5%	2.0%	2.1%	2.6%	1.7%	1.4%
		5%	5.9%	4.8%	5.0%	6.1%	4.2%	3.6%
		10%	11.3%	9.5%	10.0%	11.7%	8.0%	7.3%
		20%	21.6%	19.1%	19.7%	22.0%	14.8%	14.6%
CHS	GBR	1%	1.2%	0.9%	0.4%	0.2%	1.2%	0.1%
		2%	2.5%	1.8%	1.1%	0.8%	2.5%	0.5%
		5%	6.1%	4.8%	3.6%	2.9%	6.3%	2.0%
		10%	12.0%	9.9%	7.9%	6.6%	12.5%	4.6%
		20%	23.0%	19.8%	16.6%	14.2%	23.6%	9.4%
CHS	CHS	1%	0.9%	0.9%	0.2%	0.1%	0.9%	0.0%
		2%	1.8%	1.8%	0.7%	0.5%	1.8%	0.2%

		5%	4.7%	4.8%	2.4%	1.9%	4.7%	1.3%
		10%	9.5%	9.5%	5.2%	4.2%	9.5%	2.9%
		20%	19.1%	19.1%	10.6%	8.4%	18.9%	5.9%
CHS	YRI	1%	1.4%	0.9%	0.5%	0.3%	1.4%	0.1%
		2%	2.8%	1.9%	1.3%	1.0%	2.8%	0.5%
		5%	6.7%	4.8%	4.1%	3.3%	6.9%	2.3%
		10%	12.9%	9.8%	8.9%	7.2%	13.5%	5.4%
		20%	24.3%	19.6%	18.6%	14.2%	24.9%	12.2%
YRI	GBR	1%	1.3%	1.0%	0.6%	0.4%	0.4%	1.4%
		2%	2.5%	1.9%	1.3%	0.9%	0.8%	2.6%
		5%	6.0%	4.6%	3.8%	2.8%	2.6%	6.4%
		10%	11.5%	9.3%	8.1%	6.2%	5.6%	12.5%
		20%	21.9%	18.8%	16.7%	13.0%	11.1%	23.5%
YRI	CHS	1%	1.3%	0.9%	0.5%	0.4%	0.3%	1.3%
		2%	2.5%	1.9%	1.3%	0.9%	0.8%	2.6%
		5%	6.0%	4.7%	3.8%	2.8%	2.5%	6.3%
		10%	11.5%	9.3%	7.9%	5.9%	5.6%	12.2%
		20%	22.0%	18.8%	16.6%	12.0%	12.1%	22.9%
YRI	YRI	1%	0.9%	0.9%	0.4%	0.2%	0.2%	0.9%
		2%	1.8%	1.8%	0.9%	0.6%	0.5%	1.8%
		5%	4.4%	4.4%	2.4%	1.8%	1.6%	4.4%
		10%	8.8%	8.8%	5.1%	3.8%	3.4%	8.9%
		20%	17.6%	17.6%	10.1%	7.3%	6.6%	17.9%

Equal-Ancestry Model: Estimate from *verifyBamID2* assuming intended and contaminating samples have the same genetic ancestry (in PC coordinates)

Unequal-Ancestry Model: Estimate from *verifyBamID2* allowing intended and contaminating samples to have different genetic ancestry

Pooled AF: Estimate from original *verifyBamID* using allele frequency across all 1000 Genomes phase 3 samples

EUR AF: Estimate from original *verifyBamID* using allele frequency across European subset of 1000 Genomes phase 3 samples

EAS AF: Estimate from original verifyBamID using allele frequency across East Asian subset of 1000 Genomes phase 3 samples

AFR AF: Estimate from original verifyBamID using allele frequency across African subset of 1000 Genomes phase 3 samples

Performance on Admixed Population

Supplementary Table 3.S2: Average of estimated contamination rates across 10 *in-silico* contaminated samples from Mexican population under different models. Results are similar as Europeans, except that unequal-ancestry model slightly reduces estimated contamination rate from equal-ancestry model, unlike GBR.

Intended % Contamination	Equal-Ancestry (VB2)	Unequal-Ancestry (VB2)	Pooled AF (VB1)	EUR AF (VB1)	EAS AF (VB1)	AFR AF (VB1)
1%	1.1%	1.0%	0.8%	1.0%	0.6%	0.3%
2%	2.1%	2.1%	1.6%	2.0%	1.4%	0.9%
5%	4.8%	4.8%	3.9%	4.6%	3.5%	2.5%
10%	9.3%	9.2%	7.8%	8.8%	6.8%	4.9%
20%	18.5%	18.3%	15.4%	17.0%	13.0%	9.4%

Effect from Different Size of Marker Set

Supplementary Table 3.S3: Comparison of mean contamination rate ratio (Estimated/Intended) using different size of marker set (under Unequal-Ancestry Model). The Numbers in parenthesis represent standard deviation.

