2017

Statistical Methods For Genomic And Transcriptomic Sequencing

Yuchao Jiang
University of Pennsylvania, yuchaoj@upenn.edu

Statistical Methods For Genomic And Transcriptomic Sequencing

Abstract
Part 1: High-throughput sequencing of DNA coding regions has become a common way of assaying genomic variation in the study of human diseases. Copy number variation (CNV) is an important type of genomic variation, but CNV profiling from whole-exome sequencing (WES) is challenging due to the high level of biases and artifacts. We propose CODEX, a normalization and CNV calling procedure for WES data. CODEX includes a Poisson latent factor model, which includes terms that specifically remove biases due to GC content, exon capture and amplification efficiency, and latent systemic artifacts. CODEX also includes a Poisson likelihood-based segmentation procedure that explicitly models the count-based WES data. CODEX is compared to existing methods on germline CNV detection in HapMap samples using microarray-based gold standard and is further evaluated on 222 neuroblastoma samples with matched normal, with focus on somatic CNVs within the ATRX gene.

Part 2: Cancer is a disease driven by evolutionary selection on somatic genetic and epigenetic alterations. We propose Canopy, a method for inferring the evolutionary phylogeny of a tumor using both somatic copy number alterations and single nucleotide alterations from one or more samples derived from a single patient. Canopy is applied to bulk sequencing datasets of both longitudinal and spatial experimental designs and to a transplantable metastasis model derived from human cancer cell line MDA-MB-231. Canopy successfully identifies cell populations and infers phylogenies that are in concordance with existing knowledge and ground truth. Through simulations, we explore the effects of key parameters on deconvolution accuracy, and compare against existing methods.

Part 3: Allele-specific expression is traditionally studied by bulk RNA sequencing, which measures average expression across cells. Single-cell RNA sequencing (scRNA-seq) allows the comparison of expression distribution between the two alleles of a diploid organism and thus the characterization of allele-specific bursting. We propose SCALE to analyze genome-wide allele-specific bursting, with adjustment of technical variability. SCALE detects genes exhibiting allelic differences in bursting parameters, and genes whose alleles burst non-independently. We apply SCALE to mouse blastocyst and human fibroblast cells and find that, globally, cis control in gene expression overwhelmingly manifests as differences in burst frequency.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Nancy R. Zhang

Keywords
allele-specific gene expression, cancer genomics, copy number variation, intratumor heterogeneity, next-generation sequencing, single-cell RNA sequencing

Subject Categories
Bioinformatics | Biostatistics | Statistics and Probability

STATISTICAL METHODS FOR GENOMIC AND TRANSCRIPTOMIC SEQUENCING

Yuchao Jiang

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Nancy R. Zhang

Associate Professor of Statistics

Graduate Group Chairperson

_____

Li-San Wang, Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee

Chair: Shane T. Jensen, Associate Professor of Statistics

Mingyao Li, Associate Professor of Biostatistics

Katherine L. Nathanson, Professor of Medicine

Wei Sun, Associate Professor of Biostatistics and Bioinformatics

Li-San Wang, Associate Professor of Pathology and Laboratory Medicine

ACKNOWLEDGMENT

Everything in this dissertation I owe to the mentorship of my advisor Nancy R. Zhang, who is not only an extraordinary statistician and researcher, but also a great mentor. It is my best of luck to be advised by Nancy, from whom I learnt not only how to solve scientific problems but also how to be a better myself in various aspects. I would also like to extend my sincerest thanks to Mingyao Li, whom I have the greatest honor to work with. Thank you both for your guidance and support these past few years and in anticipation of the years to come, and for providing me many opportunities to network with the broader research community in biostatistics and genomics.

I thank my thesis committee members, Shane T. Jensen, Katherine L. Nathanson, Wei Sun, and Li-San Wang for offering invaluable insights and suggestions. I am indebted to our wonderful collaborators, Hao Chen, Kara Maxwell, Bradley Wubbenhorst, Brandon Wenz, Katherine Nathanson, Yu Qiu, Andy Minn, Derek Oldridge, Sharon Diskin, John Maris, Li-San Wang, and Gerald Schellenberg.

I am also grateful to Maja Bucan and Li-San Wang for recruiting me and guiding me as a PhD student in the Genomics and Computational Biology (GCB) graduate group and to Hannah Chervitz and Maureen Kirsch for keeping GCB running smoothly. Just as importantly, GCB students have been great friends in keeping me company and making my academic life colorful. Thank you to Mark Low, Edward George, Shane Jensen, Dylan Small, Noelle Felipe, Adam Greenberg, Sarin Sieng, Tanya Winder, and Carol Reich for their support at the Department of Statistics.

None of this work would have been possible without my parents, who have always been beside me throughout this whole process. No words can express my gratefulness towards their unconditional love. Last but not least, I am deeply grateful to Yuanshuo Qu and Jiayi Bao for being my best friends and companions. PhD is not an easy path and thank you for keeping me healthy, sane, and happy. I cherish the years we had with the many more to come.

ABSTRACT

STATISTICAL METHODS FOR GENOMIC AND TRANSCRIPTOMIC SEQUENCING

Yuchao Jiang

Nancy R. Zhang

Part 1: High-throughput sequencing of DNA coding regions has become a common way of assaying genomic variation in the study of human diseases. Copy number variation (CNV) is an important type of genomic variation, but CNV profiling from whole-exome sequencing (WES) is challenging due to the high level of biases and artifacts. We propose CODEX, a normalization and CNV calling procedure for WES data. CODEX includes a Poisson latent factor model, which includes terms that specifically remove biases due to GC content, exon capture and amplification efficiency, and latent systemic artifacts. CODEX also includes a Poisson likelihood-based segmentation procedure that explicitly models the count-based WES data. CODEX is compared to existing methods on germline CNV detection in HapMap samples using microarray-based gold standard and is further evaluated on 222 neuroblastoma samples with matched normal, with focus on somatic CNVs within the *ATRX* gene.

Part 2: Cancer is a disease driven by evolutionary selection on somatic genetic and epigenetic alterations. We propose Canopy, a method for inferring the evolutionary phylogeny of a tumor using both somatic copy number alterations and single nucleotide alterations from one or more samples derived from a single patient. Canopy is applied to bulk sequencing datasets of both longitudinal and spatial experimental designs and to a transplantable metastasis model derived from human cancer cell line MDA-MB-231. Canopy successfully identifies cell populations and infers phylogenies that are in concordance with existing knowledge and ground truth. Through simulations, we explore the effects of key parameters on deconvolution accuracy, and compare against existing methods.

Part 3: Allele-specific expression is traditionally studied by bulk RNA sequencing, which measures average expression across cells. Single-cell RNA sequencing (scRNA-seq) allows the

comparison of expression distribution between the two alleles of a diploid organism and thus the characterization of allele-specific bursting. We propose SCALE to analyze genome-wide allele-specific bursting, with adjustment of technical variability. SCALE detects genes exhibiting allelic differences in bursting parameters, and genes whose alleles burst non-independently. We apply SCALE to mouse blastocyst and human fibroblast cells and find that, globally, *cis* control in gene expression overwhelmingly manifests as differences in burst frequency.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

NORMALIZATION AND COPY NUMBER VARIATION DETECTION BY WHOLE EXOME SEQUENCING

1.1 Introduction

Copy number variants (CNVs) are large insertions and deletions that lead to gains and losses of segments of chromosomes. CNVs are an important and abundant source of variation in the human genome (1-4). Like other types of genetic variation, some CNVs have been associated with diseases, such as neuroblastoma (5), autism (6), and Crohn's disease (7). Better understanding of the genetics of CNV-associated diseases requires accurate CNV detection. Traditional genome-wide approaches to detect CNVs make use of array comparative genome hybridization (CGH) or single nucleotide polymorphism (SNP) array data (8-10). The minimum detectable size and breakpoint resolution, which are correlated with the density of probes on the array, are limited. Paired end Sanger sequencing, which is often used as the gold standard platform for CNV detection, has better resolution and accuracy but requires significant time and budget investment.

With the dramatic growth of sequencing capacity and the accompanying drop in cost, massively parallel next-generation sequencing (NGS) offers appealing platforms for CNV detection. Many current analysis methods are focused on whole genome sequencing (WGS), which allows for genome-wide CNV detection and finer breakpoint resolution than array-based approaches (11-15). Whole exome sequencing (WES), on the other hand, has been preferred as a cheaper, faster, but still effective alternative to WGS in large-scale studies, where the priority has been to identify disease associated variants in coding regions (16-19).

Due to the biases and artifacts introduced during the exon targeting and amplification steps of WES, depth of coverage in WES data is heavily contaminated with experimental noise and thus does not accurately reflect the true copy number. Here we present a novel

normalization and CNV calling method, CODEX (COpy number variation Detection by EXome sequencing) (20), to remove biases and artifacts in WES data and produce accurate CNV calls.

Several algorithms have been developed for copy number estimation with whole exome data in matched case/control settings by either directly using the matched normal (21-23) or building an optimized reference set (24, 25) to control for artifacts. Other algorithms use singular value decomposition (SVD) to extract copy number signals from noisy coverage matrices by removing K latent factors that explain the most variance (26-28). This exploratory approach assumes continuous measurements with Gaussian noise, uses an arbitrary choice of K, and doesn't specifically model known quantifiable biases, such as those due to GC content.

CODEX does not require matched normal controls, but relies on the availability of multiple samples processed using the same sequencing pipeline. Unlike current approaches, CODEX uses a Poisson log-linear model that is more suitable for discrete count data. The normalization model in CODEX includes terms that specifically remove biases due to GC content, exon length and capture and amplification efficiency, and latent systematic artifacts. We explore several different statistical approaches for choosing the number of latent factors, and discuss how one should set this crucial parameter wisely. The power of CODEX and SVD-based approaches are compared by in silico spike-in studies on the 1000 Genomes Project (29) WES data and show that CODEX offers higher power in detecting both common and rare CNVs. Also, on WES data from the 1000 Genomes Project paired with SNP array data from three previous cohort studies on the same HapMap samples (30-32), CODEX gives higher precision and recall for both rare and common CNV detection by WES data, as compared to existing methods. CODEX's normalization and segmentation accuracy is further evaluated through the analysis of the WES data of 222 neuroblastoma matched tumor/blood samples from the TARGET project (33), with a focus on the well-studied *ATRX* gene region (33-35). The cross-sample normalization procedure of CODEX, when applied to the matrix of tumor samples, is more effective in reducing noise than normalizing

each tumor to its matched normal. The somatic deletions in the *ATRX* region have a nested structure, which CODEX was able to recover.

1.2 Results

1.2.1 Overview of Analysis Pipeline

Figure 1.1 shows an overview of the analysis pipeline of CODEX. We start with mapped reads from BAM files (36) that are assembled, sorted, and indexed by the same pipeline, and compute depth of coverage after a series of quality filtering based on mappability, exon size, and a cutoff on minimum coverage (see details below). Then, we fit a normalization model based on a log-linear decomposition of the depth of coverage matrix into effects due to GC content, exon capture and amplification, and other latent systemic factors. The normalization model produces an estimated "control coverage" for each exon and each sample, which is the coverage we expect to see if there is no CNV. Next, the observed coverage for each exon and each sample is compared to the corresponding estimated control coverage in a Poisson likelihood-based segmentation algorithm, which returns a segmentation of the genome into regions of homogeneous copy number. A direct estimate of the relative copy number, in terms of fold change from the expected control value, can be used for genotyping. CODEX is freely available as a Bioconductor R package at http://bioconductor.org/packages/CODEX/.

1.2.2 Read Depth Normalization

Due to the extremely high level of systemic bias in WES data, normalization is crucial in WES CNV calling. CODEX's multi-sample normalization model takes as input the WES depth of coverage, exon-wise GC content, and sample-wise total number of reads. Specifically, we denote Y as the coverage matrix with row i $(1 \le i \le n)$ corresponding to the $i\text{th}$ exon and column j $(1 \le j \le m)$ to the jth sample, GC_i as the GC content for exon i, and N_j as the total number of mapped reads for sample j. The "null" model, which reflects the expected coverage when there is no CNVs, is

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

$$\lambda_{ij} = N_j f_j(GC_i)\beta_i \exp\left(\sum_{k=1}^{K} g_{ik}h_{jk}\right),$$

where $f_j(GC_i)$ is the bias due to GC content for exon i sample j; β_i reflects the exon-specific bias due to length and capture and amplification efficiency of exon i; and $g_{ik}h_{jk}$ $(1 \le k \le K)$ are the kth latent Poisson factors for exon i and sample j. The goal of fitting the null model to the data is to estimate the various sources of biases, which can then be used for normalization.

We adopt a robust iterative maximum-likelihood algorithm for estimating the parameters of the null model. Briefly, in each iteration, we estimate $f(GC)$ by fitting a smoothing spline of $Y/N\beta\exp(g \times h^T)$ against the GC content, using the built-in function *smooth.spline* in R. β takes the value of the median of each row in $Y/Nf(GC)\exp(g \times h^T)$. The latent variables $g_{ik}h_{jk}$ $(1 \le k \le K)$ are estimated in the following steps: (i) take known h as covariates, fit n Poisson log-linear regressions with each row of Y as the response and corresponding row of $\log(Nf(GC)\beta)$ as the fixed offset; (ii) take known g as covariates, fit m Poisson log-linear regressions with each column of Y as the response and corresponding column of $\log(Nf(GC)\beta)$ as the fixed offset; (iii) apply SVD to the row-centered matrix $g \times h^T$ to obtain the K right singular vectors to update h. The third step ensures the uniqueness and orthogonality of the updated components, which forces the identifiability of $g_{ik}h_{jk}$ $(1 \le k \le K)$ (37). We fit the Poisson log-linear models with the built-in function *glm* in R. See below for details of the maximum-likelihood algorithm. Procedures for determining K, the number of latent Poisson factors, is discussed later in 1.4.3 Poisson Latent Factors and Choice of *K*.

Initialization

$\beta^{old} = 1^n, g = 0^{n\times K}, h = 0^{m\times K}$.

Iteration

i. For each sample j, fit a smoothing spline of $\left[Y/N_j\beta^{old}\exp(g \times h^T)\right]_{:j}$ to get $f_j(GC)$.

ii. For each exon i, update β_i as $\beta_i^{new} = \text{median}([Y/Nf(GC)\exp(g \times h^T)]_{i:})$.

iii. Denote $Z = Nf(GC)\beta^{new}$. Apply SVD to row-centered $\log(Y/Z)$ to obtain the K right singular vectors and use as h^{old}.

   a. Fit n Poisson log-linear regressions with $Y_{i:}$ as response, h^{old} as covariates, $\log(Z_{i:})$ as fixed offset to obtain updated estimates as g.

   b. Fit m Poisson log-linear regressions with $Y_{:j}$ as response, g as covariates, $\log(Z_{:j})$ as fixed offset to obtain updated estimates as h^{new}.

   c. Center each row of $g \times (h^{new})^T$ and apply SVD to the row-centered matrix to obtain the K right singular vectors to update h^{new}.

   d. Repeat steps a to c with $h^{old} = h^{new}$ until convergence to obtain h and g.

iv. Repeat steps i to iii with $\beta^{old} = \beta^{new}$ until convergence.

After the normalization procedure, we obtain $\hat{\lambda} = N\hat{\beta}\hat{f}(GC)\exp(\hat{g} \times \hat{h}^T)$, which is the expected "control coverage" in the event where there is no CNV. As described later, the observed coverage Y will be compared to the corresponding estimated control coverage $\hat{\lambda}$ to test for the presence of CNVs.

For CNV detection under case-control settings (e.g. tumor with normal) involving recurrent large chromosomal aberrations, CODEX estimates the exon-wise Poisson latent factor $\{g_{ik}\}$ using only the read depths in the control cohort, and then computes the terms $\{h_{jk}\}$ for the case samples by regression. This leads to higher sensitivity for detecting variants that are present only in the case samples. CODEX also includes two modes—"integer" mode that returns copy numbers as integers for germline CNV detection and "fraction" mode that returns fractional copy numbers for CNV detection of samples with heterogeneous genetic compositions.

1.2.3 CNV Detection and Copy Number Estimation

Proper normalization sets the stage for accurate segmentation and CNV calling. For germline CNV detection in normal samples, many CNVs are short and extend over only one or two exons. In this case, simple gene- or exon-level thresholding is sufficient.

For longer CNVs and for copy number estimation in tumors where the events are expected to be large and exhibit nested structure, we propose a Poisson likelihood-based recursive segmentation algorithm. Let $y_s, \ldots, y_t$ and $\lambda_s, \ldots, \lambda_t$ be the raw and estimated control coverage of the window spanning exon s to exon t. The values $\lambda_s, \ldots, \lambda_t$ are estimated by the normalization procedure described in the previous section, but suppressing the sample indicator j since we segment each sample separately. A joint cross-sample segmentation, as proposed in Zhang *et al.* (38), can also be applied and may yield more accurate results for detection of germline CNVs. Let $y_{s:t} = \sum_{i=s}^{t} y_i$ and $\lambda_{s:t} = \sum_{i=s}^{t} \lambda_i$. The scan statistic we use is $\max_{s,t} U(s,t)$, where

$$U(s,t) = \sup_{\mu} \left(\log \left(\frac{\mu^{y_{st}} \exp(-\mu)}{\lambda_{s:t}^{\ y_{st}} \exp(-\lambda_{s:t})} \right) \right) = y_{s:t} \log \left(\frac{y_{s:t}}{\lambda_{s:t}} \right) - (y_{s:t} - \lambda_{s:t})$$

The above is the generalized log-likelihood ratio of the alternative model, $y_{s:t} \sim \text{Poisson}(\mu)$ with μ arbitrary, versus the null model, $y_{s:t} \sim \text{Poisson}(\lambda_{s:t})$. The copy number estimate for the window is given by $2y_{s:t} / \lambda_{s:t}$.

Given the scan statistic, CODEX performs a circular binary segmentation procedure (39) using $U(s,t)$. We further use a modified Bayes Information Criterion (mBIC) to determine the number of change points P in our model (40),

$$\text{mBIC}(P) = \log \left(\frac{L_\tau}{L_0} \right) - \frac{1}{2} \sum_{\rho=0}^{P} \log(\hat{\tau}_{\rho+1} - \hat{\tau}_\rho) + (\frac{1}{2} - P) \log(n),$$

where the first term is the generalized log-likelihood ratio for the model with P change points versus the null model with no change points; τ_ρ $(1 \le \rho \le P)$ is the ρth change point, $1 = \tau_0 < \tau_1 < \cdots < \tau_P < \tau_{P+1} = n$; n is the number of exons. We report the segmentation with $\hat{P} = argmax_P mBIC(P)$. Compared with algorithms based on HMM such as XHMM (28) and EXCAVATOR (25), CODEX doesn't require the user to pre-specify unknown parameters, such as expected distance between exons, exon-wise CNV rate, and average number of exons in a CNV. These quantities are often hard to set a priori without a large relevant training data set, and in

many cases have to be chosen arbitrarily. Post-segmentation, CODEX outputs an estimate of the relative copy number in terms of fold change from the expected control coverage, rather than a binary categorization of deletion and duplication as in CoNIFER (26) and XHMM (28).

1.2.4 Calling Germline Variations from HapMap Samples

To examine the accuracy of CODEX and to illustrate its application, we use a publicly available WES data set from the 1000 Genomes Project Phase 1 release (29) containing 90 healthy individuals. 46 samples are sequenced at the Washington University Genome Sequencing Center (captured by HSGC VCRome) and 44 at the Baylor College of Medicine (captured by SureSelect All Exon V2). All samples have Omni and Axiom genotypes and have more than 70% of exome targets covered to 20x or more. Sex is well balanced (44 males and 46 females) and population (40 Utah residents with northern and western European ancestry (CEU), 24 Japanese people from Tokyo (JPT), and 26 Yoruba people from Ibadan (YRI)) adds a potential source of latent variation.

Effectiveness of normalization procedure

We first examine the effectiveness of CODEX's proposed normalization model on the 1000 Genomes Project WES data set (29). Previous studies have shown that read depth has a unimodal relationship with GC content—regions with high or low GC content tend to have decreased read depth (41). In our smoothed estimates of $f_j(GC)$, we find that most but not all samples have a unimodal shape for this function. We show the predicted values of $f_j(GC)$ for 4 typical samples in Figure 1.2. Interestingly, we found that some samples have estimates with multiple peaks in $f_j(GC)$, which suggests that a parametric functional form assuming unimodality may be too simplistic. Comparing across samples, we see that the function $f_j(GC)$ changes in shape and not just by a scaling factor. Therefore, the GC content bias is not linear across samples and thus cannot be fully captured by linear latent factor models. This motivates the separate nonparametric term in our model for GC bias.

We further compare the normalization result of CODEX against that of SVD based method using array-based CNV calls from the International HapMap Consortium (30) on the same samples we analyze. For different categories of CNV events, namely, homozygous deletions, heterozygous deletions, and duplications, we use direct thresholding of $\log(Y/\hat{\lambda})$ to draw receiver operating characteristic (ROC) curves of our model, where $\hat{\lambda}$ is the estimated control coverage from CODEX's normalization procedure. The ROC curves for SVD-based normalization are drawn by thresholding on the residuals obtained by subtracting the first K PCs from the original read depth Y. Analysis is carried out for each of the following category of events separately: common homozygous deletion, common heterozygous deletion, common duplication, rare heterozygous deletion, and rare duplication (Figure 1.3). There are no rare homozygous deletions as all of the rare deletions from the HapMap CNV call set are present in only heterozygous form. We see that CODEX's normalization procedure leads to a better signal-to-noise ratio for both common and rare CNVs, and for both deletions and duplications (Figure 1.3).

Accuracy of CNV calling

We next compare the accuracy of CODEX to existing approaches that are designed for population-based CNV calling. These programs include CoNIFER (26), XHMM (28), and EXCAVATOR (25) in its "pooling" mode, for which we added four additional samples as controls.

The number of calls made by each program on each chromosome sample, broken down into common and rare calls, is shown in Table 1.1. Globally, CODEX detects twice as many CNV events as XHMM does and nearly 10 times as many as CoNIFER does, while EXCAVATOR and CODEX have comparable number of calls. CoNIFER detects the fewest CNVs in total, which agrees with comparisons against EXCAVATOR made in Magi *et al.* (25). Since CoNIFER does not automatically choose the number of PCs, we fix the number of PCs filtered out by CoNIFER at 4, agreeing with the selection made by XHMM so as to make the two SVD-based programs comparable. The choice of 4 PCs in normalization should not account for the low number of calls made by CoNIFER, since through the scree plot output by CoNIFER, we find the curve of relative

contributed variance to be still significantly decreasing at 4, indicating that the choice of 4 is conservative. A large proportion of XHMM and CoNIFER calls are rare (<5%) variants—52.46% (501/955) and 83.07% (157/189) respectively. Despite the bias in sensitivity of XHMM and CoNIFER towards rare variants, CODEX detects even more rare CNVs in total as well as proportionately more common ones. Notably, the number of latent factors K selected by CODEX is for most chromosomes one less than the number of PCs excluded by XHMM across the genome. Furthermore, CODEX and XHMM tends to detect shorter CNVs compared to CoNIFER and EXCAVATOR in units of both kb (Figure 1.4a) and exon (Figure 1.4b).

We assess the CNV calls made by the four methods by comparing to calls reported by the International HapMap Consortium (30), McCarroll *et al.* (31), and Conrad *et al.* (32) in the same 90 HapMap samples. The International HapMap 3 Consortium produced a clean CNV call set by merging and utilizing probe-level intensity from both Affymetrix and Illumina arrays, containing 856 copy number polymorphisms (CNPs) with a 99.0% mean call rate and 0.3% Mendelian inconsistency (30). Separately, McCarroll *et al.* developed a map consisting of 1320 CNVs at 2-kb breakpoint resolution by joint analysis of Affymetrix SNP array, array CGH (42) and fosmid end-sequence-pair data (31, 43). The third source of validation we use is the call set from Conrad *et al.,* who used Nimblegen tiling oligonucleotide arrays to generate a map of 11,700 CNVs greater than 443 base pairs, of which 8,599 have been validated independently (32). The genotyped CNPs from these three cohort studies that overlap with exon regions (73, 123, and 377 in total respectively) are used as “validation set” to assess sensitivity and specificity of the four methods compared in Table 1.1. Figure 1.5 shows the precision and recall rates (precision is the proportion of calls made by the program that overlap with validation set, and recall is the proportion of the CNVs in validation set that are called.) The different programs vary considerably in precision and recall rate. CODEX has the highest F-measure (harmonic mean of precision and recall) for both common and rare CNVs. XHMM performs well in detecting rare variants but is insensitive to common ones. CoNIFER has the highest precision when comparing

against calls from the International HapMap Consortium (Figure 1.5a) and McCarroll *et al.* (Figure 1.5c) but gives poor results against Conrad *et al.* (Figure 1.5b). Furthermore, the high precision of CoNIFER come with significant sacrifice on recall. See Table 1.2 for detailed comparison results based on the three SNP array metrics.

1.2.5 Sensitivity Assessment with Spike-in Study

We next conduct an in silico spike-in study to assess the sensitivity of the different methods at varying population frequencies. Starting with the WES data from chromosome 20 of the $m = 90$ HapMap samples analysed in the previous Section, we spike CNV signals in to copy-number-neutral regions. We define a region to be copy-number-neutral if it doesn't overlap with CNV calls made by CODEX, XHMM, EXCAVATOR, and CoNIFER nor with previously reported CNV regions by DGV (http://dgv.tcag.ca/dgv/app/) and dbVar (http://www.ncbi.nlm.nih.gov/dbvar/). Of the 3966 exon targets on chromosome 20, 1035 pass this criterion for copy-number-neutral. We consider only heterozygous deletions of two different lengths (5 and 10 exons) and varying population frequencies $p \in \{5\%, 10\%, \ldots, 95\%\}$. We focus on heterozygous deletions because (i) homozygous deletions are easily detectable by all methods; (ii) heterozygous deletions with frequency p in the population have exactly the same detection accuracy as duplications with frequency $1 - p$. Specifically, for deletions with population frequencies greater than 50%, copy-number-neutral states are reported as duplications whereas deletions are reported as normal events, since all copy number events are defined in reference to a population average. Events are centered at every hundredth exon and $m \times p$ samples are randomly chosen to be carriers. To generate CNV signals for heterozygous deletions, we reduce the raw depth of coverage for exons spanned by the CNV from y to $\frac{c}{2} \times y$, where c is sampled from a normal distribution with mean 1 and standard deviation 0.1.

We apply CODEX to these spike-in data sets and compare it to SVD-based normalization followed by HMM-based segmentation. For the latter, we remove the first K principal components (PCs) from the read depth matrix and transform the residuals to z-scores for each sample

separately. The z-scores are then segmented by a HMM whose parameters are set as the default values in XHMM. The specificity of both approaches is controlled to be higher than 99%. The sensitivities for short CNV (5 exons) and long CNV (10 exons) at different population frequency levels are shown in Figure 1.6. We see that both approaches attain high sensitivity for rare CNVs, and both have decreased sensitivity for common CNV events. The sensitivity of CODEX is higher than that of the existing approach for both rare and common variants (Figure 1.6). For CNV events with frequencies around 50%, both methods have the lowest power due to the fact that the CNV signals are falsely filtered out by a sample-wise latent factor (Figure 1.6). Also, shorter CNV events are more often missed by the SVD approach whereas CODEX has comparable sensitivity for short and long variants at this scale (Figure 1.6).

To gain a better understanding of what the latent factors in CODEX and SVD-based methods are capturing, we show in Figure 1.7 the correlation of the latent factors to measurable quantities. The exon-wise latent factors in both models and the estimated value of β in CODEX are compared to GC content, mean exon coverage, and true copy number. The sample-wise latent factors in both models are compared to center, batch, population, and total coverage (N). Based on these correlations, we make the following observations: First, mean exon coverage, represented by the pseudo-reference sample $\left\{(\prod_{v=1}^{m} Y_{iv})^{1/m}: 1 \le i \le n\right\}$, is captured by β in (correlation coefficient 0.99) in CODEX and the first exon-wise PC in SVD (correlation coefficient -0.98). Exon length and capture and amplification efficiency are confounded in this exon-specific bias and there is no way, nor any need, to estimate these individual quantities separately. Second, GC content is correlated with the third exon-wise PC in SVD (correlation coefficient -0.75). CODEX specifically models the GC content bias for each sample by the term $\left\{f_j(GC): 1 \le j \le m\right\}$, and as we show later, the bias cannot be fully captured by a linear PC. Third, a CNV that is more frequent in the population has higher absolute correlation between copy number state and the exon-wise latent factors in both CODEX (-0.22) and SVD (0.57). This is why sensitivity is lower for common CNVs. Finally, other known sources of bias, such as sequencing center and

batch, are captured by sample-wise latent factors in both CODEX (correlation coefficient -1 and 0.74) and SVD (correlation coefficient 0.97 and -0.71). In this data set, population doesn't seem to be captured by any of the top latent factors.

1.2.6 Analysis of Whole Exome Sequencing of Neuroblastoma

We also analyze a WES data set consisting of 222 paired tumor/normal (blood leukocyte) samples of individuals older than 18 months of age at diagnosis with stage-4 neuroblastoma from the TARGET Project (33). WES of native and whole genome amplified DNA of ~33Mb regions yields a 124X average coverage, with 87% of bases suitable for mutation detection (33). Our discussion here focuses on the well characterized *ATRX* gene region (33-35). The TARGET Project reported recurrent focal deletions with a complex nested structure spanning the *ATRX* gene. Since there are matched normal samples for this study that have also been sequenced by the same technology, the TARGET calls were made by comparing each tumor sample to its matched normal. This allows us to compare the effectiveness of CODEX's normalization model to that of taking a log ratio to the matched normal coverage. Also, focusing on this well characterized region allows us to demonstrate in accuracy of CODEX for handling recurrent complex nested events.

The RPKM (reads per kilo bases per million reads) for each exon and each sample are plotted in Figure 1.8a. The RPKM profiles are very noisy and do not show any clear decrease in this region in any of the samples, highlighting the need for normalization. For comparison, we also show the TARGET Project's initial analysis, which reported 16 multiexon deletions within *ATRX* by comparing tumor to matched normal samples (33). Specifically, we repeat their analysis by thresholding the $\log_2$-ratio of RPKM in tumor to RPKM in normal samples, illustrated in Figure 1.8b. Figure 1.8c shows the normalized intensities given by CODEX, which detects 18 samples with somatic focal deletions. We also apply XHMM to the tumor data set and detect 14 samples with focal deletions (Figure 1.8d).

Of the 18 samples with somatic deletions detected by CODEX, three are also called by the TARGET Project but missed by XHMM; one is detected by XHMM and CODEX with exactly the same breakpoints but is missed by the Target Project; one is uniquely called by CODEX (Table 1.3a). The sample uniquely called by CODEX is a small deletion that overlaps significantly with deletions called in other samples. Detailed CNV calling and genotyping results by each method are in Table 1.3b-d and the genome-wide blood and tumor CNV events discovered by CODEX are summarized in Table 1.4. The comprehensive analysis results will be published separately.

It is clear by visual comparison of Figure 1.8c to Figure 1.8b and Figure 1.8b that the read depth normalization method within CODEX gives better signal to noise ratio than the SVD based normalization method in XHMM (note the difference in range of the y-axes) and also better than the commonly prescribed method of normalizing to matched normal controls. This illustrates that by borrowing information across a large cohort, the estimated control coverage of $\hat{\lambda}$ from our normalization model is more effective in capturing the biases in whole exome sequencing than the matched normal. Whereas the matched normal sample is important to distinguish between germline and somatic variants, CODEX's normalization procedure can be used in case of unavailability of blood samples or contamination of blood samples from circulating tumor cells. When matched normal is available, somatic status can be determined by comparing CODEX calls in tumor to those in normal. This example also shows that CODEX's segmentation algorithm performs well in detecting multiexon CNVs with a nested structure, and that it successfully detected a rare CNVs (18/222=8.11%) in a clinical setting.

1.3 Discussion

Here we propose CODEX, a normalization and CNV detection method for WES data. CODEX includes a normalization model with non-parametric functional terms for GC content and Poisson latent factors for biases that are not directly quantifiable. We show that both parts of the normalization model are necessary for WES data. CODEX segments the genome using a

Poisson likelihood model based on the control coverage $\hat{\lambda}$ estimated during the normalization step. CODEX can be applied to both normal and tumor genome analysis.

We show through several data sets that CODEX's multi-sample normalization procedure offers higher sensitivity and specificity for detection and genotyping of both common and rare CNVs. The distinguishing features of CODEX compared to existing methods are: (i) CODEX doesn't require matched normal samples as controls for normalization; (ii) The Poisson log-linear model fits better with the WES count data than SVD approaches; (iii) Dependence on GC content is modelled by a flexible nonparametric function in CODEX allowing it to capture non-linear biases; (iv) CODEX implements the BIC criterion for choosing the number of latent variables, which gives a conservative normalization on simulated and real data sets; (v) Compared to HMM-based segmentation procedures, the segmentation procedure in CODEX is completely off-the-shelf and doesn't require large relevant training set; (vi) CODEX estimates relative copy number, which can be converted to genotypes by thresholding, rather than broad categorizations (deletion, duplication, and copy number neutral states).

We carry out simulation studies by spiking in CNV signals to WES read depth data from copy-number-neutral regions. We show that CODEX has higher power compared to SVD based method followed by HMM, although both methods suffer from common CNV events. We also investigate the nature of the exon- and sample-wise terms and Poisson factors in CODEX, PCs extracted by SVD, and other directly known biases and artifacts. We show that PCs from SVD obtained by unsupervised learning are correlated by the terms specifically modelled and quantified by CODEX and that the GC content correlates with one PC from SVD with correlation coefficient -0.75, which, again, is specifically modelled by CODEX. Developing a robust method that can detect common CNVs from background noise with high sensitivities may be a future direction to get focused on.

We compare CODEX's performance against direct calling results from other existing methods on the 1000 Genomes Project WES data set and show that CODEX is more accurate by

comparing CNV calls by WES against three gold standard SNP array CNV call sets. Since CoNIFER and EXCAVATOR detect a significant proportion of CNVs with lengths greater than 200 kb whereas CODEX and XHMM return much shorter CNVs (Figure 1.4), we don't exclude any CNV calls by SNP arrays so as to get more "reliable" gold standards as does Fromer *et al.* (28), despite the fact that array based methods, when compared to next-generation sequencing, don't have as good resolutions. This might explain why the overall sensitivity/recall rates are no larger than 0.6 for all methods (Figure 1.5, Table 1.2). Another possible explanation lie in that due to the discrete nature of WES data, read depth is used as the only inference to detect CNVs, which has only exon-level resolution and thus lower power in detecting short CNVs compared to split-read and paired-end-mapping methods developed for WGS. Despite the limitations, WES has been used and is still being used as a preferred method of choice for large-scale studies.

With a clinically relevant example on detecting rare somatic CNVs within *ATRX* associated with neuroblastoma, CODEX is shown to be applicable to a wide range of study designs for CNV detection using WES data. Specifically, we show that CODEX doesn't require matched normal controls for normalization and is able to detect previously reported CNVs within tumor samples more accurately compared to SVD-based method. Matched blood samples, when available, can be used to distinguish somatic CNVs from germline ones. However, under most circumstances, the normal samples are often unavailable, incomplete, or unmatched, which drives the need for normalization using cases only. The genome-wide CNV results based on this data set are available and will be compared against other metrics (matched microarrays, whole-genome sequencing, RNA-sequencing, etc.) and validated on bench. The comprehensive analysis results will be published elsewhere.

1.4 Methods

1.4.1 Sample Selection and Target Filtering

To have as much sample- and exon-wise homogeneity as possible and to make sure that our normalization algorithm converges without being deviated by extreme values, we adopt a sample

selection and target filtering strategy before applying our proposed normalization method to the read depth data. Specifically, for reducing artifacts, we recommend that all of the samples be sequenced by the same platform. We further filter out exons that: (i) have extremely low coverage (median read depth across all samples less than 20, which mostly reflect capture failure); (ii) are extremely short (less than 20 base pairs); (iii) are hard to map (mappability less than 0.9); (iv) have extreme GC content (less than 20% or greater than 80%). These default thresholds for quality control (QC) are recommended but are also user-tuneable and thus can be adapted to different sequencing protocols. We show in Table 1.1 that with the above QC thresholds, 9.74% of exon targets are excluded in the data. Details on computation of GC content, mappability, and depth of coverage are provided in 1.4.2 Depth of Coverage, GC Content, and Mappability.

1.4.2 Depth of Coverage, GC Content, and Mappability

Depth of coverage for each exon is computed as the number of reads (with mapping quality greater than a user-defined threshold) that overlap with the exon. To calculate the exonic mappability, we first construct consecutive reads that are one base pair (bp) apart along the exon. The length of the reads is set to be the same as that from the sequencing technology and the sequences are taken from the hg19 reference. We then find possible positions across the genome that the reads can map to allowing for a default number of mismatches (2 for the 1000 Genomes Project data set in our study which has read 100). Finally we compute the mean of the probabilities that the overlapped reads map to the target places where they are generated and use this as the mappability of the exon.

We compare our computed exonic mappability with the number of overlapped segmental duplications from the Segmental Duplication Database. Results show that not all segmental duplication regions are hard to map and thus it is not wise to directly filter out exons that overlap with segmental duplications (Figure 1.9a). As a comparison, we also compute the sequence complexity—percentage of bases within exons soft masked by RepeatMasker (http://www.repeatmasker.org/) using PLINK/SEQ (http://pngu.mgh.harvard.edu/purcell/plink/),

which is the filtering strategy adopted by XHMM. It turns out that not only XHMM has an overly stringent threshold on sequence complexity/mappability (Figure 1.9b), but also it includes other outlier removal steps, such as removing samples with coverage that are empirical outliers, filtering out targets with a standard deviation of PCA-normalized z-score greater than 30, etc. These additional empirical ways of excluding samples and targets might treat true signals as outliers and remove them.

1.4.3 Poisson Latent Factors and Choice of *K*

Some sources of bias in whole exome sequencing can be directly measured (GC content, mappability, and exon size). However, there are other unmeasurable sample- and target-specific biases that are amplified during the library preparation and sequencing experiment. The latent Poisson factors $\{g_{ik}\}$ and $\{h_{jk}\}$ are designed to capture and decompose these unobserved systemic bias in a log-additive manner. Such latent factor models have been shown to be effective in the analysis of microarray data (44-47), and have also recently been applied to NGS data. Both CoNIFER (26) and XHMM (28) use latent factor models to remove systemic bias, but their models assume continuous measurements with Gaussian noise structure, while CODEX is based on a Poisson log-linear model, which is more suitable for modeling the discrete counts in WES data, especially when there is high variance in depth of coverage between exons. The latent factor terms in the normalization model resemble those used in Lee *et al.* (37) for microRNA profiling. In particular, the identifiability constraints in Lee *et al.* also apply to our case, and our iterative maximum-likelihood estimation procedure ensures identifiability.

A common downfall of latent factor models is that true CNV signals may correlate with and influence the top K latent factors. Thus, the number of latent factors, K, is a crucial parameter. If K is chosen to be too large, some bona fide CNV signals, especially those for common CNVs, will be dampened during normalization. On the other hand, if K is too small, residual artifacts will remain and inflate the type I error rate. CoNIFER (26) adopts a common practice for choosing the number of factors in latent variable models, which is to draw the scree

plot with the number of components on the X-axis and the corresponding contributed variance on the Y-axis. If there is an "elbow" in the scree plot, then K is chosen at the position of the elbow (Figure 1.10a). However, in most cases there is no detectable elbow, which is why many existing methods arbitrarily set the value of K. XHMM (28) removes components with variance $0.7/m$ or higher, where m is the number of components (samples) and 0.7 is a user-tuneable parameter arbitrarily set as default.

We apply two additional statistical procedures of choosing this crucial model tuning parameter: Akaike information criterion (AIC, Figure 1.10b) and Bayes information criterion (BIC, Figure 1.10c).

$$\mathrm{AIC} = 2\ln(L) - 2k$$

$$\mathrm{BIC} = 2\ln(L) - k\ln(n)$$

where L is the likelihood for the estimated model, k is the number of parameters in the model, and n is the number of data points. Both criteria reward goodness of fit with a penalty term that is an increasing function of the number of parameters in the model. AIC penalizes the number of parameters less strongly than does BIC, and thus the model chosen by AIC removes more latent factors than that chosen by BIC. CODEX reports all three statistical metrics (AIC, BIC, percentage of variance explained) and uses BIC as the default method to determine the number of K. Since false positives can be screened out through a closer examination of the post-segmentation data, whereas CNV signals removed in the normalization step cannot be recovered, CODEX opts for a more conservative normalization that, when in doubt, uses a smaller value of K.

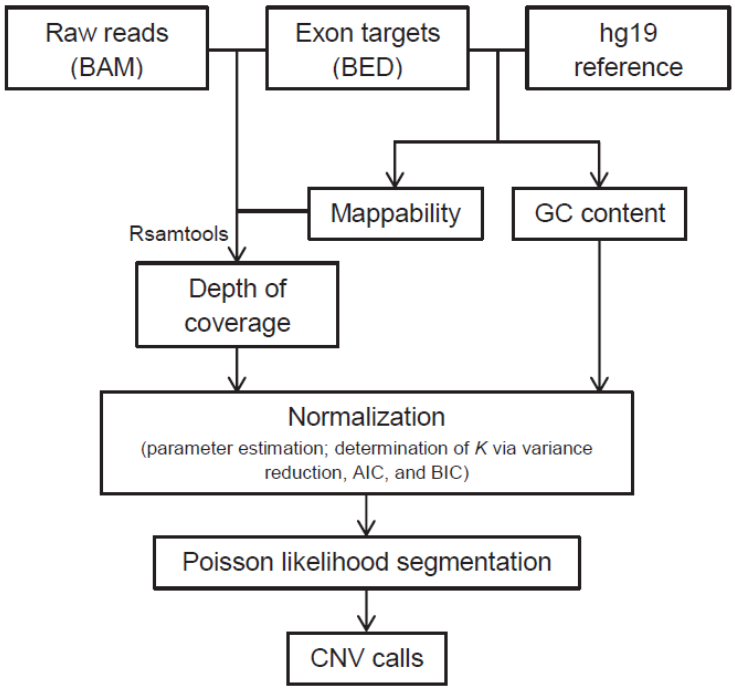


Figure 1.1: ***A flowchart outlining the procedures of CODEX in normalizing WES read depth and calling CNV.*** *The first step is computing GC content, mappability, and depth of coverage using Rsamtools with QC measures. The multi-sample normalization model by CODEX is then applied to remove biases and artifacts introduced by GC content, exon targeting and amplification efficiency, and latent systemic artifacts. The Poisson likelihood-based segmentation algorithm gives final CNV calls with copy number estimates.*

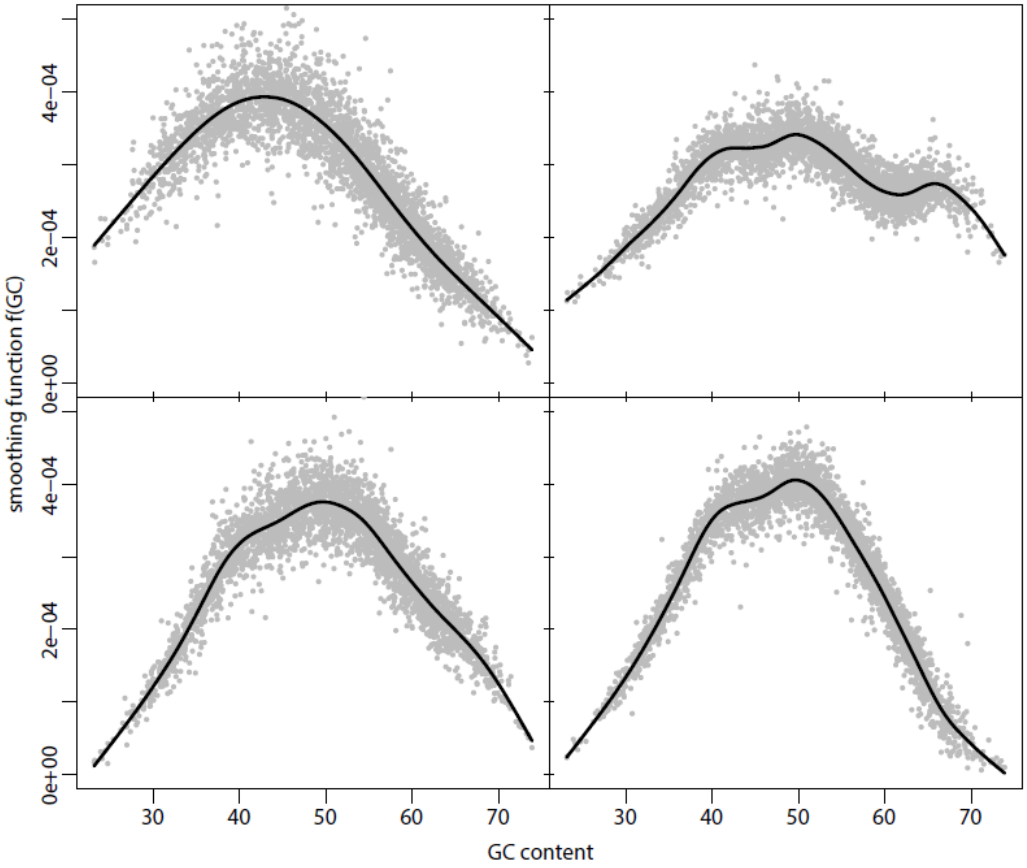


Figure 1.2: ***Predicted values of $f(GC)$ for 4 samples from the 1000 Genomes Project data set.*** *Most patterns agree with previous observations that read depth has a unimodal relationship with GC content. However, dual modality is also observed. Furthermore, the function changes in shape and not just by a scaling factor.*

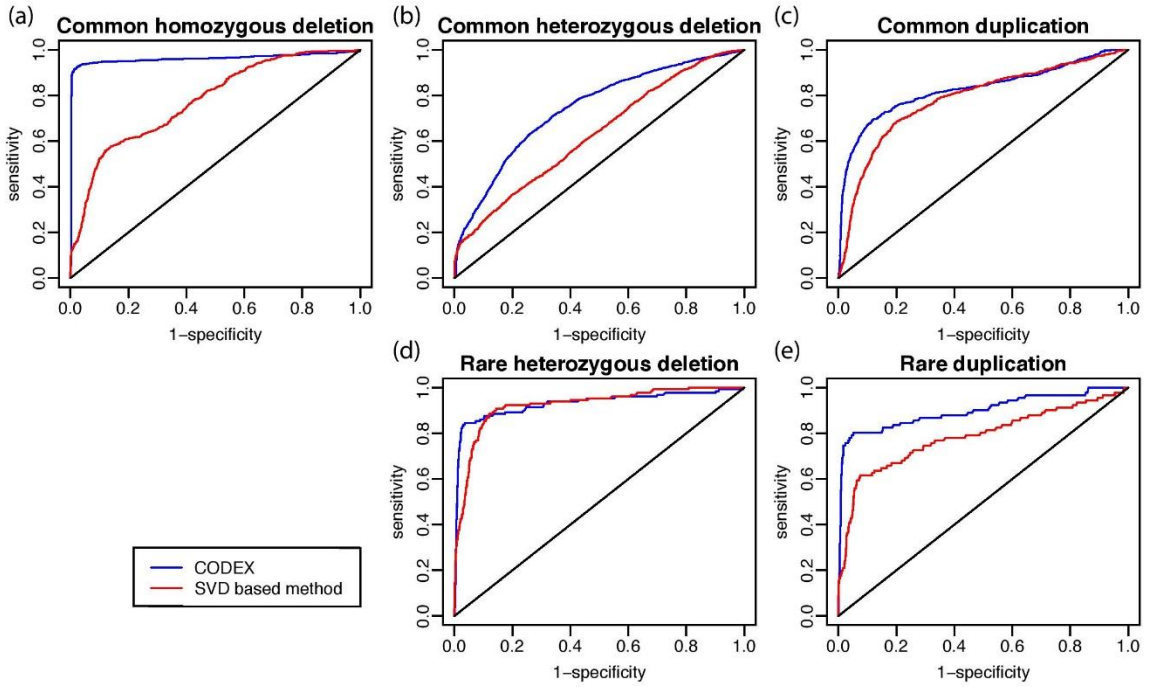


Figure 1.3: ***ROC curves of read depth normalization by CODEX and SVD-based method.*** *Gold standard is taken from the International HapMap Consortium SNP array CNV call set. The input for CODEX is the $\log_2$-ratio of the original read depth Y versus the estimated control coverage $\hat{\lambda}$; the input for SVD-based method is the residual obtained by subtracting the principal components from the original read depth Y. For common CNVs shown in (a), (b), and (c), CODEX performs significantly better since SVD-based methods are optimized for rare CNV detection; for rare CNVs shown in (d) and (e), the two methods tend to have similar power for rare heterozygous deletions whereas CODEX performs better in detecting rare duplications. Of the 90 samples we analyze, there is no rare heterozygous deletion from the HapMap call set that we can use as a gold standard.*

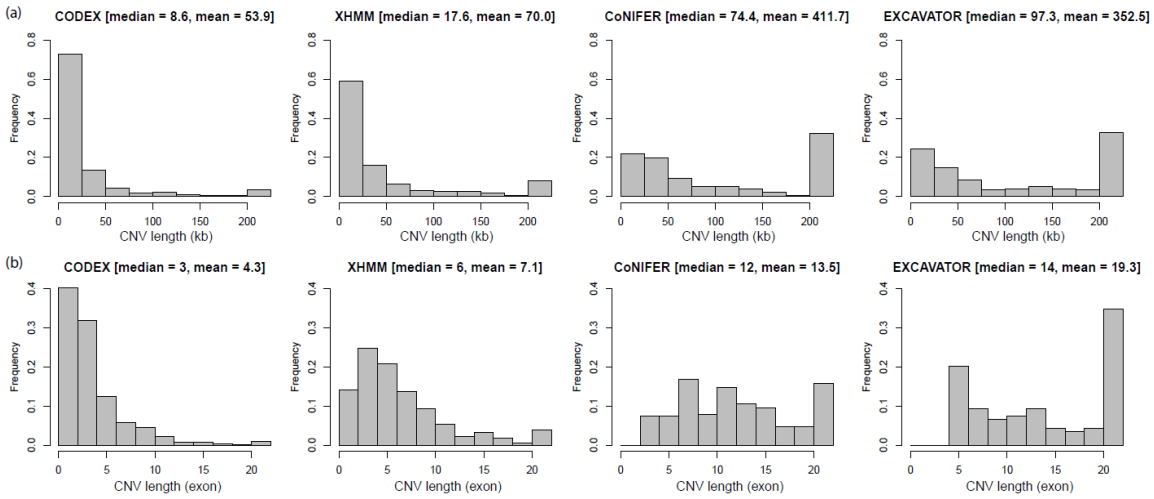


Figure 1.4: ***Lengths of CNV calls by CODEX, XHMM, CoNIFER, and EXCAVATOR.*** *Genomics lengths of CNVs (a) and number of exons in CNV regions (b) are compared across four different methods. CODEX and XHMM detects more short CNVs whereas CoNIFER and EXCAVATOR return significant proportion of CNVs with lengths greater than 200 kb/20 exons.*

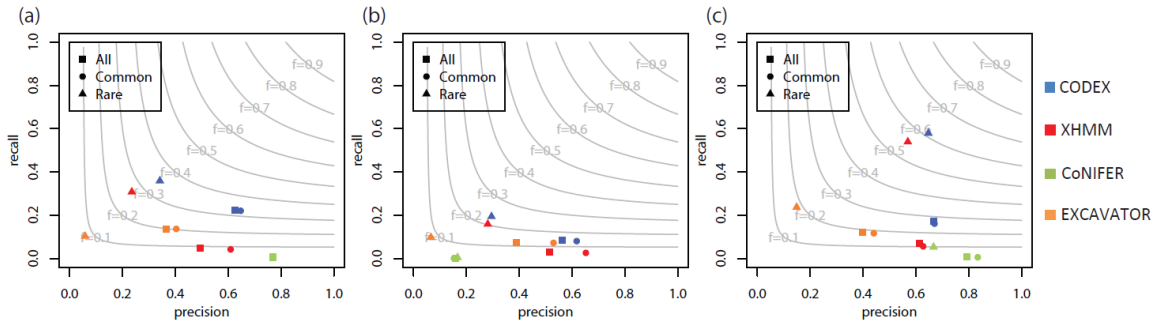


Figure 1.5: ***Assessment of CNV calls on the 1000 Genomes Project by array-based methods.*** *CNV calls by CODEX, XHMM, CoNIFER, and EXCAVATOR are validated against genotyping calls from International HapMap Consortium (a), Conrad et al. (b), and McCarroll et al. (c). CODEX returns well-balanced precision and recall rates with highest F-measures (grey contours shown harmonic means of precision and recall rates) among all methods for detection of common, rare, and all CNVs.*

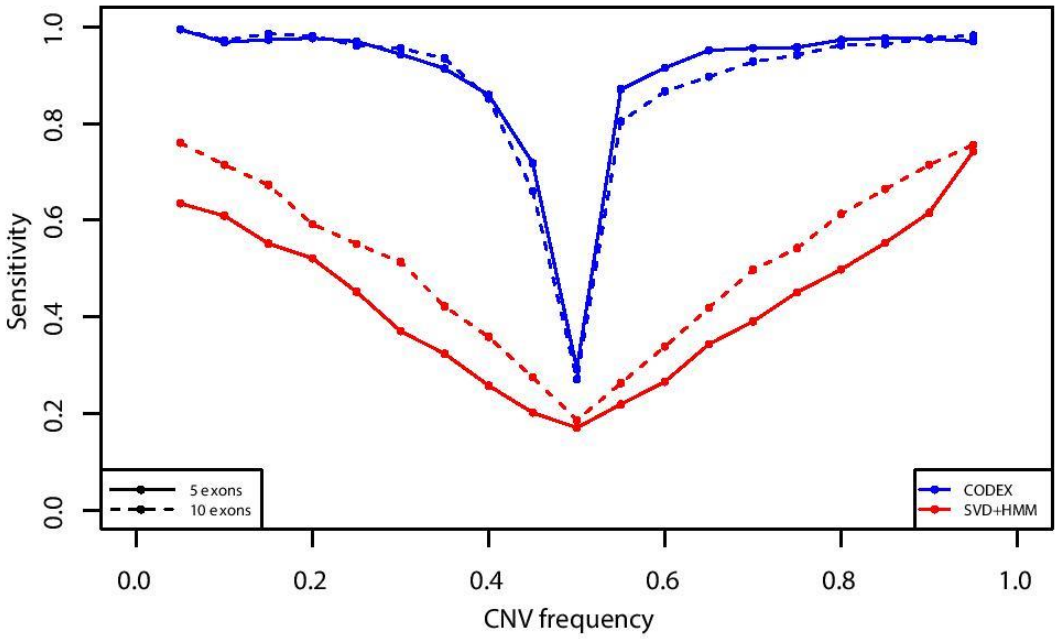


Figure 1.6: ***Power analysis of CODEX and SVD-based method on simulation data set.*** *Sensitivities are obtained by averaging results from 10 simulations. Both methods suffer from "common" CNV events (CNVs with frequencies around 50%). When CNV frequency exceeds 50%, deletions and copy-neutral states are detected as copy-neutral states and duplications instead, which recovers the sensitivities. CODEX performs better compared to SVD-based methods with higher power. Longer CNVs are generally easier to be detected.*