Sample Population		Marker Set	Intended Contamination Rate				
Intended	Contam.		0.01	0.02	0.05	0.1	0.2
GBR	GBR	1K	0.57(0.15)	0.88(0.38)	0.87(0.28)	0.92(0.18)	0.95(0.12)
		10K	0.98(0.13)	0.95(0.11)	0.93(0.09)	0.91(0.08)	0.91(0.07)
		100K	1.00(0.10)	0.96(0.09)	0.93(0.08)	0.92(0.06)	0.91(0.05)
CHS	CHS	1K	1.38(1.26)	1.09(0.63)	1.00(0.44)	0.95(0.41)	0.95(0.21)
		10K	1.08(0.48)	1.03(0.26)	1.00(0.12)	1.01(0.08)	0.96(0.06)
		100K	0.89(0.12)	0.92(0.08)	0.95(0.07)	0.95(0.05)	0.96(0.04)
YRI	YRI	1K	1.23(0.86)	0.92(0.46)	0.98(0.30)	0.95(0.16)	0.97(0.10)
		10K	0.91(0.20)	0.87(0.17)	0.89(0.05)	0.88(0.04)	0.90(0.03)
		100K	0.94(0.08)	0.92(0.07)	0.88(0.04)	0.88(0.04)	0.88(0.03)
CHS	YRI	1K	1.07(0.90)	1.03(0.61)	0.95(0.37)	0.97(0.22)	0.91(0.12)
		10K	1.00(0.46)	0.99(0.22)	1.02(0.12)	1.02(0.08)	0.99(0.06)
		100K	0.88(0.14)	0.93(0.10)	0.96(0.06)	0.98(0.05)	0.98(0.04)
YRI	CHS	1K	1.00(0.49)	1.00(0.35)	0.91(0.24)	1.00(0.17)	1.01(0.10)
		10K	1.02(0.10)	1.00(0.07)	0.95(0.03)	0.94(0.03)	0.94(0.02)
		100K	0.94(0.15)	0.95(0.09)	0.93(0.05)	0.93(0.03)	0.94(0.03)
GBR	YRI	1K	1.10(0.49)	1.10(0.28)	1.06(0.30)	0.98(0.18)	0.97(0.09)
		10K	0.94(0.23)	0.98(0.10)	0.94(0.06)	0.93(0.04)	0.94(0.03)
		100K	1.07(0.09)	1.02(0.08)	0.97(0.06)	0.95(0.05)	0.95(0.04)
YRI	GBR	1K	1.13(0.56)	0.78(0.36)	0.84(0.19)	0.93(0.11)	0.98(0.06)
		10K	0.92(0.24)	0.89(0.15)	0.91(0.06)	0.93(0.05)	0.94(0.05)
		100K	0.95(0.15)	0.93(0.08)	0.93(0.08)	0.93(0.06)	0.94(0.06)
CHS	GBR	1K	1.28(1.24)	1.12(0.70)	1.00(0.40)	0.95(0.21)	0.97(0.13)
		10K	1.06(0.54)	1.01(0.33)	1.00(0.14)	1.00(0.07)	0.98(0.05)
		100K	0.91(0.06)	0.92(0.07)	0.97(0.07)	0.99(0.06)	0.99(0.05)
GBR	CHS	1K	0.89(0.47)	0.83(0.42)	0.84(0.17)	0.91(0.14)	0.92(0.13)
		10K	0.97(0.17)	0.93(0.11)	0.94(0.08)	0.94(0.06)	0.92(0.06)
		100K	1.01(0.12)	0.97(0.10)	0.93(0.08)	0.94(0.07)	0.94(0.06)

Contamination Estimation under Different Parameter Settings

Supplementary Table 3.S4. A full table summarizing the contamination rate ratio (Estimated/Intended) across various simulation parameters, populations, and estimation methods shown in Figure 4. (100K marker sets were used.)

Sample Population		Method	Allele Frequencies	Intended % Contam.	Mean	SD	MSE
Intended	Contam.						
GBR	GBR	VB1	AFR	1%	0.52	0.11	0.242
GBR	GBR	VB1	AFR	2%	0.53	0.09	0.223
GBR	GBR	VB1	AFR	5%	0.53	0.06	0.221
GBR	GBR	VB1	AFR	10%	0.52	0.05	0.237
GBR	GBR	VB1	AFR	20%	0.48	0.04	0.276
GBR	GBR	VB1	EUR	1%	1.04	0.10	0.012
GBR	GBR	VB1	EUR	2%	0.98	0.09	0.007
GBR	GBR	VB1	EUR	5%	0.95	0.07	0.008
GBR	GBR	VB1	EUR	10%	0.94	0.06	0.008
GBR	GBR	VB1	EUR	20%	0.92	0.05	0.008
GBR	GBR	VB1	EAS	1%	0.65	0.11	0.136
GBR	GBR	VB1	EAS	2%	0.65	0.09	0.132
GBR	GBR	VB1	EAS	5%	0.64	0.06	0.135
GBR	GBR	VB1	EAS	10%	0.62	0.05	0.148
GBR	GBR	VB1	EAS	20%	0.58	0.05	0.179
GBR	GBR	VB1	Pooled	1%	0.79	0.11	0.055
GBR	GBR	VB1	Pooled	2%	0.77	0.08	0.060
GBR	GBR	VB1	Pooled	5%	0.75	0.07	0.066
GBR	GBR	VB1	Pooled	10%	0.74	0.06	0.069
GBR	GBR	VB1	Pooled	20%	0.73	0.05	0.073
GBR	GBR	VB2	ISAF (Equal-Ancestry)	1%	1.02	0.11	0.010
GBR	GBR	VB2	ISAF (Equal-Ancestry)	2%	0.96	0.09	0.009
GBR	GBR	VB2	ISAF (Equal -Ancestry)	5%	0.93	0.07	0.010
GBR	GBR	VB2	ISAF (Equal -Ancestry)	10%	0.92	0.06	0.010
GBR	GBR	VB2	ISAF (Equal -Ancestry)	20%	0.91	0.05	0.010
GBR	GBR	VB2	ISAF (Unequal-Ancestry)	1%	1.00	0.10	0.009
GBR	GBR	VB2	ISAF (Unequal-Ancestry)	2%	0.96	0.09	0.009
GBR	GBR	VB2	ISAF (Unequal-Ancestry)	5%	0.93	0.08	0.011
GBR	GBR	VB2	ISAF (Unequal-Ancestry)	10%	0.92	0.06	0.010

GBR	GBR	VB2	ISAF (Unequal-Ancestry)	20%	0.91	0.05	0.010
GBR	CHS	VB1	AFR	1%	0.59	0.08	0.172
GBR	CHS	VB1	AFR	2%	0.60	0.06	0.162
GBR	CHS	VB1	AFR	5%	0.61	0.05	0.154
GBR	CHS	VB1	AFR	10%	0.59	0.04	0.169
GBR	CHS	VB1	AFR	20%	0.55	0.03	0.206
GBR	CHS	VB1	EUR	1%	1.17	0.11	0.039
GBR	CHS	VB1	EUR	2%	1.09	0.10	0.016
GBR	CHS	VB1	EUR	5%	1.06	0.08	0.010
GBR	CHS	VB1	EUR	10%	1.04	0.07	0.006
GBR	CHS	VB1	EUR	20%	0.99	0.06	0.003
GBR	CHS	VB1	EAS	1%	0.74	0.09	0.074
GBR	CHS	VB1	EAS	2%	0.74	0.07	0.072
GBR	CHS	VB1	EAS	5%	0.76	0.06	0.063
GBR	CHS	VB1	EAS	10%	0.76	0.05	0.061
GBR	CHS	VB1	EAS	20%	0.75	0.04	0.062
GBR	CHS	VB1	Pooled	1%	0.89	0.09	0.019
GBR	CHS	VB1	Pooled	2%	0.86	0.07	0.025
GBR	CHS	VB1	Pooled	5%	0.86	0.06	0.022
GBR	CHS	VB1	Pooled	10%	0.86	0.06	0.022
GBR	CHS	VB1	Pooled	20%	0.86	0.05	0.021
GBR	CHS	VB2	ISAF (Equal-Ancestry)	1%	1.13	0.11	0.028
GBR	CHS	VB2	ISAF (Equal-Ancestry)	2%	1.06	0.09	0.011
GBR	CHS	VB2	ISAF (Equal -Ancestry)	5%	1.03	0.08	0.007
GBR	CHS	VB2	ISAF (Equal -Ancestry)	10%	1.01	0.07	0.004
GBR	CHS	VB2	ISAF (Equal -Ancestry)	20%	0.99	0.06	0.004
GBR	CHS	VB2	ISAF (Unequal-Ancestry)	1%	1.01	0.12	0.012
GBR	CHS	VB2	ISAF (Unequal-Ancestry)	2%	0.97	0.10	0.010
GBR	CHS	VB2	ISAF (Unequal-Ancestry)	5%	0.93	0.08	0.010
GBR	CHS	VB2	ISAF (Unequal-Ancestry)	10%	0.94	0.07	0.008
GBR	CHS	VB2	ISAF (Unequal-Ancestry)	20%	0.94	0.06	0.007
GBR	YRI	VB1	AFR	1%	0.67	0.11	0.119
GBR	YRI	VB1	AFR	2%	0.70	0.09	0.096
GBR	YRI	VB1	AFR	5%	0.72	0.06	0.082
GBR	YRI	VB1	AFR	10%	0.73	0.05	0.077

GBR	YRI	VB1	AFR	20%	0.73	0.04	0.074
GBR	YRI	VB1	EUR	1%	1.38	0.10	0.150
GBR	YRI	VB1	EUR	2%	1.30	0.09	0.097
GBR	YRI	VB1	EUR	5%	1.23	0.07	0.055
GBR	YRI	VB1	EUR	10%	1.17	0.05	0.032
GBR	YRI	VB1	EUR	20%	1.10	0.04	0.011
GBR	YRI	VB1	EAS	1%	0.85	0.10	0.032
GBR	YRI	VB1	EAS	2%	0.85	0.08	0.028
GBR	YRI	VB1	EAS	5%	0.83	0.06	0.031
GBR	YRI	VB1	EAS	10%	0.80	0.04	0.042
GBR	YRI	VB1	EAS	20%	0.74	0.03	0.069
GBR	YRI	VB1	Pooled	1%	1.05	0.09	0.010
GBR	YRI	VB1	Pooled	2%	1.03	0.08	0.007
GBR	YRI	VB1	Pooled	5%	1.01	0.06	0.003
GBR	YRI	VB1	Pooled	10%	1.00	0.05	0.002
GBR	YRI	VB1	Pooled	20%	0.99	0.04	0.002
GBR	YRI	VB2	ISAF (Equal-Ancestry)	1%	1.33	0.09	0.118
GBR	YRI	VB2	ISAF (Equal-Ancestry)	2%	1.26	0.09	0.074
GBR	YRI	VB2	ISAF (Equal -Ancestry)	5%	1.18	0.06	0.036
GBR	YRI	VB2	ISAF (Equal -Ancestry)	10%	1.13	0.05	0.018
GBR	YRI	VB2	ISAF (Equal -Ancestry)	20%	1.08	0.04	0.008
GBR	YRI	VB2	ISAF (Unequal-Ancestry)	1%	1.07	0.09	0.014
GBR	YRI	VB2	ISAF (Unequal-Ancestry)	2%	1.02	0.08	0.006
GBR	YRI	VB2	ISAF (Unequal-Ancestry)	5%	0.97	0.06	0.004
GBR	YRI	VB2	ISAF (Unequal-Ancestry)	10%	0.95	0.05	0.004
GBR	YRI	VB2	ISAF (Unequal-Ancestry)	20%	0.95	0.04	0.004
CHS	GBR	VB1	AFR	1%	0.07	0.04	0.868
CHS	GBR	VB1	AFR	2%	0.23	0.05	0.597
CHS	GBR	VB1	AFR	5%	0.40	0.04	0.366
CHS	GBR	VB1	AFR	10%	0.46	0.03	0.290
CHS	GBR	VB1	AFR	20%	0.47	0.02	0.282
CHS	GBR	VB1	EUR	1%	0.21	0.05	0.625
CHS	GBR	VB1	EUR	2%	0.40	0.05	0.364
CHS	GBR	VB1	EUR	5%	0.57	0.05	0.184
CHS	GBR	VB1	EUR	10%	0.66	0.04	0.119