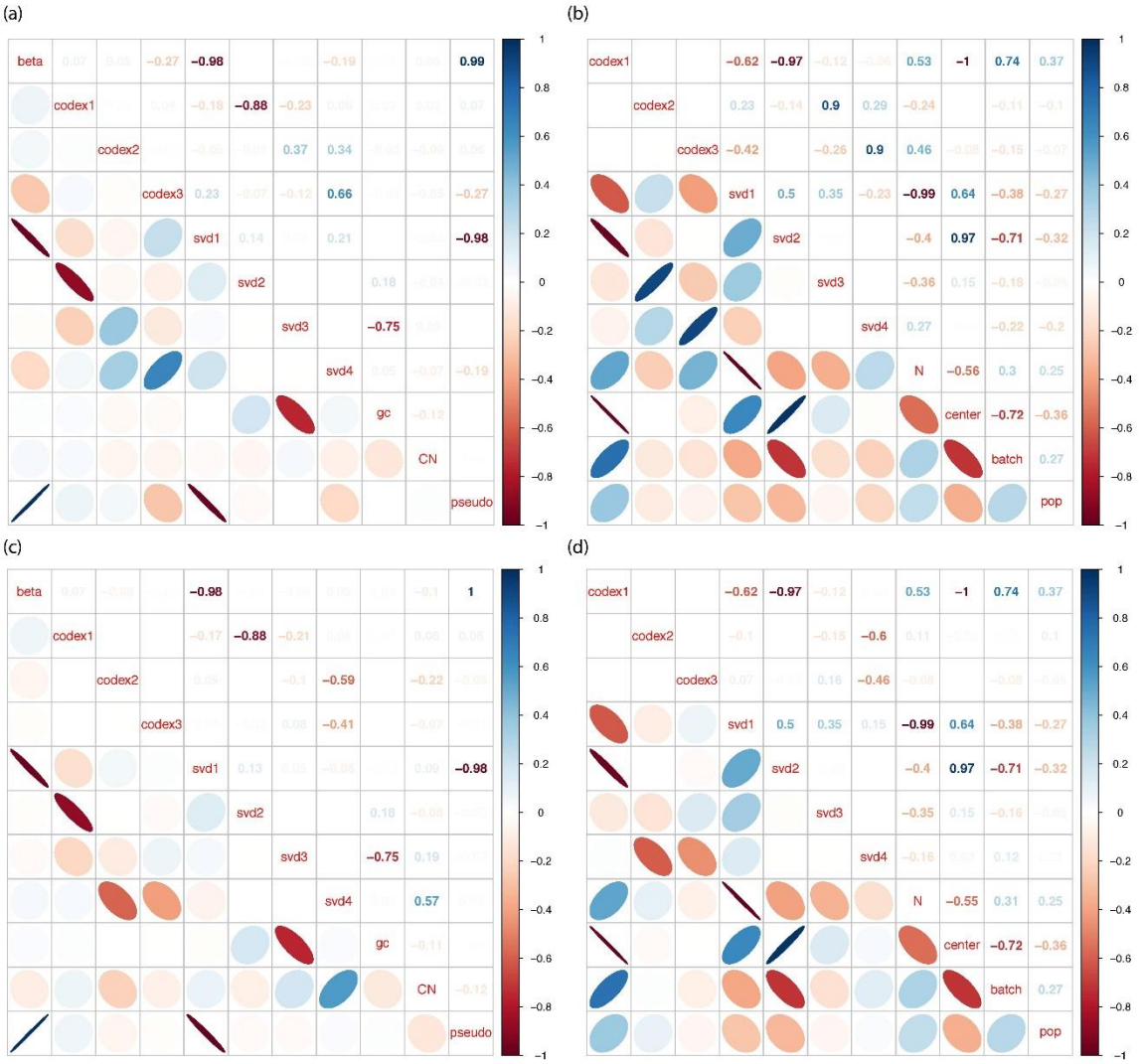


Figure 1.7: ***Correlation matrix plot of biases and artifacts shown in both exon-wise and sample-wise fashion.*** *β, exon-wise latent factors, GC content, copy-number state, and pseudo-reference genome are interrogated in (a) and (c). Sample-wise latent factors, total number of reads per sample, sequencing centers, batch effects, and population are shown in (b) and (d). (a) and (b) are for spike-in CNV events with frequency 0.1 and (c) and (d) are for spike-in CNV events with frequency 0.4. β and first exon-wise PC in SVD highly correlate with pseudo-reference genome. GC content is correlated with the third exon-wise PC in SVD with correlation coefficient -0.75. Copy-number states show higher correlation for spiked-in CNVs with higher frequencies. Sequencing centers and batch effects are captured by latent factors whereas population doesn't seem to add too much variation to the CNV signals.*

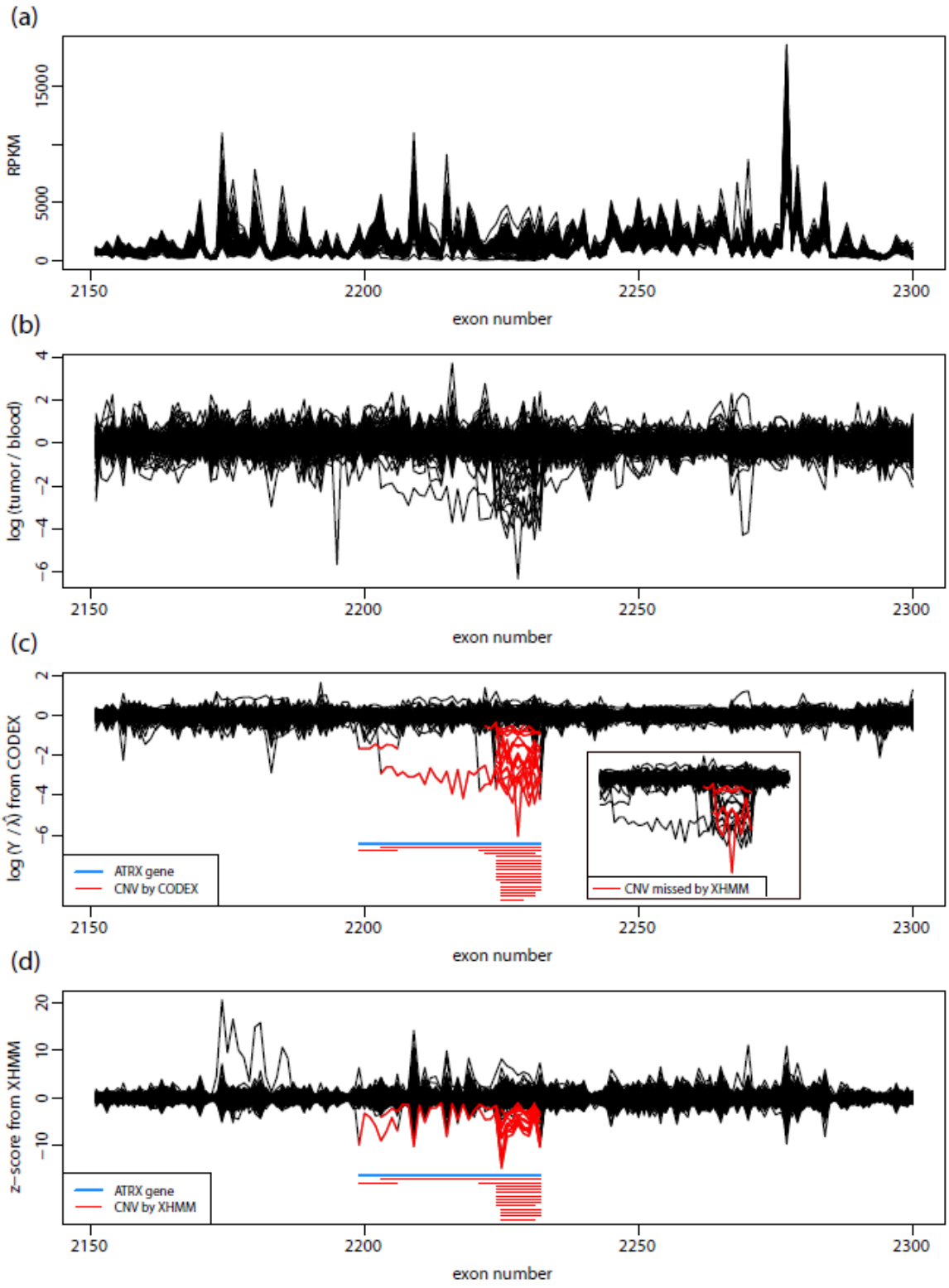

(a)
RPKM
0
5000
15000
2150
2200
2250
2300
exon number
(b)
log (tumor / blood)
4
2
0
−2
−4
−6
2150
2200
2250
2300
exon number
(c)
log (Y / λ̂) from CODEX
2
0
−2
−4
−6
ATRX gene
CNV by CODEX
CNV missed by XHMM
2150
2200
2250
2300
exon number
(d)
z−score from XHMM
20
10
0
−10
ATRX gene
CNV by XHMM
2150
2200
2250
2300
exon number

Figure 1.8: ***Detection of rare somatic deletions within ATRX by WES of 222 neuroblastoma matched tumor/blood samples.*** *Location of ATRX is shown as blue bars in c and d. (a) RPKM computed from the tumor samples. There is no clear visual indication of presence of somatic CNVs from these raw quantities. (b) Log_2-ratio of tumor versus blood read depth. Initial analysis by the TARGET Project did careful inspection of these values and discovered 17 samples with focal deletions. (c) log_2-ratio of the original tumor read depth Y versus the estimated control coverage $\hat{\lambda}$ (model fitted on tumor data set only) by CODEX. Poisson likelihood-based segmentation algorithm by CODEX discovers 18 samples (red bars) with somatic deletions that exhibit a nested structure across samples. The 4 samples that are called by CODEX but not by XHMM are colored in red in the embedded window. (d) XHMM's direct output: z-scores normalized by principal component analysis. The HMM calling algorithm by XHMM detects 14 samples (red bars) with somatic deletions.*

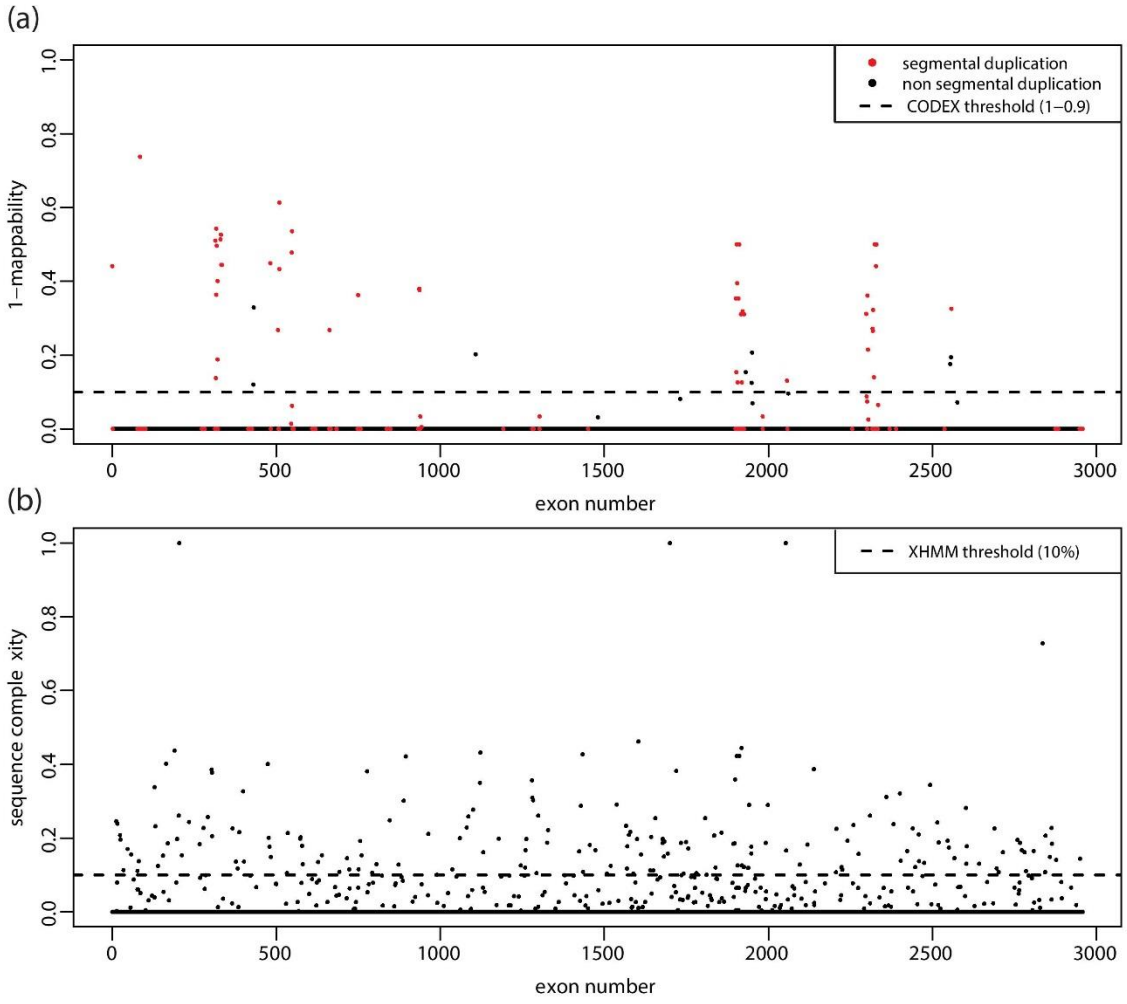


Figure 1.9: ***Filtering strategies on mappability and sequence complexity by CODEX and XHMM.*** *Computation results from chromosome 22 are shown with filtering thresholds in dashed lines. (a) Mappability computed by CODEX. Exons that overlap with previously reported segmental duplications are marked in red. (b) Sequence complexity used in pre-filtering step by XHMM.*

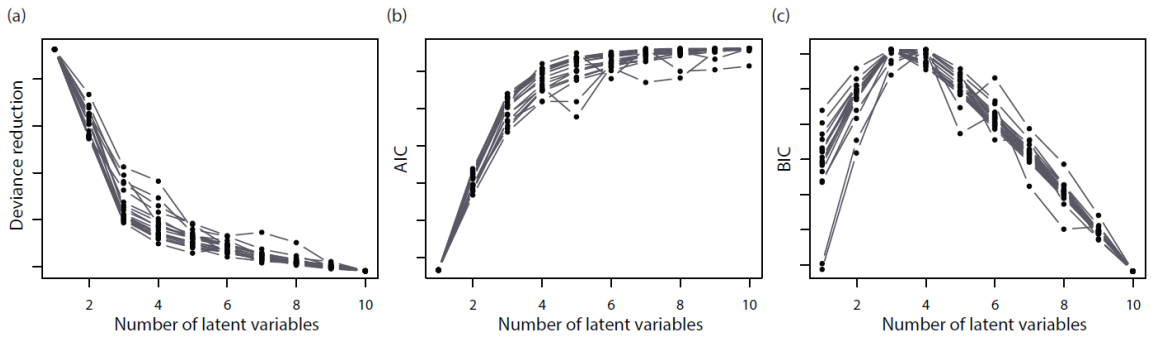


Figure 1.10: ***Choice of K, number of latent Poisson factors.*** *Remaining variance in the read depth data (a), AIC (b), and BIC (c) are used as three different metrics, which yield similar models with K optimally set at 3 or 4. Each line represents result of one chromosome. Suggested K is agreeable across the genome.*

Chr	Number of targets (before/after QC)	CODEX		XHMM		CoNIFER		EXCAVATOR CNVs
		K	CNVs	PCs	CNVs	PCs	CNVs	
1	15 426/14 101	3	361 (301-60)	4	129 (56-73)		36 (13-23)	263 (236-27)
2	9640/8956	3	54 (14-40)	4	51 (16-35)		6 (0-6)	15 (0-15)
3	8267/7775	3	13 (0-13)	4	8 (0-8)		4 (0-4)	5 (0-5)
4	5519/5157	4	27 (16-11)	4	20 (7-13)		16 (6-10)	91 (86-5)
5	6403/5950	3	163 (143-20)	4	39 (23-16)		5 (0-5)	79 (72-7)
6	6997/6569	3	115 (95-20)	4	34 (11-23)		15 (6-9)	62 (58-4)
7	6210/5546	3	164 (118-46)	4	62 (24-38)		6 (0-6)	121 (108-13)
8	4477/4118	3	51 (42-9)	4	12 (0-12)		2 (0-2)	41 (39-2)
9	5777/5136	3	27 (6-21)	4	24 (0-24)		7 (0-7)	66 (42-24)
10	6354/5759	3	28 (6-22)	4	27 (9-18)		6 (0-6)	55 (50-5)
11	7778/6979	3	77 (54-23)	4	26 (0-26)		7 (0-7)	73 (45-28)
12	7817/7261	3	35 (6-29)	4	32 (2-30)	4	9 (0-9)	25 (12-13)
13	2536/2362	3	14 (0-14)	4	7 (0-7)		0 (0-0)	3 (0-3)
14	4482/4127	3	37 (29-8)	4	16 (0-16)		8 (7-1)	56 (46-10)
15	4635/4150	3	93 (66-27)	4	40 (18-22)		5 (0-5)	73 (65-8)
16	5596/4744	4	154 (124-30)	5	86 (57-29)		9 (0-9)	112 (78-34)
17	8283/7386	3	91 (58-33)	4	49 (23-26)		10 (0-10)	124 (110-14)
18	2021/1888	3	4 (0-4)	4	5 (0-5)		1 (0-1)	3 (0-3)
19	7438/5982	4	168 (117-51)	5	135 (103-32)		15 (0-15)	197 (131-66)
20	3966/3497	3	11 (7-4)	4	9 (0-9)		1 (0-1)	18 (7-11)
21	1499/1314	4	4 (0-4)	4	29 (26-3)		0 (0-0)	61 (58-3)
22	2957/2493	4	79 (62-17)	5	55 (38-17)		12 (0-12)	124 (107-17)
X	5436/4787	3	36 (15-21)	4	60 (41-19)		9 (0-9)	248[a] (248[a]-0)
Y	281/146	3	0 (0-0)	3	0 (0-0)		0 (0-0)	144[a] (144[a]-0)
Sum	139 795/126 183	-	1806 (1279-527)	-	955 (454-501)	-	189 (32-157)	1667 (1350-317)

Table 1.1: ***CNV call sets information on the 1000 Genomes Project WES data set.*** *Number of exon targets before and after QC procedure is shown. CNVs detected by CODEX, XHMM, CoNIFER, and EXCAVATOR are shown and are further categorized into common and rare ones (common-rare in parentheses). Number of latent factors (K) and principal components (PCs) are shown for latent factor models: default values from CODEX and XHMM are adopted; number of PCs for CoNIFER is chosen at 4 so that it is conservative by the scree plot and is comparable to XHMM.* [a] *Excluded due to mis-handling of sex chromosomes by EXCAVATOR.*

(a)

Software	Sensitivity/specificity/precision/TP/FP/TN/FN compared to HapMap3 callset																			
	All							Common							Rare					
CODEX	0.22	0.95	0.63	377	225	4551	1309	0.22	0.93	0.65	363	198	2644	1284	0.36	0.99	0.34	14	27	1907	25
XHMM	0.05	0.98	0.49	82	84	4638	1658	0.04	0.98	0.61	70	45	2743	1631	0.31	0.98	0.24	12	39	1895	27
CoNIFER	0.01	1.00	0.77	10	3	4712	1737	0.01	1.00	0.77	10	3	2778	1698	0.00	1.00	NaN	0	0	1934	39
EXCAVATOR	0.14	0.92	0.37	222	386	4433	1421	0.14	0.89	0.40	218	322	2563	1386	0.10	0.97	0.06	4	64	1870	35

(b)

Software	Sensitivity/specificity/precision/TP/FP/TN/FN compared to Conrad *et al.* callset																			
	All							Common							Rare					
CODEX	0.08	0.99	0.56	493	383	26892	5378	0.08	0.95	0.62	449	278	4993	5195	0.19	1.00	0.30	44	105	21899	183
XHMM	0.03	0.99	0.52	180	169	27064	5733	0.03	0.99	0.65	144	77	5152	5542	0.16	1.00	0.28	36	92	21912	191
CoNIFER	0.00	1.00	0.16	3	16	27187	5940	0.00	1.00	0.15	2	11	5188	5714	0.00	1.00	0.17	1	5	21999	226
EXCAVATOR	0.07	0.98	0.39	424	668	26603	5451	0.07	0.93	0.53	402	357	4909	5247	0.10	0.99	0.07	22	311	21694	204

(c)

Software	Sensitivity/specificity/precision/TP/FP/TN/FN compared to McCarroll *et al.* callset																			
	All							Common							Rare					
CODEX	0.17	0.97	0.67	480	239	8404	2306	0.16	0.94	0.67	436	215	3419	2274	0.58	1.00	0.65	44	24	4985	32
XHMM	0.07	0.99	0.61	196	123	8489	2621	0.06	0.97	0.63	155	92	3511	2586	0.54	0.99	0.57	41	31	4978	35
CoNIFER	0.01	1.00	0.79	19	5	8580	2825	0.01	1.00	0.83	15	3	3573	2753	0.05	1.00	0.67	4	2	5007	72
EXCAVATOR	0.12	0.94	0.40	329	497	8205	2398	0.12	0.89	0.44	311	394	3299	2340	0.24	0.98	0.15	18	103	4906	58

Table 1.2: ***Sensitivity, specificity, and precision rate of CNV calls by CODEX, XHMM, CoNIFER, and EXCAVATOR.*** *The plot of precision and recall rates are shown in Figure 5. Three "gold-standard" CNV metrics are adopted from (a) International HapMap Consortium, (b) Conrad et al., and (c) McCarroll et al.. CODEX and XHMM performs better in detecting rare CNVs compared to common ones, with CODEX having the highest F-measure among all methods compared.*

(a)

Sample	Pugh *et al.* (tumor/normal)			CODEX			XHMM		
	Start	End	Length (Kb)	Start	End	Length (Kb)	Start	End	Length (Kb)
TARGET.30.PAIFXV				76937010	76952194	15.185			
TARGET.30.PASWLY				76931719	76972722	41.004	76931719	76972722	41.004
TARGET.30.PASRFS	76931795	76972722	40.928	76931719	76972722	41.004			
TARGET.30.PAKZRF	76931795	76972722	40.928	76931719	76972722	41.004			
TARGET.30.PALFPI	76940087	76953125	13.039	76918869	76954119	35.251			
TARGET.30.PALNLU	76940087	76972722	32.636	76937010	76972722	35.713	76931719	76972722	41.004
TARGET.30.PAMVLG	76931795	76972722	40.928	76937010	76972722	35.713	76937010	76972722	35.713
TARGET.30.PANLET	76919049	76953125	34.077	76912048	76972722	60.675	76912048	76972722	60.675
TARGET.30.PANXJL	76940087	76953125	13.039	76937010	76954119	17.11	76937010	76954119	17.11
TARGET.30.PANZVU	76940500	76972722	32.223	76931719	76972722	41.004	76937010	76972722	35.713
TARGET.30.PAPKXS	76931795	76953125	21.331	76931719	76972722	41.004	76931719	76972722	41.004
TARGET.30.PARACS	76778881	76972722	193.842	76778728	76972722	193.995	76778728	76972722	193.995
TARGET.30.PARKNP	76931795	76972722	40.928	76931719	76972722	41.004	76931719	76972722	41.004
TARGET.30.PARMLF	76931795	76972722	40.928	76931719	76972722	41.004	76931719	76972722	41.004
TARGET.30.PASAAN	76931795	76972722	40.928	76931719	76972722	41.004	76931719	76972722	41.004
TARGET.30.PAILNU	76940087	76953125	13.039	76937010	76954119	17.11	76931719	76954119	22.401
TARGET.30.PASTCN	76940087	76972722	32.636	76937010	76972722	35.713	76937010	76972722	35.713
TARGET.30.PATGLU	76764109	76845412	81.304	76763827	76829825	65.999	76763827	76829825	65.999

(b)

Sample	Chromosome	Start	End	Length (Kb)	Num_Probes	Segment_Mean
TARGET.30.PAMVLG	X	76931795	76940500	8.706	3	-2.53655
TARGET.30.PAMVLG	X	76944422	76972722	28.301	5	-26.10105
TARGET.30.PANLET	X	76919049	76953125	34.077	9	-25.48575
TARGET.30.PANXJL	X	76940087	76953125	13.039	6	-1.8623
TARGET.30.PAPKXS	X	76931795	76953125	21.331	7	-23.9693
TARGET.30.PARACS	X	76778881	76814319	35.439	3	-2.7718
TARGET.30.PARACS	X	76829825	76949428	119.604	22	-19.7517
TARGET.30.PARACS	X	76952194	76972722	20.529	3	-2.9612
TARGET.30.PARKNP	X	76931795	76944422	12.628	4	-17.2465
TARGET.30.PARKNP	X	76949428	76972722	23.295	4	-2.9277
TARGET.30.PARMLF	X	76931795	76949428	17.634	5	-13.7373
TARGET.30.PARMLF	X	76952194	76972722	20.529	3	-2.6691
TARGET.30.PASAAN	X	76931795	76972722	40.928	8	-1.06985
TARGET.30.PASRFS	X	76931795	76972722	40.928	8	-1.0058
TARGET.30.PASTCN	X	76940087	76972722	32.636	7	-2.689
TARGET.30.PATGLU	X	76764109	76845412	81.304	9	-1.5369
TARGET-30-PAILNU	X	76940087	76953125	13.039	6	-0.905
TARGET-30-PAKZRF	X	76931795	76972722	40.928	8	-2.6115
TARGET-30-PALFPI	X	76940087	76953125	13.039	6	-3.16695
TARGET-30-PALNLU	X	76940087	76972722	32.636	7	-1.659
TARGET-30-PANZVU	X	76940500	76972722	32.223	6	-2.7205

(c)

sample	chr	start	end	length(Kb)	CNV_type	copy_number_estimate	original cov	normalized cov	mBIC	log-likelihood
TARGET-30-PATGLU	23	76763827	76829825	66	del	0.66	258	776.74	188.13	209.54
TARGET-30-PARACS	23	76778728	76972722	193.99	del	0.26	808	6303.49	14341.1	2591.68
TARGET-30-PANLET	23	76912048	76972722	60.67	del	0.18	322	3607.99	1250.91	1900.32
TARGET-30-PALFPI	23	76918869	76954119	35.25	del	0.3	482	3215.44	1252.09	1273.62
TARGET-30-PARKNP	23	76931719	76972722	41	del	0.44	909	4085.92	1391.41	1412.89
TARGET-30-PASRFS	23	76931719	76972722	41	del	1.07	3368	6280.15	784.08	822.27
TARGET-30-PAPKXS	23	76931719	76972722	41	del	0.19	442	4721.97	1247.82	2377.91
TARGET-30-PAKZRF	23	76931719	76972722	41	del	0.34	574	3426.45	4850.81	1315.36
TARGET-30-PANZVU	23	76931719	76972722	41	del	1.12	2133	3804.46	402.27	434.33
TARGET-30-PARMLF	23	76931719	76972722	41	del	0.49	758	3116.96	1011.6	1033.08
TARGET-30-PASAAN	23	76931719	76972722	41	del	1.13	1455	2568.99	254.49	283.19
TARGET-30-PASWLY	23	76931719	76972722	41	del	0.71	566	1587.82	380.11	401.59
TARGET-30-PAIFXV	23	76937010	76952194	15.18	del	1.21	3304	5440.3	1694.79	446.39
TARGET-30-PANXJL	23	76937010	76954119	17.11	del	0.88	1320	2991.87	559.65	580.98
TARGET-30-PAILNU	23	76937010	76954119	17.11	del	1.09	4512	8305.7	1004.04	1047.76
TARGET-30-PAMVLG	23	76937010	76972722	35.71	del	0.21	418	3993.29	756.75	1776.51
TARGET-30-PALNLU	23	76937010	76972722	35.71	del	0.79	2762	6954.8	1541.52	1562.93
TARGET-30-PASTCN	23	76937010	76972722	35.71	del	0.26	574	4365.7	8141.52	1784.98

(d)

SAMPLE	CNV	INTERVAL	KB	CHR	MID_BP	TARGETS	NUM_TARG	Q_EXACT	Q_SOME	Q_NON_DIPLOID	Q_START	Q_STOP	MEAN_RD	MEAN_ORIG_RD
TARGET-30-PATGLU	DEL	X:76763827-76829825	66	X	76796826	2197..2204	8	69	99	99	34	12	-6.24	21.41
TARGET-30-PARACS	DEL	X:76778728-76972722	194	X	76875725	2201..2230	30	26	99	99	20	23	-3.04	10.77
TARGET-30-PANLET	DEL	X:76912048-76972722	60.67	X	76942385	2219..2230	12	11	99	99	11	22	-5.7	6.4
TARGET-30-PAILNU	DEL	X:76931719-76954119	22.4	X	76942919	2222..2229	8	7	99	99	7	8	-4.75	123.68
TARGET-30-PARKNP	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	36	99	99	16	33	-6.74	17.76
TARGET-30-PAPKXS	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	35	99	99	22	19	-7.21	9.64
TARGET-30-PASWLY	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	3	63	63	4	20	-2.78	13.47
TARGET-30-PARMLF	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	25	99	99	24	25	-5.86	16.15
TARGET-30-PASAAN	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	6	99	99	6	34	-3.45	27.47
TARGET-30-PALNLU	DEL	X:76931719-76972722	41	X	76952220	2222..2230	9	15	99	99	8	26	-4.9	62.63
TARGET-30-PANXJL	DEL	X:76937010-76954119	17.11	X	76945564	2223..2229	7	23	99	99	6	3	-5	23.68
TARGET-30-PASTCN	DEL	X:76937010-76972722	35.71	X	76954866	2223..2230	8	27	99	99	21	25	-3.3	11.88
TARGET-30-PAMVLG	DEL	X:76937010-76972722	35.71	X	76954866	2223..2230	8	96	99	99	23	32	-6.65	8.39
TARGET-30-PANZVU	DEL	X:76937010-76972722	35.71	X	76954866	2223..2230	8	13	99	99	4	31	-3.67	24.23

Table 1.3: ***Somatic deletions within*** **ATRX** ***region detected using WES data of neuroblastoma patients.*** *(a) Summary of deletions detected by tumor/normal threshold, CODEX, and XHMM with break-point and length information. Of the 18 samples detected by CODEX, 16 samples overlap with the matched tumor blood analysis result; 14 and all of XHMM's CNV events are detected; one sample is uniquely called. Breakpoints may differ slightly between different methods but are within reasonable limits. (b) Deletions detected by thresholdhing log_2-ratio of tumor RPKM to blood RPKM. (c) Deletions detected using tumor samples only by CODEX. (d) Deletions detected using tumor samples only by XHMM.*

Chr	Number of Targets	Optimal K	Blood CNVs (common-rare)	Tumor CNVs (common-rare)
1	19157	4	916 (448-468)	1303 (717-586)
2	14033	4	386 (115-271)	866 (360-506)
3	11257	4	213 (0-213)	321 (120-201)
4	7365	3	232 (98-134)	552 (295-257)
5	8472	3	384 (180-204)	755 (400-355)
6	9346	4	421 (234-187)	431 (239-192)
7	8694	4	625 (436-189)	775 (513-262)
8	6353	3	267 (106-161)	418 (179-239)
9	7523	3	185 (55-130)	307 (163-144)
10	7730	3	241 (80-161)	329 (182-147)
11	10564	4	648 (147-501)	1562 (822-740)
12	10647	4	308 (34-274)	464 (113-351)
13	3332	3	101 (13-88)	116 (20-96)
14	5676	4	159 (41-118)	173 (51-122)
15	6479	3	158 (45-113)	361 (137-224)
16	7785	4	408 (221-187)	488 (251-237)
17	11208	4	304 (138-166)	761 (500-261)
18	2807	3	45 (0-45)	84 (2-82)
19	10780	5	558 (316-242)	833 (377-456)
20	4650	3	97 (30-67)	108 (26-82)
21	1889	3	72 (24-48)	77 (39-38)
22	3929	3	116 (51-65)	146 (59-87)
X	6445	3	157 (45-112)	214 (74-140)
Y	306	2	0 (0-0)	0 (0-0)
All	186427	-	7001 (2857-4144)	11444 (5639-5805)

Table 1.4: ***Genome-wide CNVs detected by CODEX of the neuroblastoma data set.*** *Blood and tumor CNVs are reported separately by chromosome.*

Chapter 2

ASSESSING INTRA-TUMOR HETEROGENEITY AND TRACKING LONGITUDINAL AND SPATIAL CLONAL EVOLUTIONARY HISTORY BY NEXT-GENERATION SEQUENCING

2.1 Introduction

It has been long recognized that cancer is a disease driven by genetic and epigenetic alterations (48-50). These alterations confer upon its carrier cell selective advantage, and rounds of Darwinian selection produce tumor cell populations with aggressive phenotypes. High-throughput sequencing technologies have made possible the large-scale, high-resolution analysis of tumor genomes. A recurring finding of these studies is the high degree of heterogeneity – both inter-tumor heterogeneity among patients with the same clinical diagnosis (51, 52), as well as intra-tumor heterogeneity between tumor cells derived from the same patient (summarized in Table 2.1) (53-59). Heterogeneity, at all levels, confound diagnosis and treatment. Most large-scale studies to date, for example those led by the Cancer Genome Atlas Research Network (51) and the International Cancer Genome Consortium (52) have focused on inter-tumor heterogeneity. These studies typically collect and sequence bulk tissue data, usually one sample per patient, and compare the mutation profiles across patients. This study design is not optimized for the study of intra-tumor heterogeneity, which has thus received, until recently, comparatively less attention.

When *only one* sample from a tumor is sequenced, early analyses of intra-tumor heterogeneity started with the estimation of normal cell contamination and tumor ploidy (60, 61). For example, ABSOLUTE, one of the earliest methods, classifies mutations as clonal or subclonal after adjusting for the estimated purity and ploidy of the sample. Most approaches for the detection of subclonal mutations treat point mutations and copy number aberrations separately (62-65). In the case of point mutations, i.e. single-nucleotide alterations (SNAs) and small insertions and deletions (indels), most methods rely on mixture models for the variant allele frequency (VAF) under the assumption that mutations carried by the same set of cells have the same VAF. But the VAF is also affected by the copy number of the region where the point

mutation resides, and copy number aberrations (CNAs) are prevalent in cancer. Recently, Li and Li (66) and Deshwar *et al.* (67) proposed models for joint inference of SNAs and CNAs. Li and Li (66) further gave important insight into the identifiability of the underlying parameters, if one were to analyze each mutation locus separately. The many unknowns, including the number of subpopulations in the tumor, the mutation profile of each subpopulation and its contributing proportion to the sample, and the phasing of aberrations that affect the same genome locus make the estimation problem challenging in all but the simplest scenarios, *if* one were to sequence only one bulk DNA sample from the tumor. We will discuss these underlying challenges through a more thorough literature review after giving a more detailed formulation of the problem.

Ultimately, tumor evolution occurs at the single-cell level, and single-cell methods provide a powerful approach to assess tumor heterogeneity without the confounding effects of mixed cell populations (53, 68). Despite its promise, single-cell DNA sequencing data are much noisier than bulk sequencing data due to allele dropout events and amplification errors (69), and furthermore, the per-cell coverage is still limited due to constraints on budget and labor. While these single-cell sequencing studies have improved our understanding of intra-tumor heterogeneity, most current tumor studies still sequence the DNA at the bulk tissue level.

Recently, there have been increasing efforts to sequence the tumor from the same patient at multiple time-points and/or from multiple spatially separated resections (54-59). Multiple snapshots of the same tumor have proved invaluable for identifying subclonal populations and for inferring the tumor's evolutionary history. Multi-dimensional scatterplots of VAFs allow higher resolution for cluster detection than the one dimensional histogram in the single-sample case. Recent methods, such as Pyclone (62) and SciClone (63), apply Bayesian mixture models to detect these clusters. LICHeE (70) and SCHISM (71) infer phylogeny from VAFs as an acyclic directed graph network. Another recent work, Clomial (72), showed that it is possible to obtain precise and informative estimates of the underlying subpopulations through a matrix deconvolution framework. One practical drawback of Clomial (72) is that it takes only SNA input

and assumes that all mutational loci are heterozygous from copy number neutral regions. SCHISM (71), BitPhylogeny (73), PhyloWGS (67), and SPRUCE (74) adjust for CNAs in their model in different ways, but these methods still require limiting assumptions and do not make full use of the data, as we discuss in detail a bit later.

Here, we focus on the analysis of intra-tumor heterogeneity by multi-sample bulk DNA sequencing of tumor samples. We propose Canopy (***c***opy number ***a***nd single ***n***ucleotide alteration analysis ***o***f tumor ***p***h***y***logeny) (75), a statistical framework and computational procedure for identifying the subpopulations within a tumor, determining the mutation profiles of these subpopulations, and inferring the tumor's phylogenetic history. The input to Canopy are VAFs of somatic SNAs along with allele-specific coverage ratios between the tumor and matched normal sample for somatic copy number calls. These quantities can be directly taken from the output of existing software. Canopy provides a general mathematical framework for pooling data across samples and sites to infer the underlying phylogeny. For SNAs that fall within CNA regions, Canopy infers their temporal ordering and resolves their phase. When there are multiple evolutionary configurations consistent with the data, Canopy attempts to explore all configurations and assess their confidence.

Identifiability of the underlying evolutionary process and confidence in its reconstruction is an important aspect of consideration. The Bayesian framework for Canopy allows assessment of the quality of inference. The resolution at which clones can be differentiated depends on the data, and in particular, on how many slices of the tumor are taken, how genotypically different these slices are to each other, and sequencing depth. As the number of clones increase, the proportion of cells attributable to at least some of the subclones would necessarily decrease, and thus, the higher sequencing depth would be needed to detect mutations present in those clones. Under the Bayesian framework, the resolution of our estimates and the confidence in our conclusions can be quantified by the posterior distribution.

2.2 Results

We will start by giving a more precise formulation of the clonal decomposition problem along with a more in-depth discussion of existing methods and their key assumptions. We will show that, under our formulation, the likelihood of the observed sequencing data can be written in matrix form and be decomposed into terms that reflect the tumor's phylogenetic history, the phasing of overlapping SNAs and CNAs, and the contributing proportions of the admixed cell populations. Canopy assumes non-informative priors for the unknowns in the model, and explores their possible values by Markov chain Monte Carlo (MCMC). Through simulations, we explore the effects of various parameters on deconvolution accuracy, and compare Canopy against existing methods. Canopy is then applied to four datasets with different sequencing designs: the whole-exome sequencing of a heterogeneous triple-negative breast carcinoma cell line MDA-MB-231 and its derived sublines with single and mixed cell populations, the whole-genome sequencing of breast cancer patient xenografts from Eirew *et al.* (57), the whole-genome sequencing of a leukemia patient at two time-points from Ding *et al.* (54), and the multi-region sequencing of an ovarian cancer patient from Bashashati *et al.* (55).

2.2.1 Modeling of SNAs, CNAs, and Clonal Tree

Figure 2.1a shows the phylogeny of an evolving tumor, which starts from a diploid normal cell and progresses through waves of somatic mutations. The tumor's evolution is depicted as a bifurcating tree, with the ancestral normal cell population at the root, and accumulating mutations along its branches. Time runs vertically down the tree from the root, and when a sample of the tumor is taken at any point in time, the tree is sliced horizontally, cutting the branches to form leaves. The subpopulations within the sample are represented by the "leaves" in that slice. Each subpopulation contributes a fraction of cells to the sample, which, taken together, are represented by a vector of non-negative numbers that sum to one. To model normal cell contamination we restrict the left-most branch of the tree to be non-bifurcating and mutation-free. Thus, the proportion of normal cells within any sample is simply the first entry in its mixture proportion

vector. Multiple samples collected for the same tumor are represented by multiple horizontal slices of the phylogeny, each receiving its own vector of proportions.

The observed data is summarized in Figure 2.1b. We let N be the number of samples, and S and T be the number of somatic SNAs and CNAs, respectively, that were called across all samples. For SNAs, let the matrices $R \in \mathbb{R}^{S\times N}$ and $X \in \mathbb{R}^{S\times N}$ be, respectively, the number of reads containing the mutant allele and the total number of reads covering each of the S loci in each of the N samples. The ratio R/X is the proportion of reads supporting the mutant allele, known as the variant *allele* frequency (VAF). For CNAs, Canopy directly takes output from FALCON (76), FALCON-X, or other allele-specific copy number estimation methods (77). These outputs are in the form of estimated major and minor *copy number ratios,* denoted by $W^M \in \mathbb{R}^{T\times N}$ and $W^m \in \mathbb{R}^{T\times N}$ respectively, with their corresponding standard errors $\varepsilon^M \in \mathbb{R}^{T\times N}$ and $\varepsilon^m \in \mathbb{R}^{T\times N}$. See 2.4.1 Allele-Specific Copy Number for details regarding these quantities. For each SNA and each CNA, we also know whether they overlap. This information is represented by the matrix $Y \in \mathbb{R}^{S\times(T+1)}$: for column $j+1$, Y has 1's for SNAs that lie within CNA j and 0's for all other SNAs; as first column, Y has 1's for SNAs that don't reside in any CNAs and 0's otherwise (see example in Figure 2.1b).

Each sample contains a mixture of the clones that comprise the tumor, and thus these *observed* VAFs and copy number ratios rely on the mixture proportions as well as the genomic profiles of the clones, as embodied by the underlying phylogenetic tree that is shared across all samples collected for the same tumor.

2.2.2 Relationship to Existing Work

Many existing studies of tumor evolution by multi-region or multi-time-point bulk tumor DNA sequencing rely on laborious manual history reconstruction (54, 55). There have been much recent progress in the development of computational approaches for the analysis of such data. These approaches differ in the types of mutations that are modeled and the assumptions that are made. The main differences are summarized in Table 2.2 and discussed below.

TITAN (64) and THetA (65) focus on estimating cell population structure and recovering clonal evolutionary history for the case where somatic CNAs and loss of heterozygosity (LOH) distinguish subpopulations. These methods use allelic read coverage at germline heterozygous SNP loci to distinguish clonal versus subclonal CNA events. They ignore SNAs, and do not pool data across multiple samples from the same tumor.

Many programs focus specifically on SNAs. For example, SciClone (63) clusters the VAFs of SNAs in copy-number neutral and LOH-free portions of the genome using a Bayesian beta mixture model. Pyclone (62) is an extension of SciClone that adds prior information elicited from copy number estimates obtained from either genotyping arrays or whole-genome sequencing to its Bayesian nonparametric clustering method. Neither SciClone (63) nor Pyclone (62) infers the phylogenetic relationship between subclones. LICHeE (70) and SCHISM (71) take VAFs of SNAs as input and construct a phylogenetic tree via an acyclic directed graph. Clomial (72), another program designed exclusively for SNAs, performs mixture deconvolution assuming that all mutational loci are heterozygous from copy number neutral regions. Clomial decomposes the VAF matrix into a product of sample proportions and population genotypes, and uses expectation maximization (EM) to estimate both matrices.

ABSOLUTE (60) was the first software to infer subclonal heterozygosity from both SNAs and CNAs. However, taking data from only one sample, it determines whether each event is clonal or subclonal, but does not attempt to genotype or quantify the underlying subclones. In a similar fashion, Lonnstedt *et al.* (78) took a two-step approach using both SNA and CNA input, first estimating CNAs and then comparing VAF of SNA to its local copy number estimate to classify the somatic point mutation as clonal or subclonal. Recent approaches such as BitPhylogeny (73) and PhyloSub (79) detect major subclonal lineages by sampling the subclonal proportions via a tree-structured stick-breaking (TSSB) process, adjusting for overlapping CNAs. BitPhylogeny further adapts the nonparametric Bayesian mixture model to DNA methylation data from multiple microdissections from different regions of the same tumor.

As mentioned earlier, the VAF, which quantifies the proportion of *alleles* in the sample carrying a somatic mutation in the sample's DNA pool, is not the same as the proportion of *cells* in the sample carrying the somatic mutation. We call the latter, which is not directly observed in sequencing data, the mutant cell frequency (MCF). A similar quantity that is sometimes used in literature is cancer cell fraction (CCF), which is the proportion of cells among all cancer cells carrying the mutation. Given the tumor purity ϕ_C, $MCF = CCF \times \phi_C$. The MCF of a mutation directly reflects the total contributing proportion of the clone(s) that carry it, but to compute MCF from VAF, one needs to compensate for any CNAs that affect the locus. The existing methods differ by how this compensation is done. ABSOLUTE (60), EXPANDS (61), Pyclone (62) and PhyloSub (79) assume that when a CNA event overlaps a SNA, the point mutation resides in a region with homogeneous aneuploidy, a scenario where no subclonal CNA events are allowed. Li and Li (66) conducted a detailed analysis of the complete set of scenarios covering the possible order and phase of overlapping SNAs and CNAs in developing their software CHAT. However, CHAT does not pool information across sites or across samples. PhyloWGS (67) also conducts a detailed breakdown of the possible configurations of overlapping SNA and CNA and is the first method to integrate both types of mutations when reconstructing cancer phylogenies using a TSSB. However, for each CNA region, PhyloWGS requires as input the integer absolute copy number of each allele and treats CNA events as pseudo-SNA events to compute its MCF. Since knowing integer-valued copy number is akin to knowing the clonal decomposition, in essence, PhyloWGS requires a two-step procedure where the underlying clones are first identified with their absolute copy numbers estimated using CNA data only, and this information is then used to compute the MCF of SNAs. SPRUCE (74) is another recent method that analyzes both SNAs with CNAs and characterizes the tumor phylogeny as a restricted class of spanning trees. Like PhyloWGS, SPRUCE takes processed CNA calls, e.g. from THetA, and assumes known MCF for CNA events. Unlike these existing approaches, Canopy takes as input raw copy number ratios estimated by existing segmentation programs, and uses SNAs and CNAs to jointly infer the

underlying clones and their evolutionary history. Since the same clonal admixture underlie CNAs and SNAs, this integrated approach achieves more accurate estimates in complex scenarios, as we illustrate later through examples.

As with all phylogenetic inference, assumptions are needed to resolve ambiguity. The perfect phylogeny model (70, 80) assumes that all subclones share the same phylogenetic tree and that mutations don't recur independently in different subclones, that is, each mutation appears only once and once it appears, it does not revert back to its original state. This no-homoplasy assumption, also referred to as the infinite sites assumption (67, 81), is adopted by most methods to allow model identifiability. For example, it is possible to assert that under the infinite sites assumption, mutations with lower CCFs cannot be ancestral to mutations with higher CCFs. To deal with copy number changes, El-Kebir *et al.* (74) proposed instead an infinite alleles assumption, or the multi-state perfect phylogeny, where a mutation may change state more than once on the tree due to gain or loss of copy number, but changes to the same state at most once. Furthermore, Deshwar *et al.* (67) introduces the 'weak parsimony' assumption, which posits that mutations with similar CCFs across all samples lie on the same branch segment in the phylogeny. Canopy relies on both the infinite sites assumption and the weak parsimony assumption, but takes a different approach from El-Kebir *et al.* (74) in modeling CNAs: Canopy extends the infinite sites assumption to CNAs by assuming that copy number events with the exact same breakpoints and resulting in the same copy number must be the same mutation event that occurs exactly once in the tumor's evolution. CNAs that overlap but have different breakpoints or different copy number states are treated as separate events. For example, a homozygous deletion nested within a heterozygous deletion, or a series of nested amplifications, are treated as separate events rather than separate alleles of the same mutation. This assumption allows Canopy to, with appropriate data, resolve the evolutionary relationship between overlapping copy number events, as we show in the whole-exome study of breast cancer cell line MDA-MB-231.

2.2.3 Matrix Representation of a Tumor's Clonal Composition

We use K to denote the total number of clones of the tumor that have representation among the cells in our sample(s). As shown in Figure 2.1a, the tumor's evolutionary history is denoted by τ_K, a bifurcating tree with K leaves and with point mutation and copy number events assigned to its branch segments. Any τ_K gives us three matrices reflecting the mutation profiles of the underlying clones, shown in Figure 2.1c: The SNA genotypes $Z \in \mathbb{R}^{S\times K}$, where Z_{sk} is the indicator of whether the sth SNA is present at the kth clone, and the major and minor copy numbers $\tilde{C}^M \in \mathbb{R}^{T\times K}$ and $\tilde{C}^m \in \mathbb{R}^{T\times K}$, where $\tilde{C}^M_{tk}$ and $\tilde{C}^m_{tk}$ are *integer-valued* major and minor copy numbers of the tth CNA in the kth clone. Phylogenetic restrictions are imposed by Canopy in that there is a one-to-one mapping between the positions of SNAs and CNAs on the tree as well as the major and minor copies of CNA events and the matrix Z, $\tilde{C}^M$, and $\tilde{C}^m$. Furthermore, since the left most clone in the tree represents the normal cells, the first column of Z contains all zeros and the first columns of $\tilde{C}^M$ and $\tilde{C}^m$ contain all 1's. We do not directly observe the clones; instead, the samples we sequence are mixtures. We define $P \in \mathbb{R}^{K\times N}$ as the clonal frequency matrix, where P_{kj} $(1 \le k \le K, 1 \le j \le N)$ is the fraction of cells in the jth sample that belong to the kth clone (P shown in Figure 2.1a is transposed to be aligned to the phylogeny). Each column of P sums up to one with the first row corresponding to the normal cell contamination. The matrices $Z, \tilde{C}^M, \tilde{C}^m$ and P are all unobserved, as well as the number of clones K. Our goal is to estimate them from the observed data, i.e. the VAFs and the major and minor copy number ratios.

2.2.4 SNA-CNA Phase and Combined Likelihood

Here, we derive the likelihood for the data, given the model parameters $\{Z, \tilde{C}^M, \tilde{C}^m, P, K\}$. First, consider the CNA events. For CNAs, multiplication (denoted by $\times$) of the clonal integer copy number matrices $(\tilde{C}^M, \tilde{C}^m)$ and the sample proportion matrix (P) gives us the continuous-valued major and minor copy numbers for each sample:

$$\tilde{C}^M \times P = C^M \in \mathbb{R}^{T\times N}$$
$$\tilde{C}^m \times P = C^m \in \mathbb{R}^{T\times N}.$$

Since the observed copy number ratios are usually computed by averaging over a large number of loci (for microarrays), exons (for WES), or bins (for WGS), we assume that they are normally distributed with the given standard errors, that is,

$$W^M \sim N(C^M, (\varepsilon^M)^2)$$
$$W^m \sim N(C^m, (\varepsilon^m)^2).$$

For SNAs, $Z \times P$ gives the mutant *cell* frequency (MCF) of each SNA in each sample, which we denote by the matrix $MCF \in \mathbb{R}^{S \times N}$. The observed number of mutant reads R_{sj} follows a binomial distribution with total count X_{sj} and probability of success being the variant *allele* frequency (VAF), which we denote by VAF_{sj} $(1 \leq s \leq S, 1 \leq j \leq N)$. That is,

$$R_{sj} \sim \text{Binomial}(X_{sj}, VAF_{sj}).$$

Therefore, we need to convert MCF to VAF in order to calculate the binomial likelihood for SNAs. If all SNAs are heterozygous from copy number neutral regions, as assumed by SciClone (63) and Clomial (72), then $VAF = 1/2 \times MCF = 1/2 \times CCF \times \phi_C$, where ϕ_C is the cancer cell purity, MCF is the fraction of cells that have the SNA, and CCF is the fraction of cancer cell that have the mutation. Pyclone (62), PhyloSub (79) and EXPANDS (61) account for CNAs but make the assumption that was first introduced by ABSOLUTE (60), namely that there are no subclonal CNA events. Therefore,

$$VAF = \frac{C_{mut}}{2\times(1-\phi_C)+C_{total}\times\phi_C} MCF = \frac{C_{mut}\times\phi_C}{2\times(1-\phi_C)+C_{total}\times\phi_C} CCF,$$

where ϕ_C is the purity of cancer cells, which have a homogeneous CNA state with total copy number C_{total} and mutant-allele copy number C_{mut}.

To more accurately quantify the relationship between VAF and MCF, which accounts for the possible phases and temporal orders of overlapping CNAs and SNAs, we consider separately each of the three possible underlying scenarios, which were first delineated by CHAT (66) and PhyloWGS (67): (i) the CNA is ancestral to the SNA (Figure 2.2a); (ii) the CNA and SNA occur in separate branches of the tree and thus affect separate clones (Figure 2.2b); (iii) the SNA is ancestral to the CNA (Figure 2.2c). To compute the VAF, we separately calculate the *numerator*

(copy number of the affected allele at the mutational locus), and the *denominator* (total copy number at the locus) for each SNA across all samples. The *denominator* can be generalized and is the same for all three cases, being simply the total copy number at the SNA locus. Therefore, the *denominator* is, in matrix notation,

$$\left(Y \times \begin{bmatrix} \mathbb{1} \\ \cdots \\ \tilde{C}^M \end{bmatrix} + Y \times \begin{bmatrix} \mathbb{1} \\ \cdots \\ \tilde{C}^m \end{bmatrix}\right) \times P \in \mathbb{R}^{S \times N} = \begin{matrix} \\ \text{SNA1} \\ \text{SNA2} \\ \text{SNA3} \\ \text{SNA4} \end{matrix} \begin{matrix} \begin{bmatrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \\ 2 & 2 & 2 \\ 1.8 & 1.8 & 1.7 \\ 2.6 & 2.5 & 2.5 \\ 2 & 2 & 2 \end{bmatrix} \end{matrix},$$

where $\mathbb{1}$ is a vector of ones augmented to the first row of $\tilde{C}^M$ and $\tilde{C}^m$ representing the major and minor copy number for the 'non-CNA' SNA loci. The *numerator* differs across the three cases.

Case 1: the CNA is ancestral to the SNA

Only one allele of the locus is affected (e.g., SNA1 in Figure 2.2a). Therefore, the copy number of the affected allele for SNA s in each *clone* is $Z_{s,:} \in \mathbb{R}^{1 \times K}$ (the s row of the Z matrix). The *numerator*, which is the copy number of the affected allele in each sample, is thus the matrix product of $Z_{s,:}$ and P. For SNA1 in Figure 2.2a, this evaluates to

$$\left(Z_{1,:} \times P\right) = \begin{matrix} \\ \text{SNA1} \end{matrix} \begin{matrix} \begin{bmatrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \\ 0.3 & 0.5 & 0 \end{bmatrix} \end{matrix}.$$

Note that the numerator in this case is the same as the row corresponding to the SNA in the MCF matrix ($MCF = Z \times P$) since each variant cell has only one variant allele.

Case 2: the CNA and SNA occur in two non-overlapping lineages

The SNA isn't affected by the CNA, which lies on a different branch of the tree (e.g., SNA2 in Figure 2.2b). Therefore, the *numerator* is the same as is in the previous case. For SNA2 in Figure 2.2b, this evaluates to

$$\left(Z_{2,:} \times P\right) = \begin{matrix} \\ \text{SNA2} \end{matrix} \begin{matrix} \begin{bmatrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \\ 0.3 & 0 & 0.5 \end{bmatrix} \end{matrix}.$$

Case 3: the SNA is ancestral to the CNA

This is usually the most interesting case. For example, copy number loss or LOH following an SNA may delete the normal allele, or copy number gain may amplify the mutated allele (e.g.,

SNA3 in Figure 2.2c) – both scenarios having potential phenotypic consequences. Of the two underlying alleles at the SNA locus, we need to distinguish which has been lost and/or gained. That is, if the CNA confers allelic imbalance, we need to distinguish whether the major or the minor allele is the mutated allele. For SNA3 in Figure 2.2c, the major allele has the SNA and the copy number of the mutant allele in each sample, aka, the *numerator*, is

$$\left\{Z_{3,:} \cdot \left(Y_{3,:} \times \begin{bmatrix} \mathbb{1} \\ \dots \\ \tilde{C}^M \end{bmatrix}\right)\right\} \times P \in \mathbb{R}^{1\times N} = \begin{matrix} \text{SNA3} & [\end{matrix} \begin{matrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \\ 1.4 & 1.2 & 1.3 \end{matrix}].$$

Here we define the notation $\cdot$ as element-wise matrix multiplication. If the mutant allele lands on the minor copy, the numerator is simply the above with $\tilde{C}^M$ replaced by $\tilde{C}^m$.

A general formula for the *numerator* encompassing all three cases is derived in 2.4.2 Generalization of VAF and MCF Relationship for All Three Cases. Division of the *numerator* by the *denominator* gives us the *VAF*, and the likelihood can then be expressed as

$$\begin{aligned} &L\left(Z, P, \tilde{C}^M, \tilde{C}^m, \tilde{H}, \tilde{Q}, \tau_K \mid W^M, W^m, \varepsilon^M, \varepsilon^m, R, X, Y\right) \\ &\quad = \prod_{j=1}^{N} \prod_{s=1}^{S} \prod_{t=1}^{T} \Big\{ \text{pNorm}\left(W_{tj}^M, \left(\tilde{C}^M \times P\right)_{tj}, \left(\varepsilon_{tj}^M\right)^2\right) \\ &\qquad \text{pNorm}\left(W_{tj}^m, \left(\tilde{C}^m \times P\right)_{tj}, \left(\varepsilon_{tj}^m\right)^2\right) \text{pBinomial}\left(VAF_{sj}, R_{sj}, X_{sj}\right)\Big\}, \end{aligned}$$

where $\text{pNorm}(x, \mu, \sigma^2)$ is the likelihood for observing x from a Gaussian distribution with mean μ and standard deviation σ, $\text{pBinomial}(p, R, X)$ is the Binomial likelihood for observing R successes from X trials with success probability p, $\tilde{H}$ indicates the phasing of the SNAs with overlapping CNAs (whether an SNA precedes a CNA), $\tilde{Q}$ is a vector of the ordering of the SNA-CNA pair that can be directly obtained from the tree, VAF is derived in 2.4.2 Generalization of VAF and MCF Relationship for All Three Cases.