CHS	GBR	VB1	EUR	20%	0.71	0.04	0.087
CHS	GBR	VB1	EAS	1%	1.24	0.11	0.069
CHS	GBR	VB1	EAS	2%	1.26	0.10	0.075
CHS	GBR	VB1	EAS	5%	1.27	0.08	0.077
CHS	GBR	VB1	EAS	10%	1.25	0.06	0.066
CHS	GBR	VB1	EAS	20%	1.18	0.05	0.035
CHS	GBR	VB1	Pooled	1%	0.36	0.06	0.409
CHS	GBR	VB1	Pooled	2%	0.55	0.06	0.207
CHS	GBR	VB1	Pooled	5%	0.71	0.05	0.086
CHS	GBR	VB1	Pooled	10%	0.79	0.05	0.047
CHS	GBR	VB1	Pooled	20%	0.83	0.04	0.031
CHS	GBR	VB2	ISAF (Equal-Ancestry)	1%	1.22	0.11	0.060
CHS	GBR	VB2	ISAF (Equal-Ancestry)	2%	1.23	0.10	0.060
CHS	GBR	VB2	ISAF (Equal -Ancestry)	5%	1.23	0.08	0.057
CHS	GBR	VB2	ISAF (Equal -Ancestry)	10%	1.20	0.06	0.044
CHS	GBR	VB2	ISAF (Equal -Ancestry)	20%	1.15	0.05	0.025
CHS	GBR	VB2	ISAF (Unequal-Ancestry)	1%	0.91	0.06	0.011
CHS	GBR	VB2	ISAF (Unequal-Ancestry)	2%	0.92	0.07	0.010
CHS	GBR	VB2	ISAF (Unequal-Ancestry)	5%	0.97	0.07	0.005
CHS	GBR	VB2	ISAF (Unequal-Ancestry)	10%	0.99	0.06	0.004
CHS	GBR	VB2	ISAF (Unequal-Ancestry)	20%	0.99	0.05	0.003
CHS	CHS	VB1	AFR	1%	0.02	0.03	0.956
CHS	CHS	VB1	AFR	2%	0.12	0.05	0.782
CHS	CHS	VB1	AFR	5%	0.25	0.03	0.562
CHS	CHS	VB1	AFR	10%	0.29	0.02	0.500
CHS	CHS	VB1	AFR	20%	0.29	0.01	0.500
CHS	CHS	VB1	EUR	1%	0.10	0.07	0.815
CHS	CHS	VB1	EUR	2%	0.24	0.05	0.573
CHS	CHS	VB1	EUR	5%	0.38	0.03	0.385
CHS	CHS	VB1	EUR	10%	0.42	0.02	0.338
CHS	CHS	VB1	EUR	20%	0.42	0.02	0.337
CHS	CHS	VB1	EAS	1%	0.90	0.10	0.020
CHS	CHS	VB1	EAS	2%	0.92	0.07	0.011
CHS	CHS	VB1	EAS	5%	0.95	0.06	0.006
CHS	CHS	VB1	EAS	10%	0.95	0.04	0.004

CHS	CHS	VB1	EAS	20%	0.94	0.03	0.004
CHS	CHS	VB1	Pooled	1%	0.19	0.08	0.659
CHS	CHS	VB1	Pooled	2%	0.34	0.05	0.435
CHS	CHS	VB1	Pooled	5%	0.48	0.04	0.275
CHS	CHS	VB1	Pooled	10%	0.52	0.03	0.233
CHS	CHS	VB1	Pooled	20%	0.53	0.02	0.222
CHS	CHS	VB2	ISAF (Equal-Ancestry)	1%	0.90	0.11	0.021
CHS	CHS	VB2	ISAF (Equal-Ancestry)	2%	0.91	0.07	0.012
CHS	CHS	VB2	ISAF (Equal -Ancestry)	5%	0.95	0.06	0.006
CHS	CHS	VB2	ISAF (Equal -Ancestry)	10%	0.95	0.04	0.004
CHS	CHS	VB2	ISAF (Equal -Ancestry)	20%	0.95	0.03	0.003
CHS	CHS	VB2	ISAF (Unequal-Ancestry)	1%	0.89	0.12	0.024
CHS	CHS	VB2	ISAF (Unequal-Ancestry)	2%	0.92	0.08	0.012
CHS	CHS	VB2	ISAF (Unequal-Ancestry)	5%	0.95	0.07	0.006
CHS	CHS	VB2	ISAF (Unequal-Ancestry)	10%	0.95	0.05	0.004
CHS	CHS	VB2	ISAF (Unequal-Ancestry)	20%	0.96	0.04	0.003
CHS	YRI	VB1	AFR	1%	0.09	0.09	0.828
CHS	YRI	VB1	AFR	2%	0.27	0.08	0.543
CHS	YRI	VB1	AFR	5%	0.45	0.05	0.302
CHS	YRI	VB1	AFR	10%	0.54	0.04	0.210
CHS	YRI	VB1	AFR	20%	0.61	0.03	0.155
CHS	YRI	VB1	EUR	1%	0.31	0.11	0.492
CHS	YRI	VB1	EUR	2%	0.51	0.08	0.250
CHS	YRI	VB1	EUR	5%	0.67	0.05	0.114
CHS	YRI	VB1	EUR	10%	0.72	0.04	0.083
CHS	YRI	VB1	EUR	20%	0.71	0.03	0.084
CHS	YRI	VB1	EAS	1%	1.38	0.14	0.161
CHS	YRI	VB1	EAS	2%	1.39	0.10	0.163
CHS	YRI	VB1	EAS	5%	1.38	0.07	0.151
CHS	YRI	VB1	EAS	10%	1.35	0.06	0.123
CHS	YRI	VB1	EAS	20%	1.24	0.04	0.061
CHS	YRI	VB1	Pooled	1%	0.47	0.13	0.298
CHS	YRI	VB1	Pooled	2%	0.66	0.09	0.123
CHS	YRI	VB1	Pooled	5%	0.82	0.06	0.036
CHS	YRI	VB1	Pooled	10%	0.89	0.05	0.015