For cases where nested CNAs are observed, Canopy samples the temporal and spatial orders of the CNAs together with the affected SNAs in the phylogenic tree. Resolving overlapping and nested CNA events is not trivial, since in real datasets analysis we only see the major and minor allelic ratio per region per sample. By overlapping CNAs we are referring to distinct CNA

events occurring in separate samples that affect the same genomic region, more specifically, overlapping across samples (e.g., CNA event E_1 and E_2 in Figure 2.3b); by nested CNAs, on the other hand, we are referring to CNAs that may occur in different samples or within the same sample (e.g., CNA event E_2 and E_3 in Figure 2.3b). Canopy can resolve CNA events, overlapping or nested, that have representation in the data. For such events, manual inspection of the segmentation input is sometimes helpful to identify nested CNAs within the same sample and to verify the type (gain, loss, or copy neutral LOH) of each event.

We use BIC as a model selection method to determine the number of subclones K and design a Metropolis Hastings algorithm to sample the posterior distribution of the unknowns and enumerate all plausible histories in the tree space:

(i) randomly switch a CNA or SNA to another branch on the tree;

(ii) randomly select at least two clones and change their clonal frequencies;

(iii) randomly select a neighborhood for local rearrangement to generate a new tree topology (Figure 2.4);

(iv) randomly select a CNA and sample its major and minor copy number from {0,1,2,3} and update $\tilde{C}_{tk}^{M}$ and $\tilde{C}_{tk}^{m}$, $1 \leq t \leq T, 1 \leq k \leq K$;

(v) for SNA that resides in a CNA ($Q_s = 1$), randomly sample whether the major or the minor allele contains the SNA after the copy number change, aka, randomly sample the indicator random variable H_s $(1 \leq s \leq S)$.

For each run, we start with multiple chains from different start points and evaluate convergence by likelihood and acceptance rate. Posterior distribution is marginalized after combining different chains, burn-in, and thinning. When multiple posterior 'modes' exist, Canopy attempts to return all phylogenies that the data support and computes the relative confidence interval in each clonal history. Quantities that can be marginalized from the posterior distributions are obtained from subtree space with trees having the same clonal and mutational compositions.

2.2.5 Simulation Studies

As a simple illustration, we first show how Canopy successfully identifies the subclones and recovers the phylogeny for the scenario shown in Figure 2.1 and Figure 2.5, which is a simple configuration that is as typical as any other given the level of complexity. We also use this example to demonstrate the differences between Canopy and two related methods, PhyloWGS and Clomial. To generate suitable input for PhyloWGS, we converted the CNA events to pseudo-SNA events, since in this toy example we have at our disposal the true clonal proportions as well as the true SNA-CNA phasing, and thus simply used these true values as if it were known. Refer to 2.4.3 Simulation Setup for details. Canopy, starting with raw CNA estimates and assuming unknown phase, returns a tree highly concordant with the ground truth; whereas PhyloWGS, even using the true phase and clonal proportions for CNAs, returns a linear tree with incorrectly inferred cellular frequencies (Figure 2.5c). We further introduce scenarios where CNAs overlap and show that Canopy can successfully handle a fair amount of complexity. As a comparison, Clomial (72), which ignores the existence of CNAs, fails to correctly estimate the clonal frequencies and infers incorrect tumor purities (Figure 2.5b). Figure 2.5d also explores the effect of CNA estimation noise on deconvolution accuracy.

We then performed simulation studies to explore the effects of various parameters on estimation accuracy as well as computation time, and evaluate performance benchmarked against existing methods. We use the percentage of wrongly labeled Z elements (Figure 2.6) and the root-mean-square error (RMSE) of the P matrix (Figure 2.7) as a measure of the deconvolution accuracy and compare Canopy's results with those returned by Clomial (72). We use *clustering purity* as a measure of clustering quality and compare the pre-clustering results of Canopy with those of SciClone (63). We sampled systematically from a comprehensive set of possible phylogenies and P matrices. More details on simulation setup are in 2.4.3 Simulation Setup.

We ran simulations with varying number of mutations from two different sequencing pipelines: whole-genome sequencing with $d = 30$ (Figure 2.7a) and targeted sequencing with $d = 500$ (Figure 2.7b), where d is the mean sequencing depth. The results give a sense of how estimation accuracy depends on the number of *informative* mutations, the number of *genotypically distinct* samples, the sequencing depth, and the number of underlying clones. As expected, estimation accuracy increases with the number of genotypically distinct samples, the number of informative mutations, and the sequencing depth. Increasing the number of subclones makes the estimation problem harder, although this can be compensated for by a larger number of mutations. Also, in Figure 2.8, we show that the larger the difference in clonal proportions between the samples, the easier the estimation problem. Under all simulated scenarios, Canopy is as good as or better than Clomial (72) and SciClone (63) in terms of deconvolution and clustering accuracy (Figure 2.6, Figure 2.7). An interesting observation from the simulation studies is that while increasing the number of samples drives the estimation error of Z to zero, the benefit of including more mutations diminishes when there is a small number of underlying subclones. In this case, only a small high confidence set of informative mutations or mutation clusters is sufficient for recovering the underlying tree (Table 2.3); when the number of underlying subclones is large, more mutations are needed (Table 2.3).

We also performed simulations to investigate the effect of the proposed binomial mixture clustering method with varying number of mutations and clones. This pre-clustering procedure serves as an initialization step in the MCMC sampling, where mutations are first moved along tree branches in clusters and then fine-tuned individually in the later rounds. We show that this initialization method significantly reduces computation time, offers a way to clean up data by including a uniform noise component, and has similar or better deconvolution accuracy (Table 2.3, Figure 2.9).

2.2.6 Application to Transplantable Metastasis Model Derived from MDA-MB-231

Canopy is applied to a transplantable metastasis model system derived from a heterogeneous human breast cancer cell line MDA-MB-231. Cancer cells from the parental line MDA-MB-231 were engrafted into mouse hosts leading to organ-specific metastasis. Single cell populations (SCPs) or mixed cell populations (MCPs) were *in vivo* selected from either bone or lung metastasis and grew into phenotypically stable and metastatically competent cancer cell lines (Figure 2.10a, Table 2.4).

This transplantable model system has been widely used for understanding metastatic progression (82-84). Minn *et al.* (82) identified a 'poor-prognosis' gene expression signature for distinct metastatic potential by studying patterns of transcriptomic profile. Recently, Jacob *et al*. (83) performed whole-exome sequencing on a metastasis model derived from the same parental line MDA-MB-231 and found that *in vivo* selected highly metastatic cell populations showed little genetic divergence from the corresponding parental population. Their results suggest that: (i) genetic variations (including oncogenic mutations in $BRAF^{G464V}$ and $KRAS^{G13D}$, validated by Sanger sequencing) preexist in the parental line and are enriched with increased metastatic capability; (ii) metastatic competence during tumorigenesis can emerge with selection of preexisting oncogenic alleles without a need of new mutations (83).

Here we build a transplantable model from MDA-MB-231, where the parental line as well as the SCP and MCP samples are whole-exome sequenced and are used to investigate clonal evolution associated with metastatic progression on the DNA level. We only use the parental line and the MCP samples to infer metastatic phylogeny, while the SCP samples are included as a validation dataset to compare and contrast. Since SCP samples are homogeneous cell populations, their integer absolute copy numbers can be inferred by a hidden Markov model (HMM). Since we do not have a normal control for MDA-MB-231, the integer absolute copy numbers for the SCP samples are used as controls to infer copy number ratios in the MCP

samples (Figure 2.11). SNAs and indels are called by the UnifiedGenotyper in the Genome Analysis Toolkit (GATK) (85) and are further annotated by ANNOVAR (86).

In addition to the oncogenic point mutations in *BRAF* and *KRAS* reported by Jacob *et al*. (83), our analysis pipeline identified two nonsynonymous mutations in *ALPK2* and *RYR1* that are deleterious by functional annotation. VAFs of the four mutations vary between bone and lung metastasis samples: *BRAF* and *KRAS* mutations are enriched in the bone samples; *ALPK2* mutation is enriched in the lung samples; *RYR1* mutation is additionally acquired by the bone samples from the parental line (Figure 2.10b). These four mutations also overlap with six CNAs events, with regions in chromosome (chr) 7q and 12 being double 'hit' by two non-identical overlapping CNA events in separate samples (Figure 2.12).

The a posteriori most likely phylogenetic tree inferred by Canopy using the parental line and the MCP samples only has four subclones guided by BIC (Figure 2.10c) and is shown in Figure 2.10d. As expected, our results show that the bone and lung metastatic sublines acquire additional mutation from the parental line and form organ-specific subclones that dominate the metastasis. All samples, except MCP2287, are almost 100% comprised of cells from a single clone. Clone 2 is unique to the lung subline and clone 3 is unique to the brain subline (Figure 2.10d). MCP2287 partially retains the parental line and is a mixture of two subclones, which, upon detailed visual inspection, is supported by the raw SNA and CNA input (Figure 2.10b, Figure 2.12a-b). For CNAs, Canopy successfully resolves overlapping CNAs with correctly inferred copy number states (Figure 2.12); among the SNAs, *BRAF*, *KRAS*, and *ALPK2* each undergo a duplication event that amplifies the mutant allele, with *BRAF* and *KRAS* further losing the reference allele via a second LOH event (Figure 2.10d) that occurs later in the evolutionary process. All sublines share chr 12 duplication, while the bone and lung sublines gain additional mutations that mark and/or drive their divergence (Figure 2.10d).

Canopy's inferred phylogeny is confirmed by the SCP samples, which we use as validation. The two SCP samples derived from the lung metastasis are 100% identical to clone 2,

and the two derived from the bone metastasis are 100% identical to clone 3 (Figure 2.10d). Similar to Jacob *et al*. (83), Canopy's inferred phylogeny shows that amplification of oncogenic signals preexisting in the parent cell line (*KRAS*, *BRAF*, and *ALPK2*) leads to higher tumor-initiating fitness. Nevertheless, in contradiction to the proposed model by Jacob *et al*. (83) where no new mutations are needed, here we report additionally acquired SNA and CNAs as DNA signatures that mark and/or drive the divergence between the lung and bone sublines. These mutation signatures—chr 18q deletion, *RYR1* point mutation, and chr 7q and 12 LOH—can indicate breast cancer metastatic potentials and serve as prognostic markers for the development of distant metastasis.

2.2.7 Application to Breast Cancer Patient Xenografts

We further applied Canopy to a deep-genome sequencing dataset of breast cancer patient xenografts from Eirew *et al.* (57). Xenografts of a patient line were generated by serially transplanting breast cancer tissue organoid suspensions into immunodeficient mice (57). Whole-genome sequencing was performed on the initial engraftment (SA494T) and its subsequent propagation of metastatic xenograft (SA494X4). Targeted-amplicon deep sequencing was performed to validate somatic SNAs; TITAN (64) was applied to infer CNAs and LOH (57). We adopt bivariate clustering and stringent quality control procedures to remove experimental noise (Figure 2.13a). Canopy takes as input SNAs from four clusters that are CNA-free, three SNAs that overlap with CNAs, and four CNAs (chr1p, 3p, and 19p deletion and chr5q duplication) to reconstruct phylogeny.

The number of subclones is chosen at 4 by BIC as a criterion for model selection (Figure 2.13b). The most likely tree returned by Canopy is shown in Figure 2.13c. Clone 2 and clone 3 (2% and 1% of the starting population, SA494T) undergo a one-copy loss event and additionally acquire SNAs in cluster 3, indicating extreme selective engraftment of minor clones (Figure 2.13c). These two clones are further separated by SNAs in cluster 4 and become dominant in the subsequent metastatic xenograft SA494X4 with high prevalence (77% and 23% shown in Figure

2.13c). For SNAs that overlap with CNAs, only SNA2 precedes its affecting CNA2 (one-copy loss) and has a higher mutational multiplicity after losing the healthy allele. Both SNA1 and SNA3 arise after one-copy-loss events, resides in clone 1 only, and thus are present in sample SA494T but not in SA494X4.

We compared our analysis result to the SNA clustering result achieved by Pyclone (62). SNA clusters 1-4 correspond to the four clusters inferred by Pyclone shown in Figure 2.13c, which is expected since the SNAs within these clusters are CNA-free and these cell lines are expected to have no normal cell contamination ($CCF = MCF = 2 \times VAF$). While Pyclone outputs the clustering of these MCFs, Canopy also infers the evolutionary relationship between the clones represented by these clusters. Thus, from this analysis we can be quite confident that the mutations in cluster 2 are ancestral to the mutations cluster 4, that is, cells which carry the mutations in cluster 4 must also carry the mutations in cluster 2. Also, Pyclone uses CNA-corrected VAFs of SNAs as input whereas Canopy uses both SNAs and CNAs simultaneously to infer tumor phylogeny. This allow us to infer the temporal order of the CNA events in relation to the SNA events. For example, we are quite confident that CNAs 1 and 3 are clonal events, while CNA 2 and CNA 4 came later affecting separate subclones.

Canopy's results are confirmed by single-cell sequencing carried out by Eirew *et al.* (57)—two mutually exclusive sets of mutant alleles from SA494 tumor and passage 4 xenograft respectively were identified in addition to a set of shared alleles (57).

2.2.8 Application to Normal, Primary Tumor, and Relapse Genome of Leukemia Patients

As proof of principle and to further illustrate our method, we apply Canopy to the longitudinal dataset from Ding *et al.* (54), where whole-genome sequencing was performed on the normal tissue, the primary tumor, and the relapse genome of leukemia patients. 1292 and 412 candidate somatic SNAs and indels were identified in sample AML43/UPN869586 and AML1/UPN933142 respectively and were confirmed by deep sequencing (54). CNAs (total copy numbers) were also predicted (54).

By the weak parsimony assumption and in a similar fashion to Pyclone (62) and SciClone (63), we first adopt a binomial mixture clustering method to cluster all the mutations into mutational 'waves' and give an estimate of the VAF of each cluster (Figure 2.14a). To gain robustness against false positives calls, we add a mixture component (shown as pink dots in Figure 2.14a), with a small weight, that is uniform on the unit interval. Our clustering results show that in both patients, there is a one unique mutation cluster identified in the primary tumor, and one found at relapse. Furthermore, all mutation clusters are heterozygous and diploid, except mutation cluster 1 (mut1) in AML43, the mutations in which all reside in a copy number neutral LOH region from chr 16. The VAFs of the SNAs as well as the absolute copy number of the LOH (major copy 2, minor copy 0) are used as input for Canopy to infer phylogenetic trees.

For this dataset, Canopy returns only one plausible clonal history that can explain the observed mutation profiles, shown in Figure 2.14b. We re-parameterize the model to accommodate a redistribution event between the two time-points to improve interpretability. The tree is observed twice, first at the collection of the primary tumor, and then at the stage of relapse malignancy. Through the selection bottleneck that is imposed between the two time-points, mutations can arise (e.g., mutation cluster 5 (mut5) shown in red in Figure 2.14b) and clonal frequencies (shown in blue in Figure 2.14b) can change—some subclones expand while others become extinct or remain dormant. Meaningful quantities can be marginalized from the posterior distribution in the tree space. For example, Figure 2.14c shows the posterior distribution of the clonal frequencies through the selection bottleneck—a minor clone (clone 3 in AML43 and clone 4 in AML1) carrying the vast majority (but not all) of the primary tumor mutations survives the chemotherapy and becomes dominant at relapse by acquiring additional mutations (mut5 in both samples) while the remaining clones diminish (Figure 2.14c). Normal cell fractions are also estimated with their posterior distributions shown in the first columns of Figure 2.14c.

While Ding *et al.* (54) arrived at this same clonal history manually (a minor clone carrying the vast majority of the primary tumor mutations survived and expanded at relapse), we automate

the analysis pipeline via Canopy and allow the inclusion of both SNAs and CNAs. Canopy's inferred phylogenies shown in Figure 2.14b, as well as its estimated clonal frequencies and tumor purities shown in Figure 2.14c, are concordant with the results and conclusions in Ding *et al.* (54).

2.2.9 Application to ten spatially separated samples of ovarian cancer

We further evaluate Canopy's performance on a data set with spatial experimental design from Bashashati *et al.* (55). 63 somatic mutations (SNAs and indels) were confirmed by deep amplicon resequencing in ten tumor samples from different dissections (4a-4e, right ovary; 4f-4i, left ovary; 4j, left fallopian tube) of a high-grade serous ovarian cancer patient (Figure 2.15a-b). We keep the same assumption as in Bashashati *et al.* (55) that the 63 SNAs across all samples are heterozygous from copy number neutral regions as in the original studies: (i) CNAs weren't profiled in all samples by Affymetrix SNP genotyping arrays; (ii) for the samples with CNA calls, only total copy number is available (55).

BIC for model selection is shown in Figure 2.15c and the number of subclones is chosen at 5. Canopy returns posterior trees with one configuration and it is shown in Figure 2.15d. Different mutations correspond to rows in the heatmap in Figure 2.15a and are grouped on branches with different colors. Specifically, all ten samples share and acquire somatic mutations in *TP53* and *DHX8*, along with 13 other mutations in mutation set 2, 3, and 4 shown in light blue, green, and orange, indicating a common cell of origin. It is also observed that there is a clear separation between the samples from the right ovary and the samples from the left ovary in the clonal frequency matrix P. *GLDC*, *LIG1* as well as the rest mutations in mutation set 5 shown in blue drive and/or mark the divergence and thus have the potential to serve as a biomarker to indicate whether distal metastasis is formed in ovarian cancer patient. Mutation set 7 in red further distinguish case4a from 4b-be and form a unique subclone in case4a.

Collectively, our results suggest that multiple sublcones migrate from the left ovary to the right ovary and that both sample sets are mixtures of different subclones with diversified mutational profiles. These mutational profiles from spatially separated samples correlate with

spatial distribution due to regional evolutionary selection and reflect different histological evolutionary trajectories within a single patient.

Notably, spatial distribution of the samples in the phylogeny is concordant with the tree configuration inferred by Bashashati *et al.* (55) (Figure 2.15e). Nevertheless, the neighbor joining method with Pearson correlation distance metric doesn't account for many aspects including: (i) varying standard errors in the estimates for mutational frequencies due to varying sequencing depths; (ii) each spatial sample offers a snapshot of different combinations of subclones and therefore they cannot be treated as homogeneous samples at the tips of the tree branches; (iii) there is no placement of mutation along the tree; (iv) the inference of branch lengths assumes a constant biological clock, which doesn't hold in cancer genomes. Popic *et al.* (70) also reconstructed a clonal tree (Figure 2.15f) that is highly similar to the one returned by Bashashati *et al.* (55). Somatic mutations arise from the germline (GL) sample and are placed in the phylogeny with numbers shown on tree branches. There are three subclones with distinct mutational. The proportion of the subclonal admixtures, however, remains unknown with samples at tree tips.

2.3 Discussion

Intra-tumor heterogeneity contributes to drug resistance and failures of targeted therapies (87). To gain a comprehensive understanding of the evolutionary dynamics of tumors, it is important not only to determine which alterations drive the progression of a tumor but also to understand their relative temporal and spatial order during tumor evolution. Here we propose a novel method, Canopy, to assess intra-tumor heterogeneity and infer clonal evolutionary history. The distinguishing features of Canopy compared to existing methods are: (i) SNAs and CNAs are jointly modeled and overlapping events are phased and temporally ordered; (ii) The SNA input can be taken directly from the GATK (85) or MuTect (88) and the CNA input are continuous-valued allele-specific copy number ratios, which can be directly obtained from allele-specific copy number estimation methods (76); (iii) A pre-clustering initialization step for SNAs improves

robustness to noise and significantly reduces computation time; (iv) CNA events are allowed to be subclonal (66, 67); (v) overlapping and nested CNA events with different breakpoints affecting the same region are treated as separate evolutionary events, as illustrated by our analysis of MDA-MB-231; (vi) the Bayesian framework reconstructs the phylogeny together with posterior confidence assessment, which is useful when the data supports multiple configurations.

Despite the fact that Canopy starts with a pre-clustering, an input that contains too many false detections can still lead to unreliable phylogeny inference. Most current CNA and SNA detection algorithms still have a high false positive rate, and thus we suggest rigorous quality checking of input before a Canopy analysis. As we showed in our simulations, Canopy does not require a large set of variant loci to attain precise phylogeny inference; that is, the payoff for including multiple variants derived from the same clone quickly diminishes. A Canopy analysis should start with manual inspection and visualization of the input data, followed by removing short CNAs that may be unreliably called, and utilizing the pre-clustering procedure with a multivariate uniform component on SNAs, as illustrated in our analysis of the data from Ding *et al.* (54) and Eirew *et al.* (57).

Canopy has been demonstrated on four cancer sequencing datasets of varying study design, as well as on extensive simulation data. On a whole-exome study of breast cancer cell line MDA-MB-231, Canopy successfully deconvolved the mixed cell sublines, identifying subclones which were validated by comparing to single-cell sublines as ground truth. On a whole-genome sequencing dataset of the breast cancer tumor and its subsequent metastatic xenograft, Canopy's inferred clonal phylogeny is concordant with genomic markers of major clonal genotype and is confirmed by single-cell sequencing. On a whole-genome sequencing dataset of the primary tumor and relapse genome of a leukemia patient, and on a spatially sampled targeted sequencing study of ovarian cancer, Canopy predicted phylogenetic histories in concordance with existing knowledge. Finally, through simulations, we explored the effects of various parameters on deconvolution accuracy, and evaluate performance with comparison against existing methods.

Collectively, Canopy provides a rigorous foundation for statistical inference on repeated sequencing data from evolving populations.

Many factors determine the accuracy of Canopy's results: higher sequencing depth allows for higher sensitivity for detection of rare subclones; more samples and more difference between samples in their clonal composition allow for higher accuracy in estimating the phylogeny. In particular, the maximum number of subclones that can be reliably inferred depends on all of these factors. As the number of subclones increase, the proportion of cells attributable to at least some of the subclones would necessarily decrease, and higher coverage would be needed to detect mutations present in those smaller subclones. A survey of recent multi-region and multi-timepoint cancer genome sequencing studies shows that, even in scenarios where up to 11 bulk samples were analyzed from the same patient, the number of subclones identified was typically less than 8 (summarized in Table 2.1). A similar range for the number of subclones was found by single cell sequencing. To increase resolution for rare subclones, deeper sequencing or sequencing of a larger number of single cells is needed.

Most current cancer sequencing studies sequence only one sample from each patient, from which it is difficult to deconvolve clonal mixtures. The recent advances in single-cell sequencing technologies make possible a different approach to study tissue heterogeneity at higher resolution. Nevertheless, reliable simultaneous profiling of copy number and single nucleotide mutations by single-cell sequencing is still at infancy. Here, we show that traditional bulk sequencing can lead to accurate subclone identification and phylogenetic inference, if only the researcher is willing to sequence multiple slices of the tissue. Thus, bulk tissue sequencing can play an important part in our understanding of tumor heterogeneity, and in the coming years experimental designs that combine bulk tissue sampling and single cell analysis needs to be better explored.

2.4 Methods

2.4.1 Allele-Specific Copy Number

For the tth ($1 \le t \le T$) CNA, we let N_t be the number of germline heterozygous loci within its segment (segmentation carried out by FALCON (76) or FALCON-X). From FALCON's segmentation and phasing outputs, we can get for each tumor-normal pair the read counts of major and minor allele in the jth tumor slice, M_{ij} and m_{ij}, and in the matched normal sample, M_{i0} and m_{i0}, where $1 \le i \le N_t$ is the germline SNP index and $1 \le j \le N$ is the sample index.

For CNA events that are non-overlapping (Figure 2.3a), we use the germline heterozygous loci within each CNA segment to compute major and minor copy number input across all samples:

$$W_{tj}^{M} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(M_{ij}/M_{i0}\right), \left(\varepsilon_{tj}^{M}\right)^2 = \frac{\sum_{i=1}^{N_t}\left(M_{ij}/M_{i0}\right)^2 - N_t\left(W_{tj}^{M}\right)^2}{N_t(N_t-1)};$$

$$W_{tj}^{m} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(m_{ij}/m_{i0}\right), \left(\varepsilon_{tj}^{m}\right)^2 = \frac{\sum_{i=1}^{N_t}\left(m_{ij}/m_{i0}\right)^2 - N_t\left(W_{tj}^{m}\right)^2}{N_t(N_t-1)}.$$

In the above, W_{tj}^{M}, W_{tj}^{m} are the estimates of the major and minor copy numbers, respectively, and $\varepsilon_{tj}^{M}, \varepsilon_{tj}^{m}$ can be considered as their standard errors.

For CNA events that are overlapping or nested (Figure 2.3b), we propose a new algorithm that automates the pre-processing of allele-specific copy number for input to Canopy. If external ploidy information is available, this can be added as a fixed CNA event (e.g., a genome doubling event for tetraploidy). Specifically, we propose a 4-step prioritization algorithm to get the major and minor copy numbers for each event, briefly summarized as follows: (i) Merge CNA events where both endpoints are close, e.g. within 1 kb of each other; (ii) Identify nested CNA events, e.g., a homozygous deletion residing in a one-copy deletion region; (iii) Rank overlapping and nested CNA events by a Chi-square score, details below; (iv) Get major and minor copy number estimates through a recursive procedure. Now we expand on the details, with an illustrative example shown in Figure 2.3. Let $E_1, E_2, \ldots, E_T$ be the CNA events collected across all

samples after the merging step (i), which may contain nested or overlapping events; let $\pi_t^{(j)}$ ($1 \leq \pi_t^{(j)} \leq T$) be the ranking of event t ($1 \leq t \leq T$) in sample j ($1 \leq j \leq N$) (Figure 2.3c) based on its Chi-squared statistic,

$$Q_{tj} = \left(\frac{W_{tj}^M - 1}{\varepsilon_{tj}^M}\right)^2 + \left(\frac{W_{tj}^m - 1}{\varepsilon_{tj}^m}\right)^2 \sim \chi_2^2,$$

with larger Chi-square ranked higher (i.e. smaller $\pi_t^{(j)}$ value), but with an important caveat that nested events always takes precedence over the event that it resides in regardless of their Chi-square values, e.g., homozygous deletion event E_3 always has a higher ranking than heterozygous deletion E_2 (Figure 2.3c). Another important detail is that, at this point, the input values W^M, W^m, ε^M, and ε^m used to compute Q_{tj} are estimated from segments with shared breakpoints across all samples due to the preceding merging step. As a result, in some samples certain segments may have a mixture of more than one copy number state if it overlaps with a different CNA from another sample, e.g., in Figure 2.3b sample 1 has three copy number states in the segment that corresponds to event E_1. These segments won't have the highest Chi-squared values so they should be ranked low, as desired. To get an accurate estimate of major and minor copy numbers for overlapping and nested CNAs we adopt the algorithm outlined below, the result of which on the illustrative example is also shown in Figure 2.3c. For each sample j,

(1) Start with event t with the highest ranking: $\pi_t^{(j)} = 1$, get W_{tj}^M, W_{tj}^m, ε_{tj}^M, and ε_{tj}^m by taking the mean and standard error across all heterozygous loci that reside within this event;

(2) For event t: $\pi_t^{(j)} > 1$, in computing the major and minor copy number input, use segment E_t excluding all segments of lower rank, that is,

$$E_t \setminus \bigcup_{\pi_{t'}^j < \pi_t^j} E_{t'}.$$

2.4.2 Generalization of VAF and MCF Relationship for All Three Cases

Here we derive a general formula for the *numerator* encompassing all three cases. We denote $H \in \mathbb{R}^S$ as a vector of indicator of whether an SNA is from the major or the minor copy of the CNA

that affects it and occurs after it. We further define Q as a vector indicating whether an SNA precedes the CNA it resides in, which can be directly obtained from the tree τ_K. Let $\widetilde{H} = [H', H', \dots, H']_K \in \mathbb{R}^{S\times K}$ and $\tilde{Q} = [Q', Q', \dots, Q']_K \in \mathbb{R}^{S\times K}$. Then, the *numerator* for all three cases shown in Figure 2.2 can be generalized and division of the *numerator* by the *denominator* gives us the VAF matrix:

$$VAF = \frac{\left\{Z \cdot \left(Y \times \begin{bmatrix} \mathbb{1} \\ \dots \\ \tilde{C}^M \end{bmatrix}\right)^{\tilde{Q}} \cdot \widetilde{H} + Z \cdot \left(Y \times \begin{bmatrix} \mathbb{1} \\ \dots \\ \tilde{C}^m \end{bmatrix}\right)^{\tilde{Q}} \cdot \left(1 - \widetilde{H}\right)\right\} \times P}{\left(Y \times \begin{bmatrix} \mathbb{1} \\ \dots \\ \tilde{C}^M \end{bmatrix} + Y \times \begin{bmatrix} \mathbb{1} \\ \dots \\ \tilde{C}^m \end{bmatrix}\right) \times P} \in \mathbb{R}^{S\times N}.$$

Note that the exponentiation and division are carried out in an element-wise fashion and that 0^0 is defined to be equal to 1. This generalized matrix representation form to get VAFs of SNAs only apply to SNAs that are CNA-free or those that are affected by a single CNA event. For SNAs that are affected by more than one CNA event, VAFs are obtained iteratively for each SNA with adjustment of the affecting CNA events that are overlapping or nested.