CHS	YRI	VB1	Pooled	20%	0.93	0.04	0.007
CHS	YRI	VB2	ISAF (Equal-Ancestry)	1%	1.37	0.14	0.157
CHS	YRI	VB2	ISAF (Equal-Ancestry)	2%	1.38	0.11	0.154
CHS	YRI	VB2	ISAF (Equal -Ancestry)	5%	1.34	0.07	0.122
CHS	YRI	VB2	ISAF (Equal -Ancestry)	10%	1.29	0.06	0.085
CHS	YRI	VB2	ISAF (Equal -Ancestry)	20%	1.22	0.05	0.048
CHS	YRI	VB2	ISAF (Unequal-Ancestry)	1%	0.88	0.14	0.033
CHS	YRI	VB2	ISAF (Unequal-Ancestry)	2%	0.93	0.10	0.014
CHS	YRI	VB2	ISAF (Unequal-Ancestry)	5%	0.96	0.06	0.005
CHS	YRI	VB2	ISAF (Unequal-Ancestry)	10%	0.98	0.05	0.002
CHS	YRI	VB2	ISAF (Unequal-Ancestry)	20%	0.98	0.04	0.002
YRI	GBR	VB1	AFR	1%	1.43	0.23	0.236
YRI	GBR	VB1	AFR	2%	1.31	0.14	0.113
YRI	GBR	VB1	AFR	5%	1.28	0.08	0.081
YRI	GBR	VB1	AFR	10%	1.25	0.07	0.065
YRI	GBR	VB1	AFR	20%	1.17	0.06	0.034
YRI	GBR	VB1	EUR	1%	0.36	0.22	0.453
YRI	GBR	VB1	EUR	2%	0.46	0.19	0.329
YRI	GBR	VB1	EUR	5%	0.56	0.11	0.201
YRI	GBR	VB1	EUR	10%	0.62	0.07	0.153
YRI	GBR	VB1	EUR	20%	0.65	0.05	0.124
YRI	GBR	VB1	EAS	1%	0.32	0.20	0.504
YRI	GBR	VB1	EAS	2%	0.41	0.18	0.380
YRI	GBR	VB1	EAS	5%	0.52	0.11	0.245
YRI	GBR	VB1	EAS	10%	0.55	0.07	0.204
YRI	GBR	VB1	EAS	20%	0.56	0.04	0.198
YRI	GBR	VB1	Pooled	1%	0.57	0.26	0.248
YRI	GBR	VB1	Pooled	2%	0.67	0.17	0.138
YRI	GBR	VB1	Pooled	5%	0.76	0.10	0.066
YRI	GBR	VB1	Pooled	10%	0.80	0.07	0.043
YRI	GBR	VB1	Pooled	20%	0.83	0.05	0.030
YRI	GBR	VB2	ISAF (Equal-Ancestry)	1%	1.30	0.15	0.107
YRI	GBR	VB2	ISAF (Equal-Ancestry)	2%	1.25	0.11	0.072
YRI	GBR	VB2	ISAF (Equal -Ancestry)	5%	1.20	0.09	0.048
YRI	GBR	VB2	ISAF (Equal -Ancestry)	10%	1.15	0.07	0.028

YRI	GBR	VB2	ISAF (Equal -Ancestry)	20%	1.10	0.07	0.013
YRI	GBR	VB2	ISAF (Unequal-Ancestry)	1%	0.95	0.15	0.021
YRI	GBR	VB2	ISAF (Unequal-Ancestry)	2%	0.93	0.08	0.012
YRI	GBR	VB2	ISAF (Unequal-Ancestry)	5%	0.93	0.08	0.011
YRI	GBR	VB2	ISAF (Unequal-Ancestry)	10%	0.93	0.06	0.008
YRI	GBR	VB2	ISAF (Unequal-Ancestry)	20%	0.94	0.06	0.006
YRI	CHS	VB1	AFR	1%	1.32	0.13	0.120
YRI	CHS	VB1	AFR	2%	1.30	0.11	0.098
YRI	CHS	VB1	AFR	5%	1.26	0.07	0.071
YRI	CHS	VB1	AFR	10%	1.22	0.05	0.049
YRI	CHS	VB1	AFR	20%	1.14	0.04	0.022
YRI	CHS	VB1	EUR	1%	0.35	0.20	0.455
YRI	CHS	VB1	EUR	2%	0.46	0.18	0.319
YRI	CHS	VB1	EUR	5%	0.56	0.10	0.204
YRI	CHS	VB1	EUR	10%	0.59	0.06	0.168
YRI	CHS	VB1	EUR	20%	0.60	0.03	0.159
YRI	CHS	VB1	EAS	1%	0.29	0.17	0.528
YRI	CHS	VB1	EAS	2%	0.40	0.18	0.395
YRI	CHS	VB1	EAS	5%	0.51	0.11	0.252
YRI	CHS	VB1	EAS	10%	0.56	0.06	0.194
YRI	CHS	VB1	EAS	20%	0.61	0.03	0.157
YRI	CHS	VB1	Pooled	1%	0.55	0.24	0.259
YRI	CHS	VB1	Pooled	2%	0.66	0.16	0.136
YRI	CHS	VB1	Pooled	5%	0.75	0.09	0.068
YRI	CHS	VB1	Pooled	10%	0.79	0.04	0.044
YRI	CHS	VB1	Pooled	20%	0.83	0.02	0.030
YRI	CHS	VB2	ISAF (Equal-Ancestry)	1%	1.29	0.13	0.098
YRI	CHS	VB2	ISAF (Equal-Ancestry)	2%	1.25	0.11	0.074
YRI	CHS	VB2	ISAF (Equal -Ancestry)	5%	1.20	0.07	0.043
YRI	CHS	VB2	ISAF (Equal -Ancestry)	10%	1.15	0.04	0.023
YRI	CHS	VB2	ISAF (Equal -Ancestry)	20%	1.10	0.04	0.011
YRI	CHS	VB2	ISAF (Unequal-Ancestry)	1%	0.94	0.15	0.023
YRI	CHS	VB2	ISAF (Unequal-Ancestry)	2%	0.95	0.09	0.010
YRI	CHS	VB2	ISAF (Unequal-Ancestry)	5%	0.93	0.05	0.007
YRI	CHS	VB2	ISAF (Unequal-Ancestry)	10%	0.93	0.03	0.005

YRI	CHS	VB2	ISAF (Unequal-Ancestry)	20%	0.94	0.03	0.004
YRI	YRI	VB1	AFR	1%	0.95	0.09	0.010
YRI	YRI	VB1	AFR	2%	0.92	0.06	0.009
YRI	YRI	VB1	AFR	5%	0.89	0.05	0.014
YRI	YRI	VB1	AFR	10%	0.89	0.04	0.013
YRI	YRI	VB1	AFR	20%	0.89	0.03	0.012
YRI	YRI	VB1	EUR	1%	0.22	0.13	0.619
YRI	YRI	VB1	EUR	2%	0.29	0.14	0.518
YRI	YRI	VB1	EUR	5%	0.36	0.09	0.423
YRI	YRI	VB1	EUR	10%	0.38	0.06	0.391
YRI	YRI	VB1	EUR	20%	0.36	0.03	0.405
YRI	YRI	VB1	EAS	1%	0.18	0.11	0.680
YRI	YRI	VB1	EAS	2%	0.25	0.13	0.575
YRI	YRI	VB1	EAS	5%	0.32	0.09	0.474
YRI	YRI	VB1	EAS	10%	0.34	0.06	0.438
YRI	YRI	VB1	EAS	20%	0.33	0.03	0.452
YRI	YRI	VB1	Pooled	1%	0.36	0.18	0.433
YRI	YRI	VB1	Pooled	2%	0.44	0.14	0.333
YRI	YRI	VB1	Pooled	5%	0.49	0.08	0.267
YRI	YRI	VB1	Pooled	10%	0.51	0.05	0.242
YRI	YRI	VB1	Pooled	20%	0.51	0.03	0.245
YRI	YRI	VB2	ISAF (Equal-Ancestry)	1%	0.94	0.10	0.012
YRI	YRI	VB2	ISAF (Equal-Ancestry)	2%	0.92	0.06	0.011
YRI	YRI	VB2	ISAF (Equal -Ancestry)	5%	0.88	0.04	0.016
YRI	YRI	VB2	ISAF (Equal -Ancestry)	10%	0.88	0.04	0.015
YRI	YRI	VB2	ISAF (Equal -Ancestry)	20%	0.88	0.03	0.015
YRI	YRI	VB2	ISAF (Unequal-Ancestry)	1%	0.94	0.08	0.010
YRI	YRI	VB2	ISAF (Unequal-Ancestry)	2%	0.92	0.07	0.011
YRI	YRI	VB2	ISAF (Unequal-Ancestry)	5%	0.88	0.04	0.016
YRI	YRI	VB2	ISAF (Unequal-Ancestry)	10%	0.88	0.04	0.015
YRI	YRI	VB2	ISAF (Unequal-Ancestry)	20%	0.88	0.03	0.015

Appendix C

Illustration of *freemuxlet* workflow

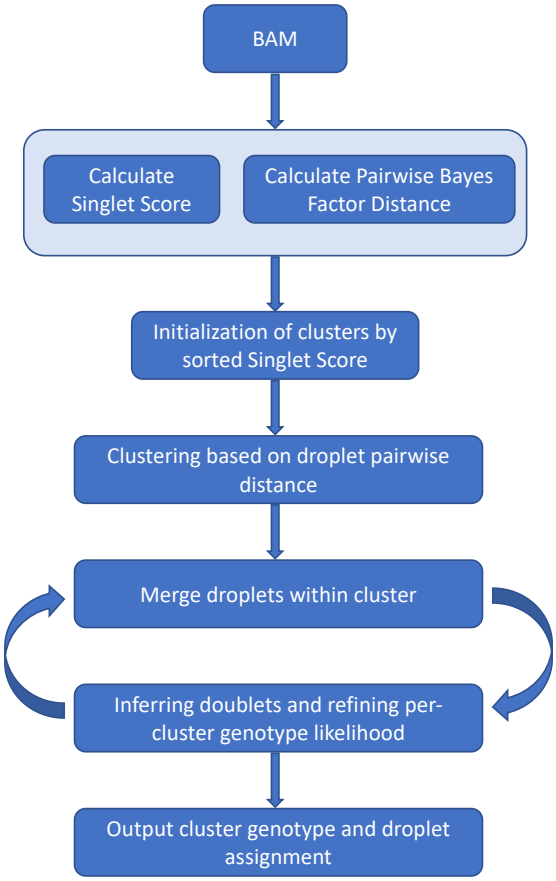

Figure 4.S1.Workflow of *freemuxlet*. The *freemuxlet* takes a BAM file of multiplexed sample sequences as input, and outputs pileup information of reads overlapping with known genetic variants. *freemuxlet* then calculate singlet score and pairwise Bayes Factor distance. Based on sorted singlet score, clusters are initialized by preferably choosing droplets being more likely to be a singlet. Droplets are then clustered based on BF distance and later merged based on membership to aggregately call genotypes of each cluster. Genotypes of each cluster are used to detect and remove non-singlets.