2.4.3 Simulation Setup

We firstly generate input data from the true underlying tree with and without overlapping CNAs respectively and apply Canopy to reconstruct the phylogeny. For SNAs, the total read depth matrix X has each of its column sampled from a multinomial distribution

$$X_{:,j} \sim \text{Multinomial}\left(d \times S, \frac{1}{S}, \dots, \frac{1}{S}\right),$$

where d is the mean sequencing depth and $1 \le j \le N$. The mutant read depth matrix R is sampled from a binomial distribution indexed at X with success probabilities VAF derived in 2.2.4 SNA-CNA Phase and Combined Likelihood (numerator divided by denominator). For CNAs, the input matrix W^M and W^m are sampled from a normal distribution with mean $\tilde{C}^M \times P$ and $\tilde{C}^m \times P$ and standard deviation ε^M and ε^m ranging from 0.001 to 0.64 (Figure 2.5d). The matrices X, R, W^M, W^m, ε^M, and ε^m are then used as input for Canopy to infer phylogeny with output shown in

Figure 2.5a. For Clomial (72), we keep its assumptions and use X and R as input to infer phylogeny with result shown in Figure 2.5b.

We then separately investigate the effects of the number of mutations, the sequencing depth, the number of samples, the number of subclones, and the pre-clustering procedure as an initialization step on deconvolution and pre-clustering accuracy and computation time. Without loss of generality, we focus on using SNAs to reconstruct phylogeny and compare against two existing methods, Clomial (72) and SciClone (63). For each investigation, we control for confounding parameters, run 30 simulations in parallel, and integrate results from each run. Within each simulation, we run 10 Markov chains with random starts and correspondingly choose $binomTryNum = 10$ for Clomial (72), a parameter specifying the number of random starts for the EM algorithm. The true clonal frequency matrix P is pre-fixed but varies between different runs with a perturbation added to each of its element from a Gaussian distribution with mean 0 and standard deviation 0.01. The generated matrix is then scaled so that each element is non-negative and that the columns sum up to one. We calculate the percentage of wrongly labeled elements in Z (Figure 2.6) and the RMSE of the inferred P matrix (Figure 2.7) across all simulation runs.

Number of mutations and sequencing depth

We start with constructing a true underlying tree with a fixed number of subclones. Various numbers of mutations are placed on branches of the tree (except for the leftmost one) with equal probabilities and as a result we can get a true genotyping matrix Z. The clonal frequency matrix P is fixed so that we can control for the number of subclones, the number of samples, and the clonal compositions. Here we mimic two different sequencing pipelines—whole-genome sequencing with $d = 30$ and targeted sequencing with $d = 500$. The input matrix X is sampled from the multinomial distribution and the mutant read depth matrix R is then sampled from a binomial distribution

$$R \sim \text{Binomial}\left(X, \frac{1}{2}Z \times P\right).$$

Number of samples

We evaluate the effect of the number of samples by running parallel simulations with fixed number of subclones ($K = 5$) and mutation clusters ($S = 7$) but varied number of samples, which correspond to columns of the clonal frequency matrix P. Since adding a same sample doesn't guarantee adding additional information for phylogeny reconstruction, we choose and fix the elements of the P matrix so that the additive summation result is the most distinct in the unit space and that different combinations of subclones are present across different samples. We further measure the deconvolution difficulty quantitatively from the P matrix itself. Specifically, we define $q \in \mathbb{R}^{(2K-3)\times N}$ as the summation of the offspring subclonal frequencies at each of the $(2K-3)$ internal edges across all samples,

$$q_{ij} = \sum_{\{s:\, s \text{ is descedant of edge } i\}} P_{sj}\ (1 \le i \le 2K-3, 1 \le j \le N).$$

The statistic that we use to measure the deconvolution difficulty of the the P matrix is

$$q_{min} = \min_{\{i \ne i'\}} \|q_i - q_{i'}\|^2,$$

where $q_i = (q_{i1}, q_{i2}, \dots, q_{iN})$ (Figure 2.8).

Number of subclones

We study the effect of the number of subclones by keeping the P matrix the same with varied number of rows ($3 \le K \le 10$). The number of samples is fixed at 3, among which there is the greatest distinction of clonal compositions; the number of mutations is set at $K + 2$. In addition to measure the accuracy of the inferred Z and P matrix, we also compare Canopy's pre-clustering result against that of SciClone's (63). We use *clustering purity* as a measure of clustering quality. To compute *clustering purity*, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned mutations and dividing by the total number of mutations. We further carry out simulations to examine a larger subclone space and to investigate the tradeoff between the number of subclones, the sequencing depth, as well as the number of mutations. Running time, estimation errors of the Z and the P matrix are recorded (Table 2.3).

Binomial mixture clustering

We investigate the effect of the binomial mixture clustering on computation time and deconvolution accuracy. The binomial mixture clustering is carried out as an initialization step to guide the MCMC sampling procedure – we firstly move the mutational clusters along the tree branches and then fine tune every mutation within each cluster. Simulation is carried out with varying number of mutations $N \in \{25,50,100,200\}$ along trees with different number of clones $K \in \{3,4,5,6\}$ from three samples. The true underlying clonal frequency matrix P is the same as is in the previous section. Convergence is measured by both the log-likelihood and the acceptance rate (Figure 2.9), with running time recorded and estimation errors measured (Table 2.3).

2.4.4 WES of Transplantable Metastasis Model Derived from MDA-MB-231

The parental cell line MDA-MB-231 was obtained from the American Type Tissue Collection. Its derivative cell lines (both SCPs and MCPs) were described previously (82, 89, 90). Cells were grown in high-glucose DMEM medium with 10% fetal bovine serum. Genomic DNA was harvested with Purelink genomic DNA kit (Invitrogen). Exome libraries were prepared with SureSelect Human All Exon kit (Agilent) and were sequenced on an Illumina Hiseq-2000 sequencer. The WES data have been deposited in the BioProject database with accession number PRJNA315318.

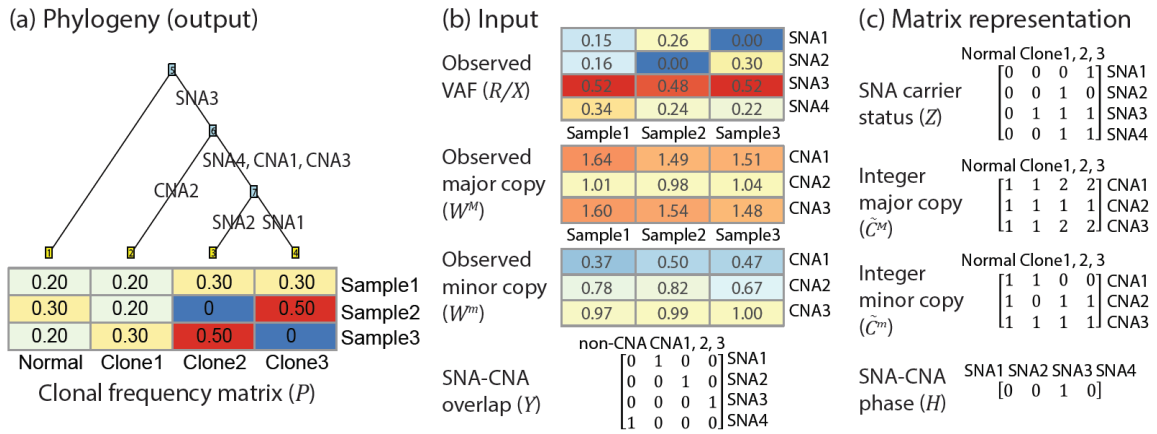


Figure 2.1: ***Tumor phylogeny, observed input, and inferred output of Canopy.*** *(a) Phylogeny of tumor progression as a bifurcating tree with SNAs and CNAs along the branches. Longitudinal and/or spatial samples offer different snapshot of subpopulations, represented by tree leaves. The lengths of the branches are arbitrary—since without further strong assumptions, we cannot infer branch length from this data. (b) Observed VAFs, major copies, and minor copies across samples. Matrix Y indicates whether an SNA resides in a CNA. (c) Matrix decomposition by Canopy. Genotyping matrix Z represents the positions of the SNAs in the phylogeny. $\tilde{C}^M$ and $\tilde{C}^m$ encode major and minor copy number of each clone. H specifies SNA-CNA phasing—whether SNAs reside in major or minor copies. Clonal frequency matrix P is shown as part of (a).*

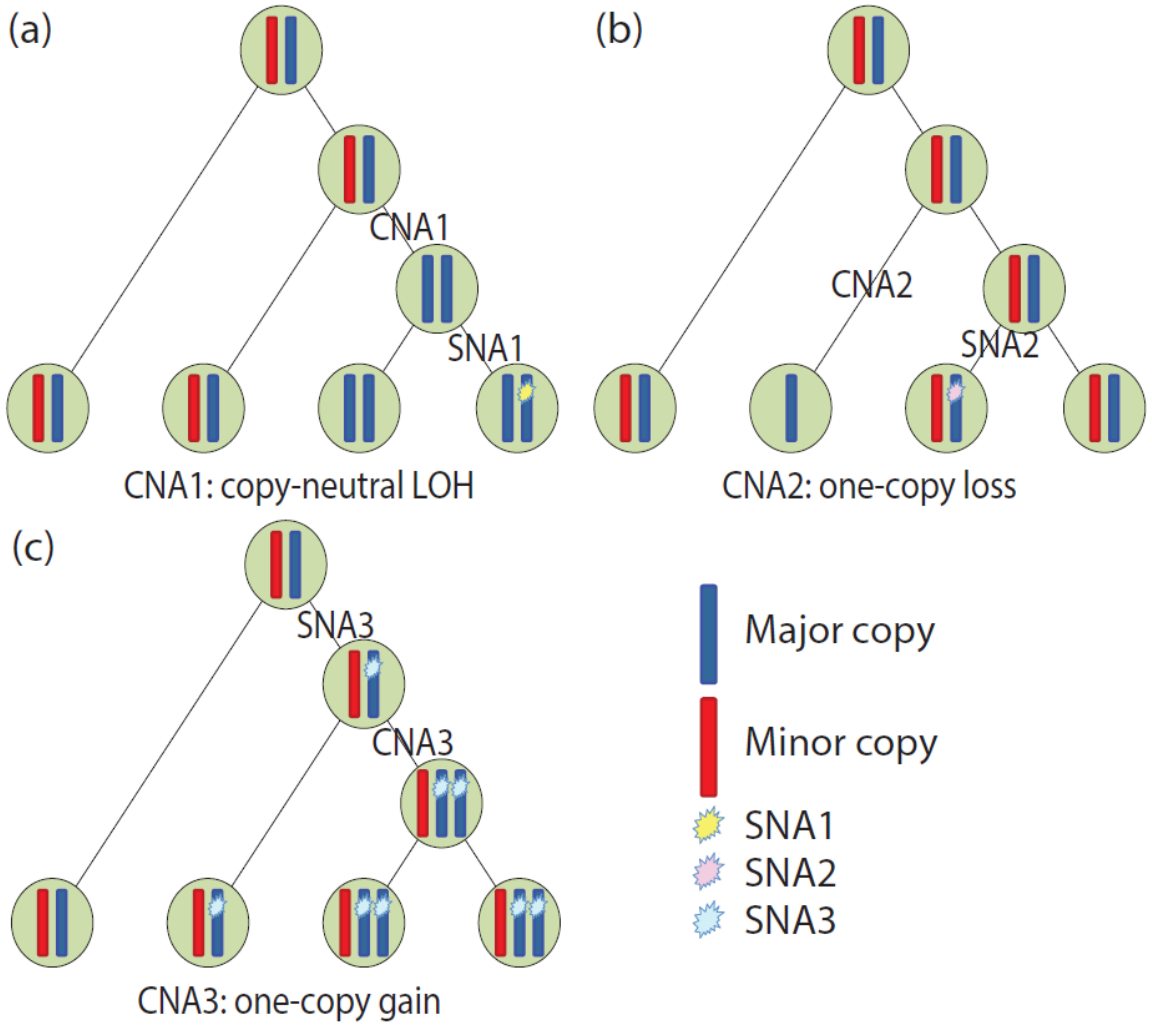


Figure 2.2: ***Three cases of SNA-CNA phase and order.*** *Different phases and orders of CNA and the SNA it affects are shown with clonal histories concordant with Figure1. Major and minor copies are in blue and red respectively; SNA mutational loci are shown as stars. (a) CNA precedes SNA. SNA resides in only one chromosomal copy. (b) CNA and SNA are on two separate branches. SNA is unaffected by CNA. (c) SNA precedes CNA. Scenario where major copies contain the SNA is shown. SNA4 from Figure 2.1 is unaffected by CNA and is not shown.*

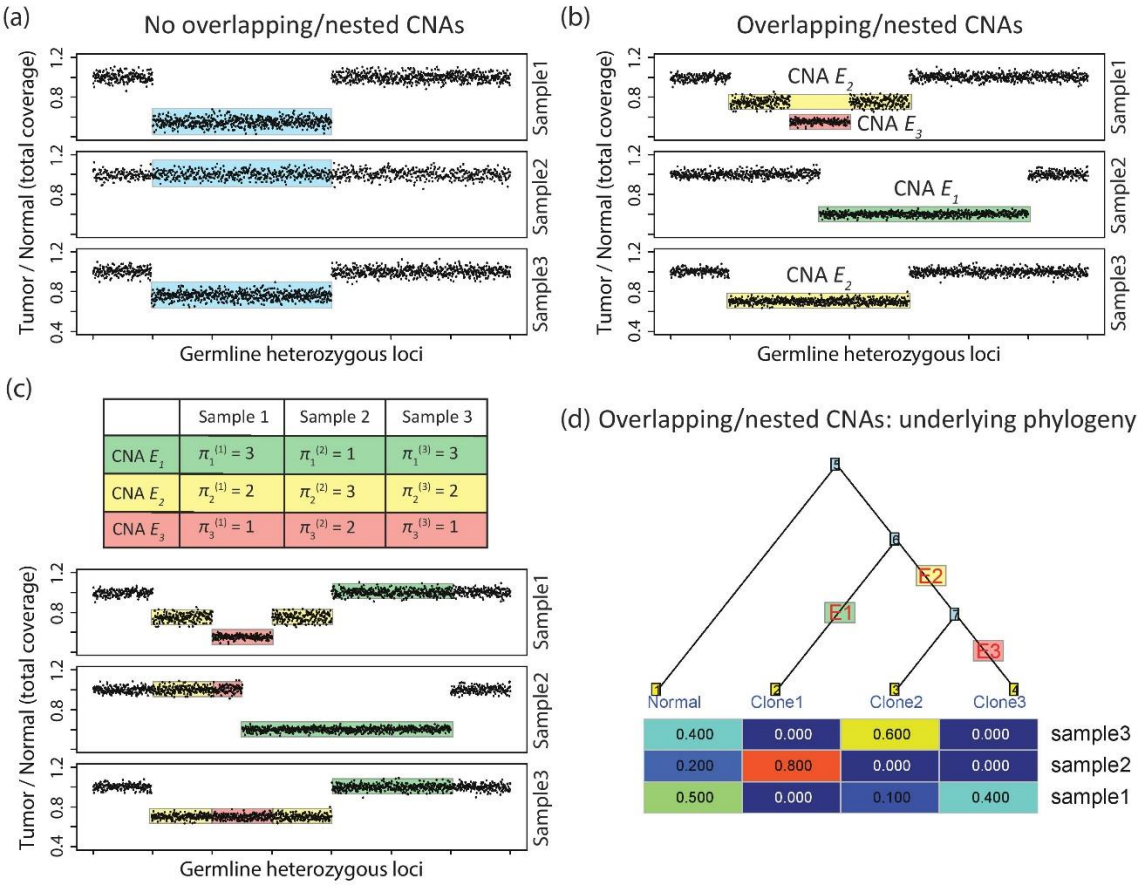


Figure 2.3: ***Illustration on generating CNA input for Canopy.*** *Initial segmentation is performed by FALCON-X. (a) For CNAs that aren't overlapping or nested, the segment mean and standard error are computed for each segment across all samples (Methods in main manuscript). (b) For CNA events that display overlapping/nested structure, a four-step CNA prioritization algorithm (Supplementary Methods) is adopted. (c) The ranking of CNA events in each sample and the segments that are used to generate allele-specific copy number calls. (d) The underlying tree structure for samples and CNA events shown in (b).*

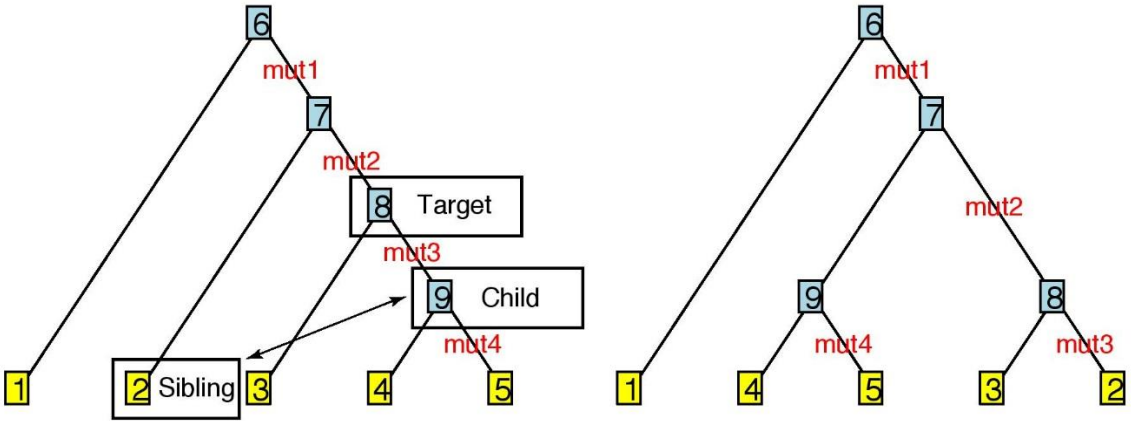


Figure 2.4: ***Generating new tree topology by local rearrangement.*** *A neighborhood—an internal node that has both a parent and two children—is selected for local rearrangement. Switch the sibling with one of the children to generate a new tree topology (91).*

(a) Canopy's esimates

$$\hat{Z} \in \mathbb{R}^{S\times K} = \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix}$$
$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{SNA1} \\ \text{SNA2} \\ \text{SNA3} \\ \text{SNA4} \end{matrix}$$

$$\hat{P} \in \mathbb{R}^{K\times N} = \begin{matrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \end{matrix}$$
$$\begin{bmatrix} 0.18 & 0.30 & 0.19 \\ 0.19 & 0.18 & 0.31 \\ 0.33 & 0.00 & 0.50 \\ 0.30 & 0.52 & 0.00 \end{bmatrix} \begin{matrix} \text{Clone1} \\ \text{Clone2} \\ \text{Clone3} \\ \text{Clone4} \end{matrix}$$

$$\hat{\tilde{C}}^{M} \in \mathbb{R}^{T\times K} = \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix}$$
$$\begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 \end{bmatrix} \begin{matrix} \text{CNA1} \\ \text{CNA2} \\ \text{CNA3} \end{matrix}$$

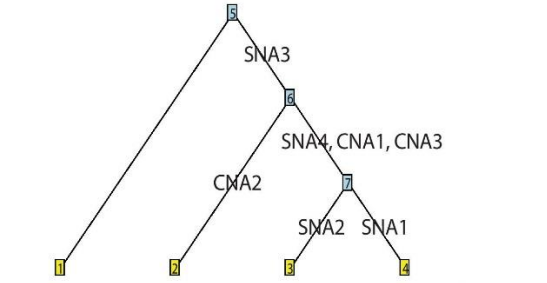
$$\hat{\tilde{C}}^{m} \in \mathbb{R}^{T\times K} = \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix}$$
$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{CNA1} \\ \text{CNA2} \\ \text{CNA3} \end{matrix}$$

(b) Clomial's esimates

$$\hat{Z} \in \mathbb{R}^{S\times K} = \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix}$$
$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{SNA1} \\ \text{SNA2} \\ \text{SNA3} \\ \text{SNA4} \end{matrix}$$

$$\hat{P} \in \mathbb{R}^{K\times N} = \begin{matrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \end{matrix}$$
$$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.37 & 0.47 & 0.49 \\ 0.36 & 0.00 & 0.51 \\ 0.27 & 0.53 & 0.00 \end{bmatrix} \begin{matrix} \text{Clone1} \\ \text{Clone2} \\ \text{Clone3} \\ \text{Clone4} \end{matrix}$$

(c) True phylogeny (Figure 1)



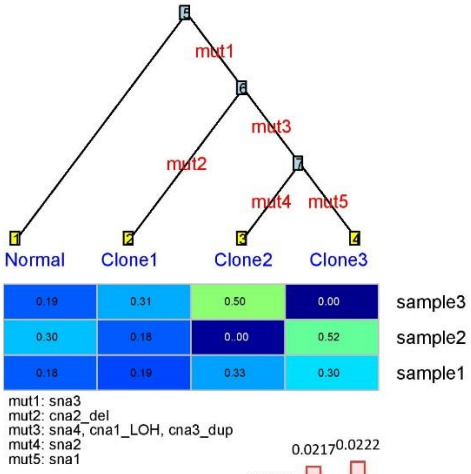
0.20	0.20	0.30	0.30	Sample1
0.30	0.20	0	0.50	Sample2
0.20	0.30	0.50	0	Sample3
Normal	Clone1	Clone2	Clone3	

Clonal frequency matrix (P)

Estimated phylogeny (PhyloWGS output)

Node	Cellular prevalence	CCF	SSMs	CNVs
0	1	0	0	0
1	0.948	1	2	2
2	0.44	0.464	1	1
3	0.147	0.155	1	0

Estimated phylogeny (Canopy output)



(d)

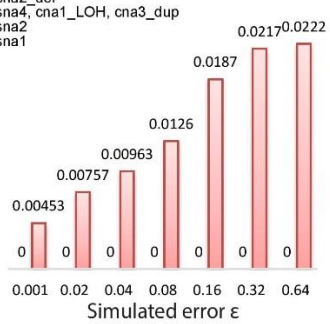


Figure 2.5: ***Inferred phylogenies by Canopy, Clomial and PhyloWGS.*** *(a) Canopy successfully decomposes all matrices with confidence assessment. (b) Clomial doesn't utilize somatic CNA information and fails to estimate the clonal frequencies with zero normal cell contaminations in all three samples. The true quantities are shown in Figure 1. (c) True phylogeny and estimated phylogeny by Canopy and PhyloWGS. Canopy returned a tree highly concordant with the ground truth whereas PhyloWGS returned a linear tree with incorrectly inferred cellular frequencies. The input for this dataset can be found in the Canopy R-package. (d) Higher noise (spiked-in error term ε) doesn't seem to affect Canopy's estimation of the genotyping matrix Z but leads to higher estimation error of the clonal proportion P. The estimation error is taken as the median across ten parallel runs.*

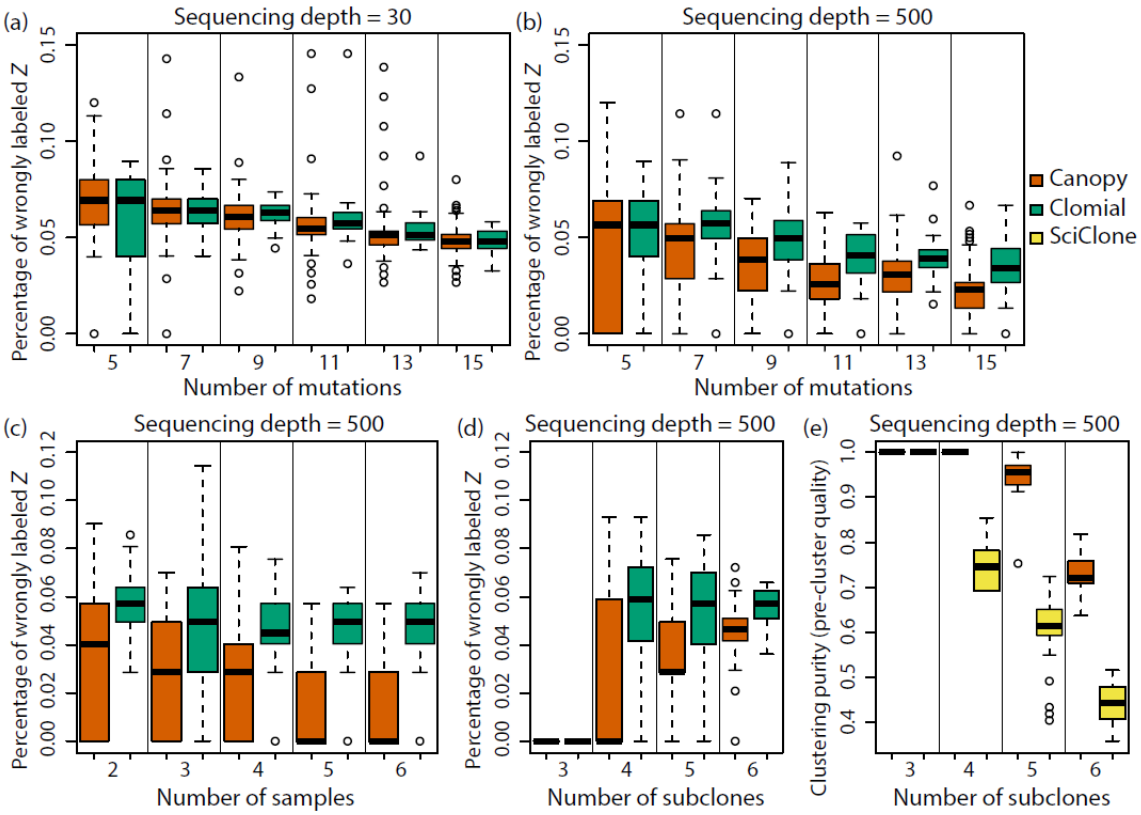


Figure 2.6: ***Deconvolution accuracy and clustering quality via simulation studies.*** *Various parameters show effects on deconvolution accuracy (measured by the percentage of wrongly labeled Z elements) and pre-clustering quality (measured by the clustering purity). Canopy is compared against Clomial and SciClone and is shown to have better performance. (a-b) Whole-genome sequencing compensates its low sequencing depth with more profiled mutations. (c) Increasing sample size helps solve reconstruction ambiguity. (d-e) Number of subclones is negatively correlated with deconvolution accuracy and pre-clustering quality.*