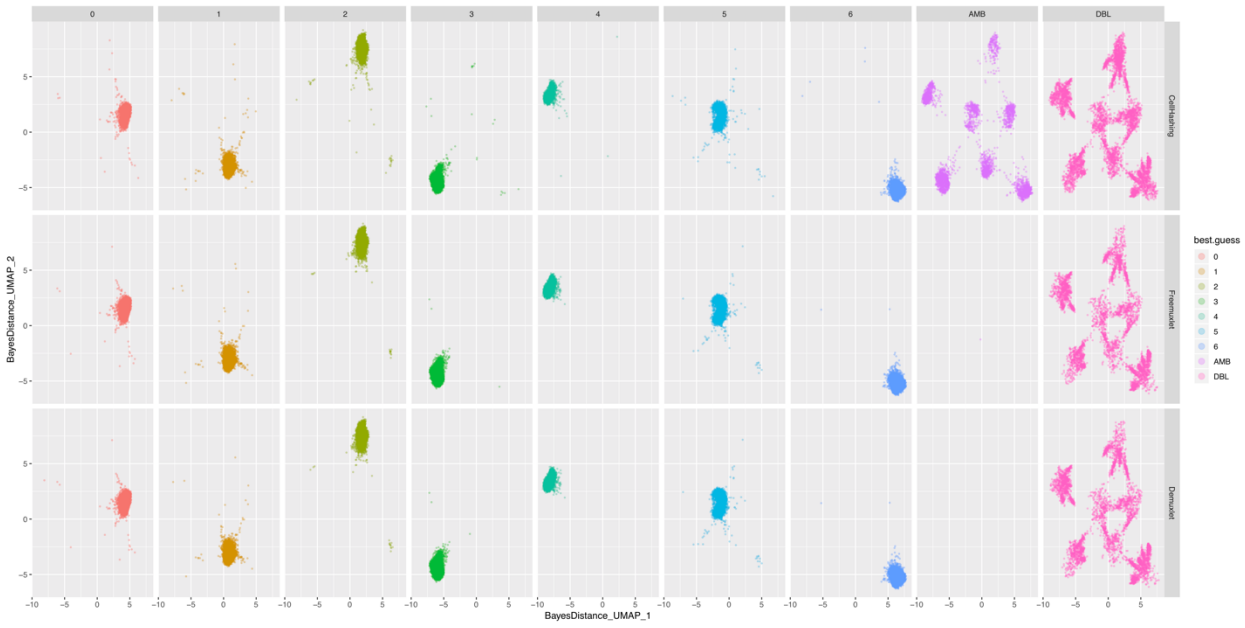

Figure 4.S2. Comparison of droplets assignment from 3 methods visualized based on UMAP clustering result. Droplets are clustered using UMAP based on BF distance. Each color represents one individual assignment (Ambiguous as purple and Doublet as pink). Assignments are estimated from CellHashing (upper panel), *freemuxlet* (middle panel), and *demuxlet* (lower panel).

t-SNE Clustering Result Based on Bayes Distance

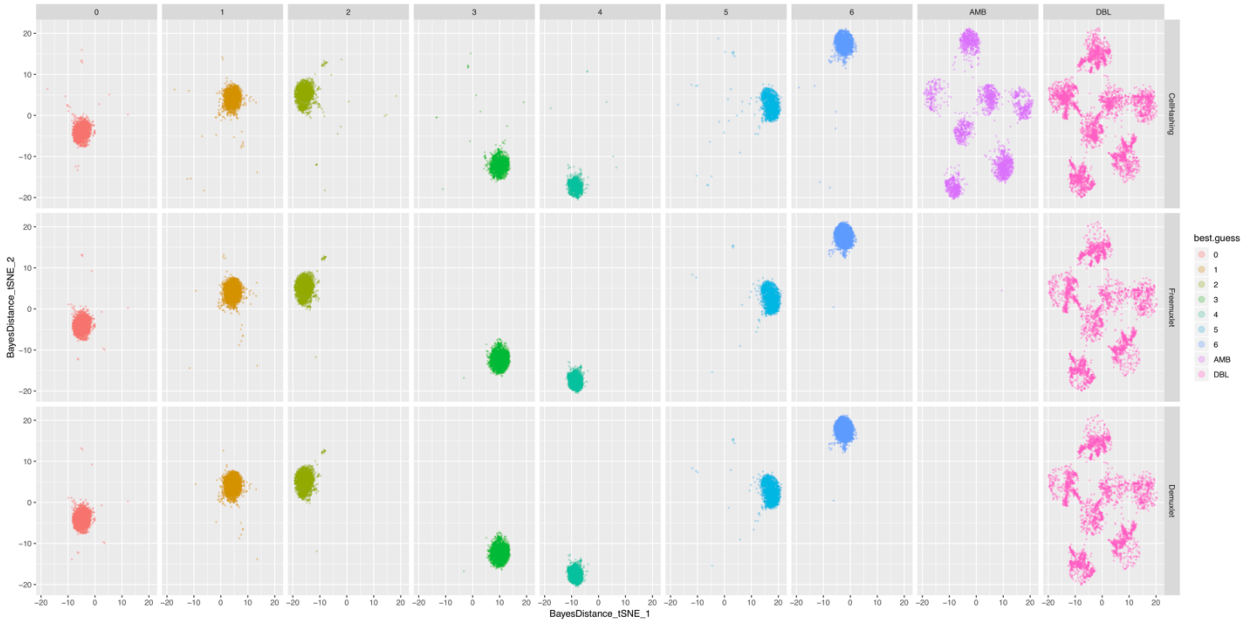

Figure 4.S3. Comparison of droplets assignment from 3 methods visualized based on tSNE clustering result. Droplets are clustered using t-SNE based on BF distance. Each color represents one individual assignment (Ambiguous as purple and Doublet as pink). Assignments are estimated from CellHashing (upper panel), *freemuxlet* (middle panel), and *demuxlet* (lower panel).

Experiment Design for Sample Identity Recovering

Batch1	ID1	ID3	ID5
Batch2	ID1	ID4	ID6
Batch3	ID2	ID3	ID6
Batch4	ID2	ID4	ID5

Figure 4.S4 Example configuration of pair-identifiable mux-seq experiment with 6 samples and 4 batches.

Sequence Error Model Used in Genotype Likelihood

Table 4.S1. Sequence error model. Conditional probability $P(b_{ij}| g_i, e_{ij})$ of read b_{ij} given true genotype g_i and the variable representing the event of base calling error e_{ij}

True Genotype g_i	**Base Calling Error Event e_{ij}**	**Pr(b_{ij} = R)**	**Pr(b_{ij} = A)**	**Pr(b_{ij} = O)**[b]
g_i = RR[a]	$e_{ij} = 0$	1	0	0
	$e_{ij} = 1$	0	1/3	2/3
g_i = RA[a]	$e_{ij} = 0$	1/2	1/2	0
	$e_{ij} = 1$	1/6	1/6	2/3
g_i = AA[a]	$e_{ij} = 0$	0	1	0
	$e_{ij} = 1$	1/3	0	2/3

[a] RR, RA, AA: homozygous reference, heterozygous, and homozygous non-reference genotypes

[b] O: alleles other than R or A; assumes four possible alleles (bases)

BIBLIOGRAPHY

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2016.49
2. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* (2018). doi:10.1038/nprot.2017.149
3. Allyse, M. *et al.* Non-invasive prenatal testing: A review of international implementation and challenges. *International Journal of Women's Health* (2015). doi:10.2147/IJWH.S67124
4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012). doi:10.1038/nature11247
6. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* (2008). doi:10.1038/nature07517
7. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* (1988). doi:10.1126/science.3413476
8. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* (2011). doi:10.1038/nature10242
9. Deamer, D. W. & Akeson, M. Nanopores and nucleic acids: Prospects for ultrarapid sequencing. *Trends in Biotechnology* (2000). doi:10.1016/S0167-7799(00)01426-8
10. Andrews, S. & Babraham Bioinformatics. FastQC: A quality control tool for high throughput sequence data. *Manual* (2010). doi:citeulike-article-id:11583827
11. Broad Institute. Picard Tools. *Broad Institute, GitHub repository*
12. Li, B. *et al.* QPLOT: A quality assessment tool for next generation sequencing data. *BioMed Research International* (2013). doi:10.1155/2013/865181
13. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics* (2012). doi:10.1016/j.ajhg.2012.09.004
14. Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M. & Kang, H. M. Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *American Journal of Human Genetics* (2015). doi:10.1016/j.ajhg.2015.07.002
15. Fiévet, A. *et al.* ART-DeCo: easy tool for detection and characterization of cross-contamination of DNA samples in diagnostic next-generation sequencing analysis. *European Journal of Human Genetics* (2019). doi:10.1038/s41431-018-0317-x
16. Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V. & Zody, M. C. Conpair: Concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw389

17. Lee, S. *et al.* NGSCheckMate: Software for validating sample identity in Next-generation sequencing studies within and across data types. *Nucleic Acids Research* (2017). doi:10.1093/nar/gkx193
18. Cibulskis, K. *et al.* ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr446
19. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* (2006). doi:10.1038/ng1847
20. Wang, C., Zhan, X., Liang, L., Abecasis, G. R. & Lin, X. Improved Ancestry Estimation for both Genotyping and Sequencing Data using Projection Procrustes Analysis and Genotype Imputation. *American Journal of Human Genetics* (2015). doi:10.1016/j.ajhg.2015.04.018
21. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics* (2014). doi:10.1038/ng.2924
22. Cox, T. & Cox, M. *Multidimensional Scaling, Second Edition*. *New York* (2000). doi:10.1201/9781420036121
23. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics* (2017). doi:10.1038/ng.3818
24. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* (2017). doi:10.1038/ncomms14049
25. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* (2015). doi:10.1016/j.cell.2015.05.002
26. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* (2018). doi:10.1038/nbt.4042
27. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* (2018). doi:10.1186/s13059-018-1603-1
28. McGinnis, C. S. *et al.* MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *bioRxiv* (2018). doi:10.1101/387241
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
30. Zhang, F., Flickinger, M., Abecasis, G., Boehnke, M. & Kang, H. M. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *bioRxiv Bioinformatics* 1–42 (2018). doi:10.1101/466268
31. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics* **98**, 127–148 (2016).
32. Hao, W., Song, M. & Storey, J. D. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics (Oxford, England)* **32**, 713–721 (2015).
33. Martínez-Alcántara, A. *et al.* PIQA: Pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* **25**, 2438–2439 (2009).
34. Yang, X. *et al.* HTQC: A fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* **14**, (2013).
35. Broad Institute. Picard:A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. *http://broadinstitute.github.io/picard/* (2016). doi:10.1007/s00586-004-0822-1
36. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis

framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research* **25**, 918–925 (2015).
37. Satten, G. A. & Datta, S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician* **55**, 207–210 (2001).
38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
39. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* **36**, 875–881 (2018).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
42. Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M. & Kang, H. M. Correcting for sample contamination in genotype calling of DNA sequence data. *American Journal of Human Genetics* **97**, 284–290 (2015).
43. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics* **91**, 839–848 (2012).
44. Cavalli-Sforza, L. L. The human genome diversity project: past, present and future. *Nat Rev Genet* **6**, 333–340 (2005).
45. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
46. Wang, C., Zhan, X., Liang, L., Abecasis, G. R. & Lin, X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *American Journal of Human Genetics* **96**, 926–937 (2015).
47. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications* **9**, (2018).
48. Malaspinas, A. S. *et al.* bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics (Oxford, England)* **30**, 2962–2964 (2014).
49. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909 (2006).
50. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology* **28**, 289–301 (2005).
51. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664 (2009).
52. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* **36**, 89–94 (2018).
53. Brent, R. P. Algorithms for minimization without derivatives. *IEEE Transactions on Automatic Control* (1974). doi:10.1109/TAC.1974.1100629
54. Yang, W. W.-Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics* **44**, 725–731 (2012).
55. Wang, C. *et al.* Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics* **46**, 409–15 (2014).
56. Pearson, K. LIII. *On lines and planes of closest fit to systems of points in space*. *Philosophical Magazine Series 6* **2**, 559–572 (1901).

57. Nelder, J. A. & Mead, R. A simplex method for function minimization. *The Computer Journal* **7**, 308–313 (1965).
58. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).
59. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* (2015). doi:10.1016/j.cell.2015.05.002
60. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* (2015). doi:10.1016/j.cell.2015.04.044
61. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* (2017). doi:10.1038/ncomms14049
62. Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods* (2017). doi:10.1038/nmeth.4179
63. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods* (2017). doi:10.1038/nmeth.4154
64. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* (2018). doi:10.1126/science.aam8999
65. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods* (2017). doi:10.1038/nmeth.4407
66. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *Nature Communications* (2018). doi:10.1038/s41467-018-05887-x
67. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* (2018). doi:10.1126/science.aau0730
68. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* (2018). doi:10.1038/nbt.4042
69. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* (2018). doi:10.1186/s13059-018-1603-1
70. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* (2019). doi:10.1038/nbt.4314
71. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
72. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* (2014).
73. Tate, J. G. *et al.* COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky1015
74. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012). doi:10.1038/nature11003
75. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* (2018). doi:10.1038/nbt.4096
76. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* (1998). doi:10.1101/gr.8.3.186
77. Waltman, L. & Van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B* (2013). doi:10.1140/epjb/e2013-40829-0
78. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants

using mapping quality scores. *Genome Research* **18**, 1851–1858 (2008).
79. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* (2009). doi:10.1126/science.1162986
80. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* (2010). doi:10.1093/nar/gkq543
81. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* (2018). doi:10.1038/nbt.4227
82. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq559
83. Hu, Y. J., Sun, W., Tzeng, J. Y. & Perou, C. M. Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. *Journal of the American Statistical Association* (2015). doi:10.1080/01621459.2015.1038449