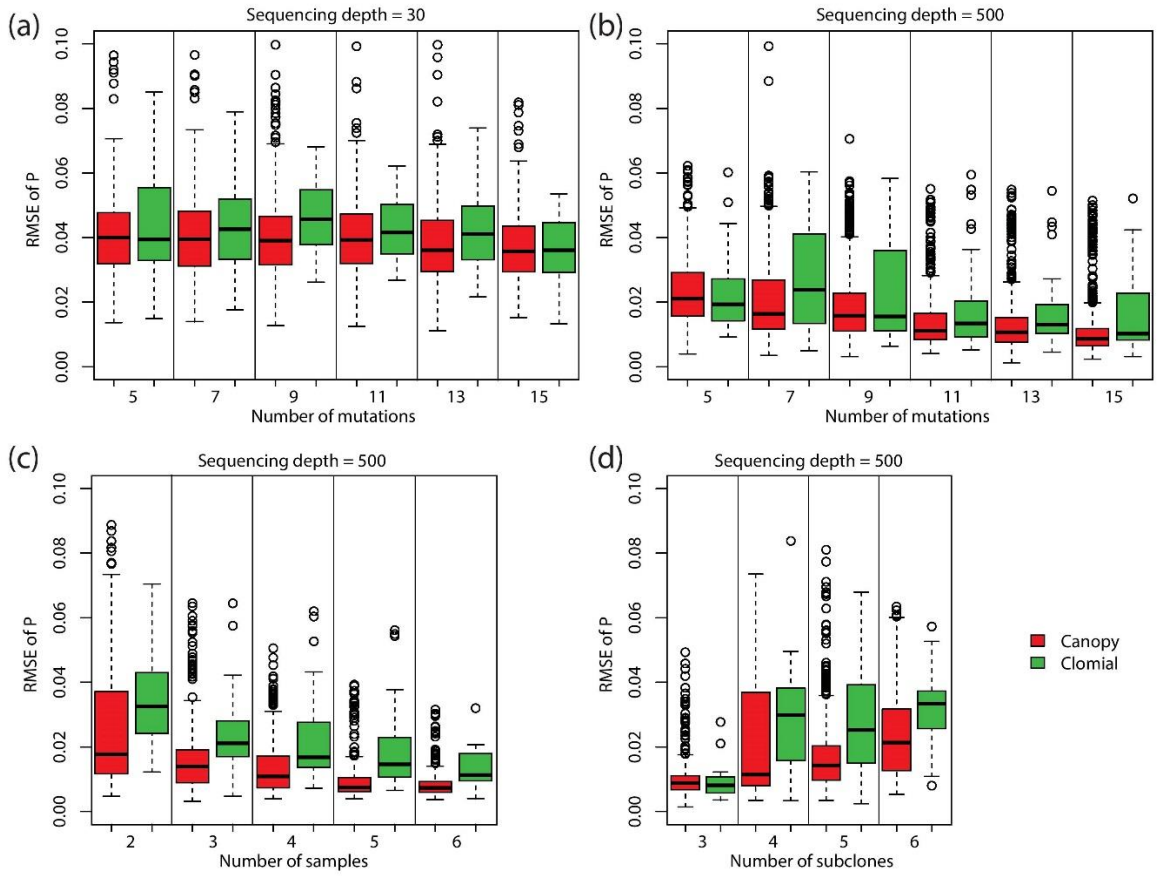


Figure 2.7: ***Deconvolution accuracy via simulation studies.*** *Various parameters show effects on deconvolution accuracy (measured by RMSE of the P matrix). (a-b) Whole-genome sequencing compensates the lower sequencing depth with more profiled mutations. (c) Large number of samples helps solve reconstruction ambiguity. (d) Number of subclones is negatively correlated with deconvolution accuracy. Canopy outperforms Clomial under all settings.*

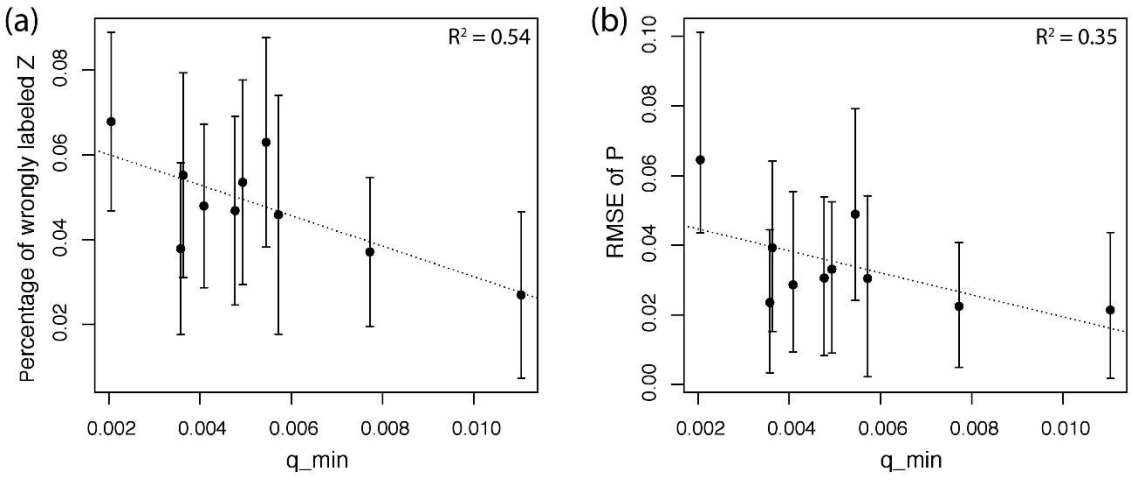


Figure 2.8: ***q_{min} as a measure of deconvolution difficulty from the clonal frequency matrix P.*** *The larger the q_{min} is, the more distinct the clonal frequencies at the tree edges are, and thus the more difficult the deconvolution problem is.*

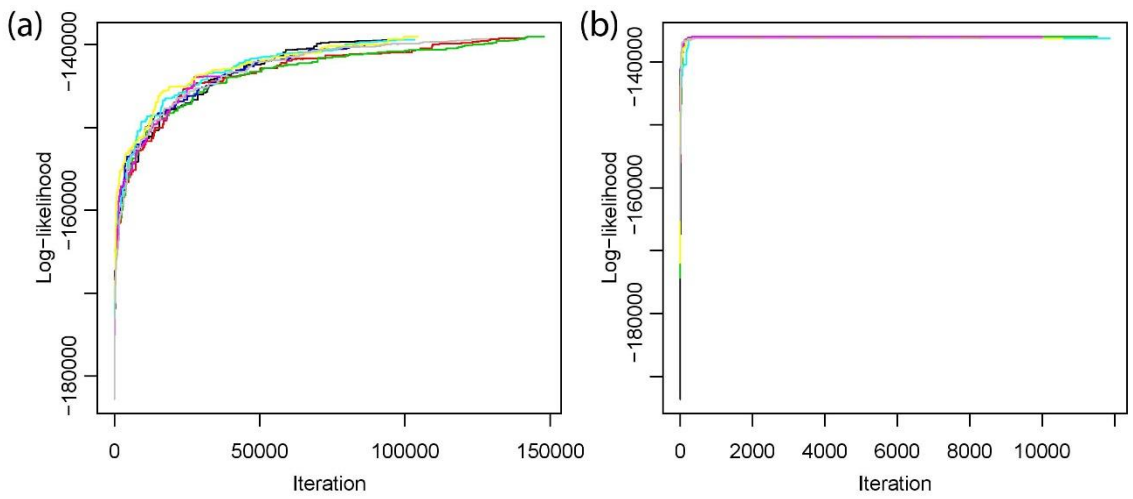


Figure 2.9: ***Log-likelihood of MCMC sampling with and without pre-clustering step.*** *Simulation is carried out with 200 mutations along a five-branch tree using three samples. Ten chains shown in different color are randomly started with (a) and without (b) a Binomial mixture clustering step. Convergence is measured by both the log-likelihood and the acceptance rate. Pre-clustering step significantly reduces computation time with MCMC converging faster.*

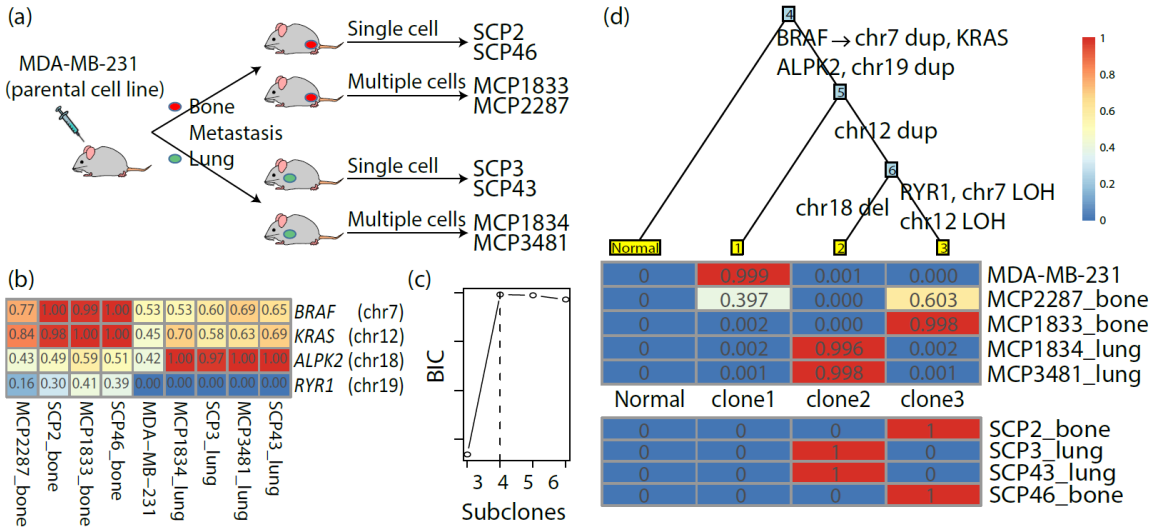


Figure 2.10: ***Clonal history of transplantable metastasis model MDA-MB-231 with validation by SCP samples.*** *(a) Transplantable model system of MDA-MB-231. Parental line is injected into mouse models and induces organ-specific metastasis. Sublines are derived from single or multiple cell(s) from different metastatic sites. (b) Observed VAFs of somatic SNAs, which reside in nested CNAs. Canopy takes both SNA and CNA input. (c) BIC as a model selection method to determine the number of subclones. (d) Clonal tree reconstructed by Canopy. Sublines acquire additional mutation from the parental line and form organ-specific subclones that dominate the metastasis. SCP samples successfully validate the subclones and confirm Canopy's inferred phylogeny.*

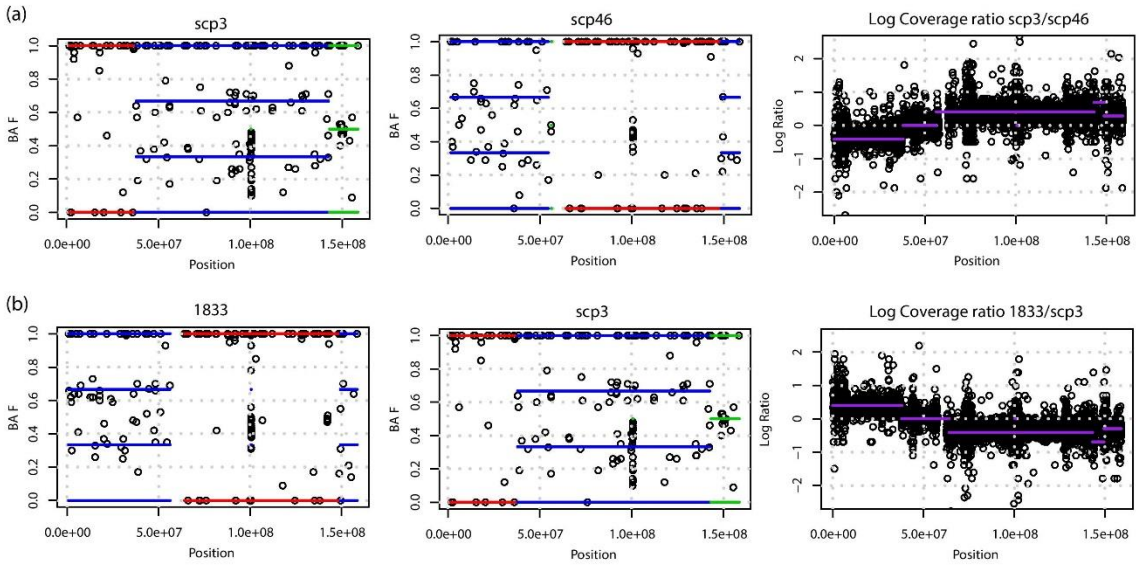
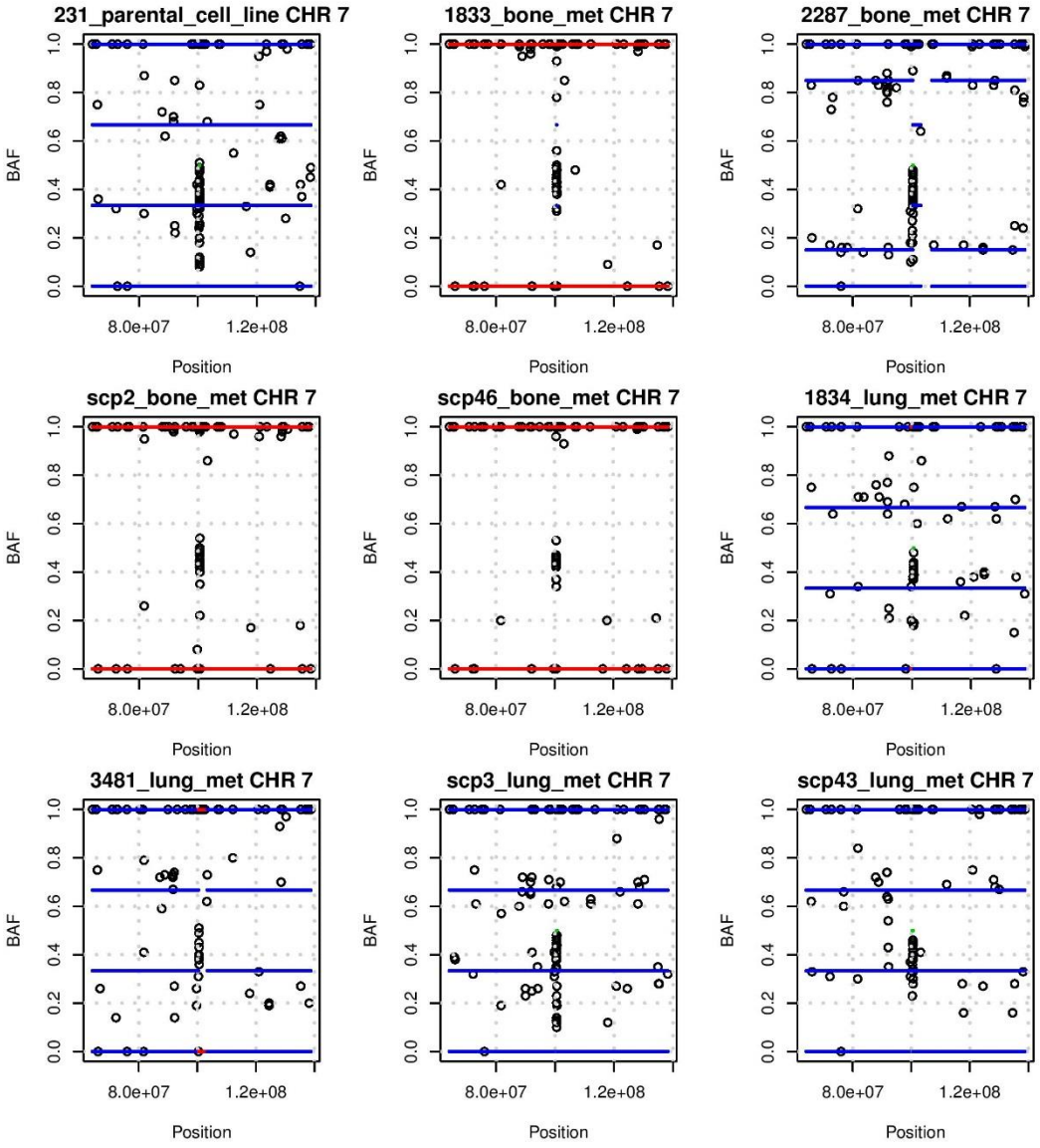
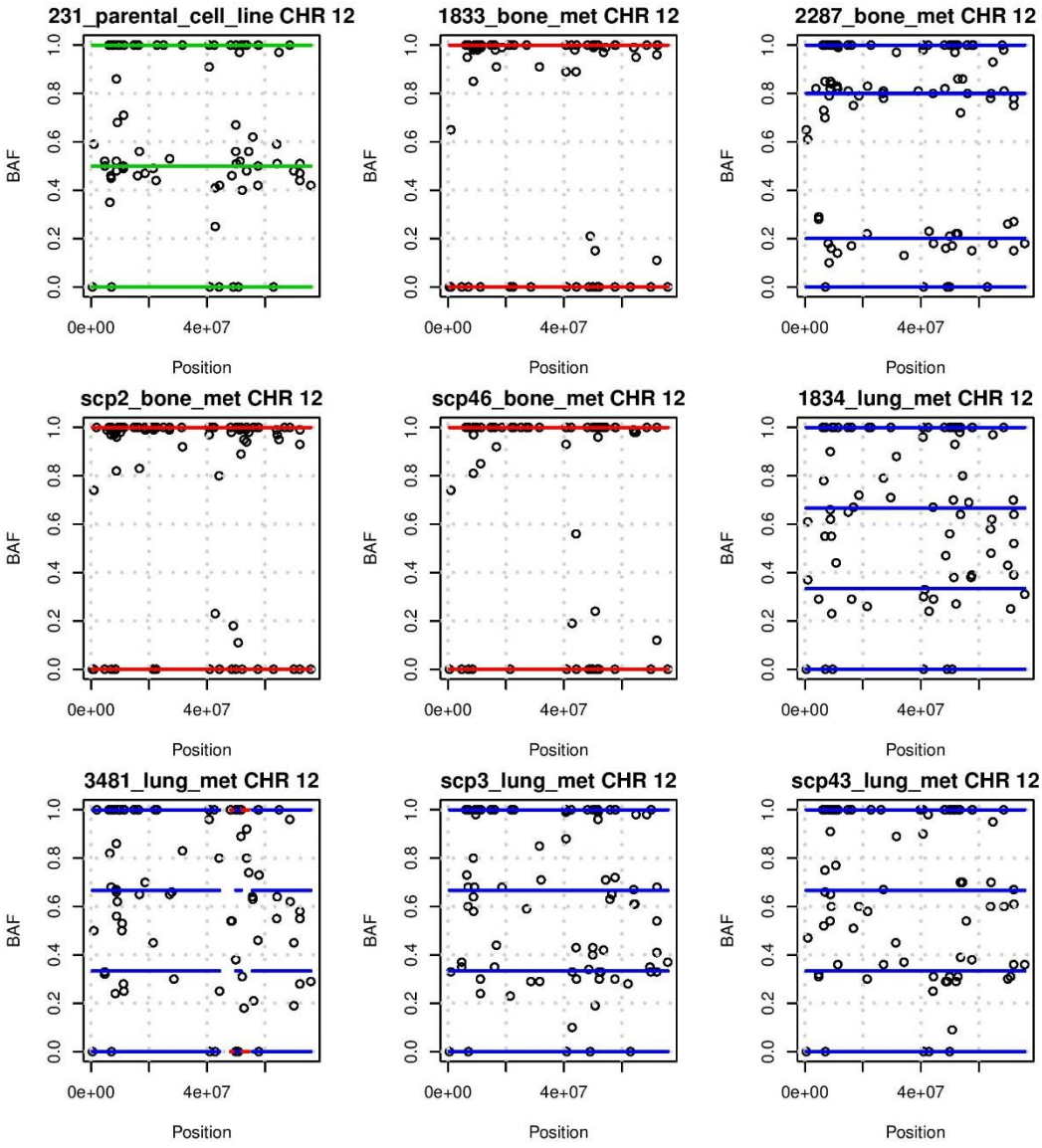


Figure 2.11: ***CNA inference by HMM.*** *(a) HMM is applied to segment the genome in SCP samples and manually corrected by the exonic coverage ratios between two SCPs. B allele frequencies (BAFs) are used as input. Deletion/LOH is shown in red, duplication in blue, and copy number neutral region in green. Purple line is the log ratio of the segmented total copy numbers, overlaid by the corresponding depth of coverage ratio. (b) Using SCP as a normal control, CNAs for the MCP sample is called.*

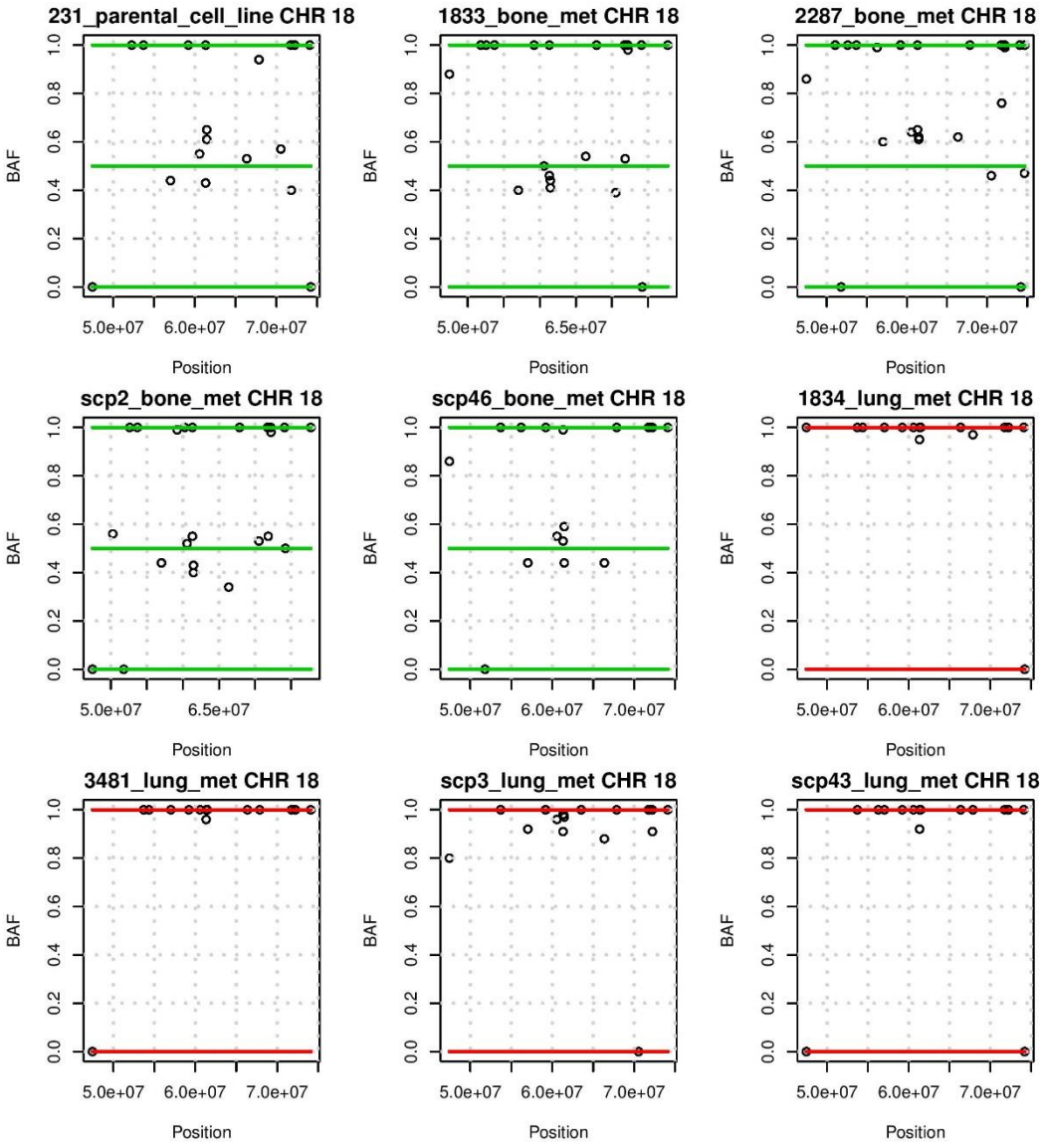
(a)


231_parental_cell_line CHR 7
1833_bone_met CHR 7
2287_bone_met CHR 7
scp2_bone_met CHR 7
scp46_bone_met CHR 7
1834_lung_met CHR 7
3481_lung_met CHR 7
scp3_lung_met CHR 7
scp43_lung_met CHR 7
BAF
Position
0.0
0.2
0.4
0.6
0.8
1.0
8.0e+07
1.2e+08

(b)


231_parental_cell_line CHR 12
1833_bone_met CHR 12
2287_bone_met CHR 12
scp2_bone_met CHR 12
scp46_bone_met CHR 12
1834_lung_met CHR 12
3481_lung_met CHR 12
scp3_lung_met CHR 12
scp43_lung_met CHR 12
BAF
Position
0e+00
4e+07
0.0
0.2
0.4
0.6
0.8
1.0

(c)


231_parental_cell_line CHR 18
1833_bone_met CHR 18
2287_bone_met CHR 18
scp2_bone_met CHR 18
scp46_bone_met CHR 18
1834_lung_met CHR 18
3481_lung_met CHR 18
scp3_lung_met CHR 18
scp43_lung_met CHR 18
BAF
Position
0.0
0.2
0.4
0.6
0.8
1.0
5.0e+07
6.0e+07
6.5e+07
7.0e+07

(d)

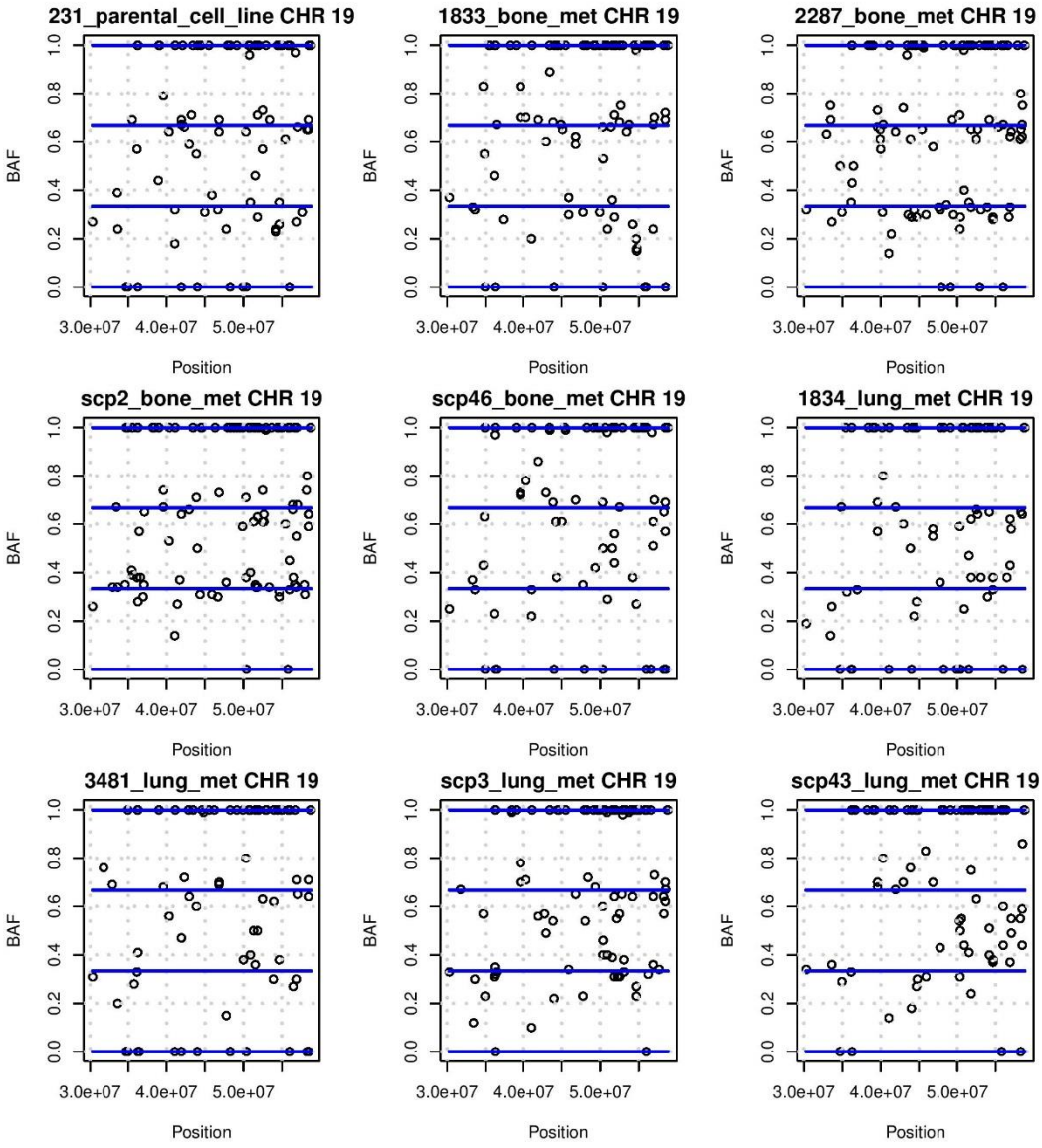


Figure 2.12: ***Canopy's CNA input to infer phylogeny in the parental cell line and its sublines.*** *Six somatic CNAs from four different chromosomes—(a) chr7, (b) chr12, (c) chr18, (d) chr19. Chr7 and chr12 are double 'hit' by two CNAs; chr18 and chr19 undergo one-copy loss and gain respectively. CNA subclonal events result in different allele specific copy number states across different samples. The observed B allele frequencies (BAFs), i.e.,* $W^M/(W^M + W^m)$ *and* $W^m/(W^M + W^m)$*, are used as input for Canopy to infer the clonal tree.*

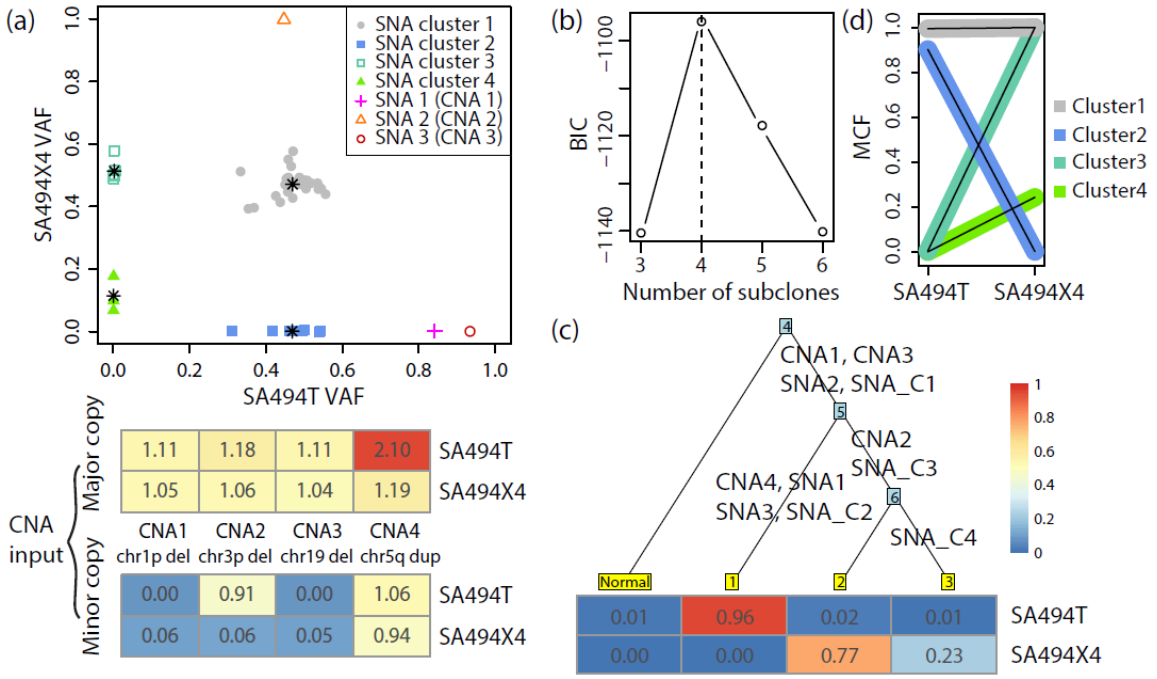


Figure 2.13: ***Clonal architecture of breast cancer initial engraftment and passage xenograftment.*** *Tumor sample SA494T and its subsequent xenograft SA494X4 are whole-genome sequenced with SNAs validated by deep amplicon resequencing and CNAs inferred by TITAN. (a) SNA and CNA input of Canopy. VAFs of four SNA clusters and three CNA-affected SNAs are shown in the top panel. Heatmap of observed major and minor copy numbers are shown in the bottom panel. (b) BIC as a model selection metric to determine the number of subclones. (c) The most likely tree returned by Canopy based on the mutational profiling. Extreme selection of minor clones is imposed on engraftment. SA494T and SA494X4 bear two mutually exclusive sets of mutations in addition to shared ancestral mutations. (d) Mutation clusters inferred by the Pyclone model.*

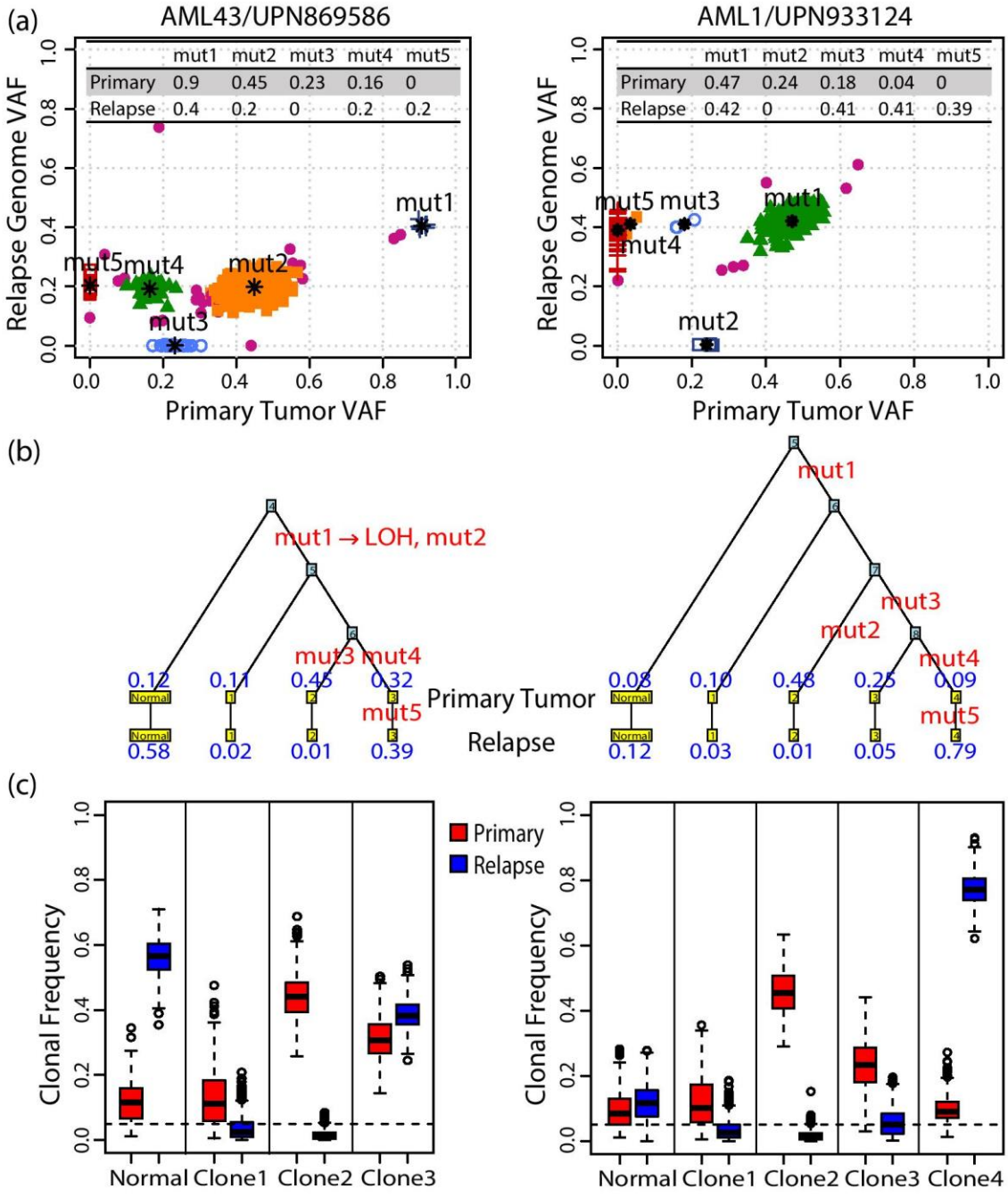

(a)
AML43/UPN869586
mut1 mut2 mut3 mut4 mut5
Primary 0.9 0.45 0.23 0.16 0
Relapse 0.4 0.2 0 0.2 0.2
Relapse Genome VAF
Primary Tumor VAF
mut1
mut2
mut3
mut4
mut5
AML1/UPN933124
mut1 mut2 mut3 mut4 mut5
Primary 0.47 0.24 0.18 0.04 0
Relapse 0.42 0 0.41 0.41 0.39
(b)
mut1 → LOH, mut2
0.12 0.11 0.45 0.32
0.58 0.02 0.01 0.39
mut3 mut4
mut5
Primary Tumor
Relapse
0.08 0.10 0.48 0.25 0.09
0.12 0.03 0.01 0.05 0.79
(c)
Primary
Relapse
Clonal Frequency
Normal Clone1 Clone2 Clone3
Normal Clone1 Clone2 Clone3 Clone4

Figure 2.14: ***Clonal history reconstructed from primary tumor and the relapse genome of leukemia patients.*** *(a) VAFs of SNAs and indels of the primary tumor and the relapse genome of patient AML43 and AML1 are clustered into mutational waves shown in different colors. A mixture component with a small weight shown as pink dots is included to gain robustness against false positives. CNAs for each mutational cluster are profiled. SNAs and CNAs are used as input for Canopy. (b) Plausible phylogenies inferred by Canopy, observed at two time-points. Mutations and clonal proportions are shown in red and blue respectively. Both trees support the model that a subclone from the primary tumor gains additional mutations and expands at relapse. (c) Inference of clonal frequency from the posterior distribution. One subclone survives the chemotherapy and becomes dominant. Normal cell contaminations/tumor purities are estimated as the first columns.*

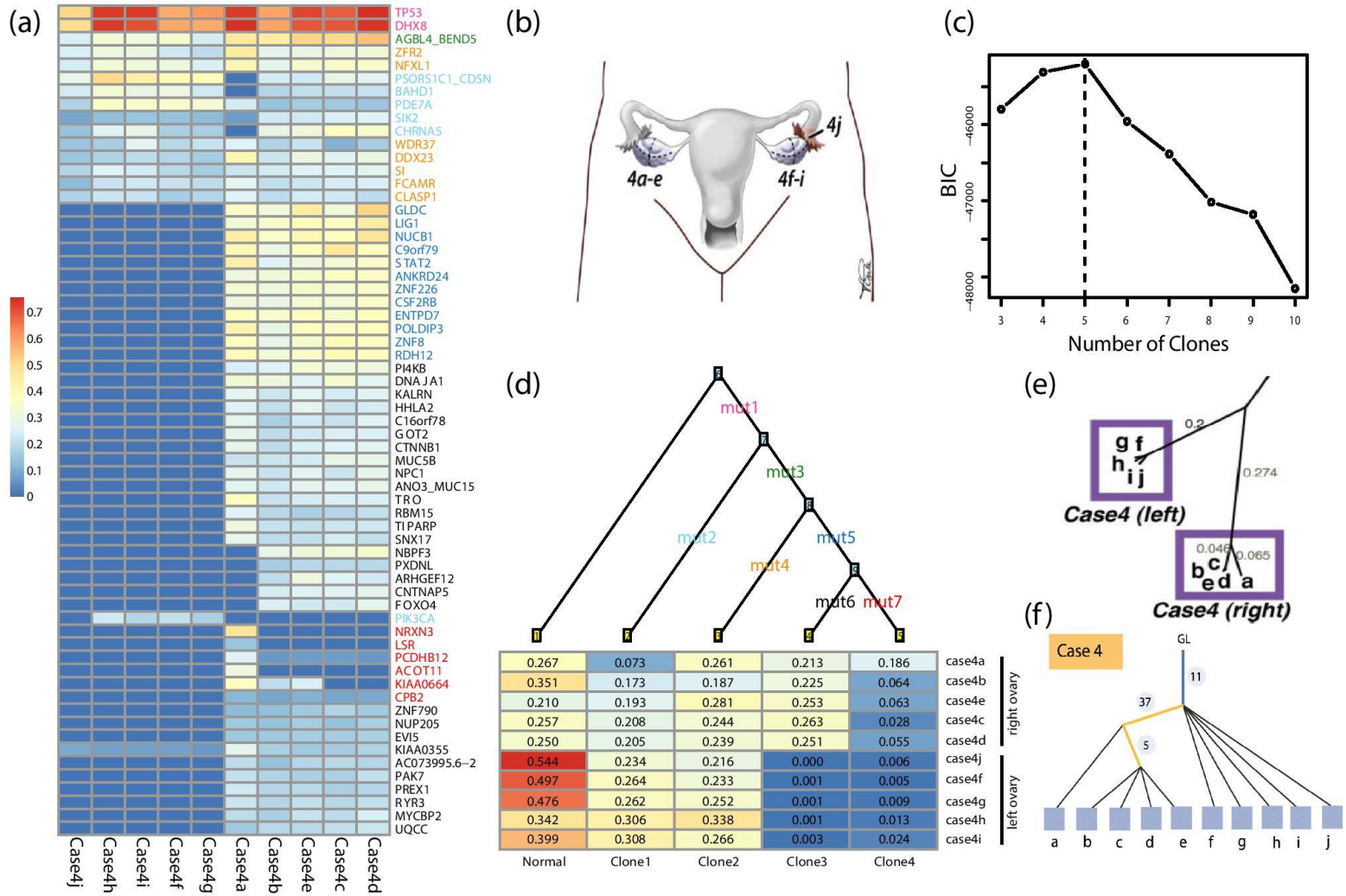

(a)
TP53
DHX8
AGBL4_BEND5
ZFR2
NFXL1
PSORS1C1_CDSN
BAHD1
PDE7A
SIK2
CHRNA5
WDR37
DDX23
SI
FCAMR
CLASP1
GLDC
LIG1
NUCB1
C9orf79
STAT2
ANKRD24
ZNF226
CSF2RB
ENTPD7
POLDIP3
ZNF8
RDH12
PI4KB
DNAJA1
KALRN
HHLA2
C16orf78
GOT2
CTNNB1
MUC5B
NPC1
ANO3_MUC15
TRQ
RBM15
TIPARP
SNX17
NBPF3
PXDNL
ARHGEF12
CNTNAP5
FOXO4
PIK3CA
NRXN3
LSR
PCDHB12
ACOT11
KIAA0664
CPB2
ZNF790
NUP205
EVI5
KIAA0355
AC073995.6--2
PAK7
PREX1
RYR3
MYCBP2
UQCC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Case4j
Case4h
Case4i
Case4f
Case4g
Case4a
Case4b
Case4e
Case4c
Case4d
(b)
4j
4a-e
4f-i
(c)
BIC
−46000
−47000
−48000
3
4
5
6
7
8
9
10
Number of Clones
(d)
mut1
mut3
mut2
mut5
mut4
mut6
mut7
0.267 0.073 0.261 0.213 0.186 case4a
0.351 0.173 0.187 0.225 0.064 case4b
0.210 0.193 0.281 0.253 0.063 case4e
0.257 0.208 0.244 0.263 0.028 case4c
0.250 0.205 0.239 0.251 0.055 case4d
0.544 0.234 0.216 0.000 0.006 case4j
0.497 0.264 0.233 0.001 0.005 case4f
0.476 0.262 0.252 0.001 0.009 case4g
0.342 0.306 0.338 0.001 0.013 case4h
0.399 0.308 0.266 0.003 0.024 case4i
Normal
Clone1
Clone2
Clone3
Clone4
right ovary
left ovary
(e)
0.2
0.274
0.046
0.065
Case4 (left)
Case4 (right)
(f)
Case 4
GL
11
37
5
a
b
c
d
e
f
g
h
i
j

Figure 2.15: ***Clonal history reconstructed from ten spatially separated samples.*** *Ten ovarian cancer tumor samples from different regions (4a-4e, right ovary; 4f-4i, left ovary; 4j, left fallopian tube) from case 4 in Bashashati* et al. *(55) are whole exome sequenced. 63 mutations are confirmed by deep amplicon resequencing. (a) Heatmap of mutational profiling across 63 genes, 10 samples. (b) Anatomical sites of the ten spatially separated samples. (c) BIC as a model selection metric to determine the number of subclones. (d) The most likely tree returned by Canopy based on the mutational profiling. Mutations in blue are additionally acquired by the right ovary samples from the left ovary samples and drive the divergence. Mutations in red further distinguish case4a from the rest of the samples from the right ovary. Each sample offers a snapshot of different combinations of the subclones that is correlated with their spatial distribution. (e) Tree reconstructed by Bashashati* et al. *(55) by a nearest neighbor method. (f) Tree reconstructed by Popic* et al. *(70) as an acyclic directed graph. Both methods put samples at the tree leaves as homogeneous populations.*

Literature	Cancer type	Sequencing	Number of tumor samples from the same individual	Number of clones
Navin *et al.*, 2011 (10)	Breast cancer	Single-cell	100 single cells	5
Ding *et al.*, 2012 (3)	Acute myeloid leukaemia	Whole-genome	2 bulk samples	2-5
Bashashati *et al.*, 2013 (6)	Ovarian cancer	Whole-exome + deep amplicon	10 bulk samples	NA*
Gerlinger *et al.*, 2014 (11)	Clear cell renal cell carcinoma	Whole-exome + ultra deep	5-10 bulk samples	NA*
Zare *et al.*, 2014 (9)	Breast cancer	Whole-exome + targeted	12 bulk samples	4-6
Eirew *et al.*, 2015 (12)	Breast cancer xenografts	Whole-genome + targeted + single-cell	11 bulk samples and 90 single cells (SA501)	5
Sottoriva *et al.*, 2015 (13)	Colorectal adenomas and carcinomas	Whole-exome + targeted; single-cell FISH	On average 23 tumor glands and 2 bulk samples	2-7
Boutros *et al.*, 2015 (14)	Prostate cancer	Whole-genome	4 bulk samples (CPCG0183)	<6

Table 2.1: ***Cancer genomic studies by sequencing multiple samples from the same patients.*** Multiple types of cancer were sequenced by different platforms by a longitudinal (multi-time point) or a spatial experimental (multi-region) design. For bulk-tissue sequencing, 5 to 12 samples were sequenced from the same individual; for single-cell sequencing, ~100 single cells were sequenced. Across all studies, less than 8 cancer clones were identified. NA*: Bashashati *et al.* (55) and Gerlinger *et al.* (56) constructed phylogenetic tree by neighbour-joining and maximum parsimony method and put bulk-tumor samples as tree leaves.

Property/assumption	TITAN THetA	SciClone Clomial	PyClone	LICHeE	SCHISM PhyloSub BitPhylogeny	CHAT	PhyloWGS SPRUCE	Canopy
Takes raw SNAs calls as input	N	Y	Y	Y	Y	Y	Y	Y
Takes raw copy number estimates as input	Y	N	Y	N	Y	Y	N	Y
Allows CNAs to be subclonal	Y	N	N	N	N	Y	Y	Y
Resolves SNA and CNA that overlap	N	N	N	N	N	Y	Y	Y
Resolves overlapping CNAs with different endpoints	N	N	N	N	N	N	N	Y
Pools information across samples	N	Y	Y	Y	Y	N	Y	Y
Pools information across sites	Y	Y	Y	Y	Y	N	Y	Y
Reconstructs phylogeny	N	N	N	Y	Y	N	Y	Y
Quantifies uncertainty in phylogeny	N	N	N/A	N	Y	N/A	Y	Y

Table 2.2: ***Properties and assumptions of cancer clonal phylogeny reconstruction methods.*** *Y, yes; N, no; N/A, not applicable.*

Number of mutations N	Number of branches K	Run time (C, sec)	Run time (NC, sec)	Z error (C)	Z error (NC)	P error (C)	P error (NC)
25	3	84.1	57.0	0	0	0.003	0.003
	4	124.6	87.3	0.006	0.003	0.008	0.005
	5	142.0	126.0	0.022	0.019	0.01	0.008
	6	180.1	143.2	0.025	0.023	0.012	0.011
50	3	145.6	134.8	0	0	0.003	0.003
	4	235.3	295.6	0.013	0.009	0.013	0.007
	5	191.1	360.4	0.015	0.02	0.009	0.013
	6	261.2	429.5	0.019	0.019	0.013	0.009
100	3	348.5	436.5	0.003	0.005	0.005	0.005
	4	374.7	911.2	0.012	0.012	0.016	0.009
	5	334.9	1011.7	0.011	0.016	0.009	0.011
	6	372.6	1191.3	0.012	0.016	0.008	0.011
200	3	498.2	1463.6	0.002	0.007	0.007	0.011
	4	512.6	2454.9	0.008	0.012	0.017	0.015
	5	558.0	2871.0	0.009	0.014	0.010	0.017
	6	643.1	3580.9	0.010	0.013	0.014	0.014

Table 2.3: ***Running time and estimation error with and without pre-clustering step.*** *Simulation is carried out with varying number of mutations* $N \in \{25,50,100,200\}$ *along trees with different number of branches* $K \in \{3,4,5,6\}$ *from three samples. Canopy is run with and without a Binomial clustering procedure (C for clustering and NC for non-clustering) as an initialization step for MCMC. Convergence is measured by both the log-likelihood and the acceptance rate. Run time is measured in seconds; estimation error of the genotyping matrix* Z *is measured as the percentage of wrongly labeled elements; RMSE is used to measure the estimation error of the clonal proportion matrix* P*. Pre-clustering step significantly reduces computation time for larger number of mutations and results in comparable or smaller estimation errors.*

	Cell type	Metastatic outcome
MDA-MB-231	Parental line	-
1833	Mixed-cell subline (MCP)	Bone
2287	Mixed-cell subline (MCP)	Bone
SCP2	Single-cell subline (SCP)	Bone
SCP46	Single-cell subline (SCP)	Bone
1834	Mixed-cell subline (MCP)	Lung
3481	Mixed-cell subline (MCP)	Lung
SCP3	Single-cell subline (SCP)	Lung
SCP43	Single-cell subline (SCP)	Lung

Table 2.4: ***Metastatic outcomes and cell population types of MDA-MB-231 and its sublines.***

CHAPTER 3

MODELING ALLELE-SPECIFIC GENE EXPRESSION BY SINGLE-CELL RNA SEQUENCING

3.1 Introduction

In diploid organisms, two copies of each autosomal gene are available for transcription, and differences in gene expression level between the two alleles are widespread in tissues (92-98). Allele-specific expression (ASE), in its extreme, is found in genomic imprinting, where the allele from one parent is uniformly silenced across cells, and in random X-chromosome inactivation, where one of the two X-chromosomes in females is randomly silenced. During the last decade, using single-nucleotide polymorphism (SNP)-sensitive microarrays and bulk RNA sequencing (RNA-seq), more subtle expression differences between the two alleles were found, mostly in the form of allelic imbalance of varying magnitudes in mean expression across cells (99-102). In some cases such expression differences between alleles can lead to phenotypic consequences and result in disease (94, 103-105). These studies, though revelatory, were at the bulk tissue level, where one could only observe average expression across a possibly heterogeneous mixture of cells.

Recent developments in single-cell RNA sequencing (scRNA-seq) have made possible the better characterization of the nature of allelic differences in gene expression across individual cells (97, 106, 107). For example, recent scRNA-seq studies estimated that 12-24% of the expressed genes are monoallelically expressed during mouse preimplantation development (93) and that 76.4% of the heterozygous loci across all cells express only one allele (108). These ongoing efforts have improved our understanding of gene regulation and enriched our vocabulary in describing gene expression at the allelic level with single-cell resolution.

Despite this rapid progress, much of the potential offered by scRNA-seq data remains untapped. ASE, in the setting of bulk RNA-seq data, is usually quantified by comparing the mean expression level of the two alleles. However, due to the inherent stochasticity of gene expression across cells, the characterization of ASE using scRNA-seq data should look beyond mean

expression. A fundamental property of gene expression is transcriptional bursting, in which transcription from DNA to RNA occurs in bursts, depending on whether the gene's promoter is activated (Figure 3.1A) (109, 110). Transcriptional bursting is a widespread phenomenon that has been observed across many species including bacteria (111), yeast (112), Drosophila embryos (113), and mammalian cells (114, 115), and is one of the primary sources of expression variability in single cells. Figure 3.1B illustrates the expression across time of the two alleles of a gene. Under the assumption of ergodicity, each cell in a scRNA-seq sample pool is at a different time in this process, implying that for each allele, some cells might be in the transcriptional "ON" state, whereas other cells are in the "OFF" state. While in the "ON" state, the magnitude and length of the burst can also vary across cells, further complicating analysis. For each expressed heterozygous site, a scRNA-seq experiment gives us the bivariate distribution of the expression of its two alleles across cells, allowing us to compare the alleles not only in their mean, but also in their distribution. In this paper, we will use scRNA-seq data to characterize transcriptional bursting in an allele-specific manner and detect genes with allelic differences in the parameters of this process.

Kim and Marioni (116) first studied bursting kinetics of stochastic gene expression from scRNA-seq data, using a Beta-Poisson model and estimated the kinetic parameters via a Gibbs sampler. In this early attempt, they assumed shared bursting kinetics between the two alleles and modeled total expression of a gene instead of allele-specific expression. Current scRNA-seq protocols often introduce substantial technical noise (Figure 3.2) (117-121), and these noise (e.g., gene dropouts, amplification and sequencing bias) are largely ignored in Kim and Marioni (116) and another recent scRNA-seq study Borel et al. (108), where, in particular, gene dropout may have led to overestimation of the pervasiveness of monoallelic expression (ME). Realizing this, Kim et al. (122) incorporated measurements of technical noise from external spike-in molecules into the identification of stochastic ASE (defined as excessive variability in allelic ratios among cells), and concluded that more than 80% of stochastic ASE in mouse embryonic stem cells are

due to scRNA-seq technical noise. Kim et al.'s analysis was restricted to the identification of random monoallelic expression (RME) and did not consider more general patterns of ASE such as allele-specific transcriptional bursting.

ScRNA-seq also enables us to quantify the degree of dependence between the expressions of the two alleles. A previous RNA fluorescence *in situ* hybridization (FISH) experiment fluorescently labeled 20 genes in an allele-specific manner and showed that there was no significant deviation from independent bursting between the two alleles (123). A recent scRNA-seq study of mouse cells through embryonic development (93) produced similar conclusions on the genome-wide level: They modeled transcript loss by splitting each cell's lysate into two fractions of equal volume and controlling for false discoveries by diluting bulk RNA down to single-cell level. Their results suggest that on the genome-wide scale, assuming both alleles share the same bursting kinetics, the two alleles of most genes burst independently. Deviation from the theoretical curve in Deng et al. (93) for independent bursting with shared allele-specific kinetics, however, can be due to not only dependent bursting, but also differential bursting kinetics.

In this paper, we develop SCALE (**S**ingle-**C**ell **AL**lelic **E**xpression) (124), a systematic statistical framework to study ASE in single cells by examining allele-specific transcriptional bursting kinetics. Our main goal is to detect and characterize differences between the two alleles in their expression distribution across cells. As a by-product, we will also quantify the degree of dependence between the expressions of the two alleles. SCALE is comprised of three steps. First, an empirical Bayes method determines, for each *gene*, whether it is silent, monoallelically expressed, or biallelically expressed, based on its allele-specific counts across cells (Figure 3.1C). Next, for genes determined to be biallelic bursty (i.e., both alleles have zero expression level in some but not all cells), a Poisson-Beta hierarchical model is used to estimate allele-specific transcriptional kinetics while accounting for technical noise and cell size differences. Finally, resampling-based testing procedures are developed to detect allelic differences in

transcriptional burst size or burst frequency, and identify genes whose alleles exhibit non-independent transcription.

In silico simulations are conducted to investigate estimation accuracy and testing power. The stringency of model assumptions, and the robustness of the proposed procedures to the violation of these assumptions, will be discussed as they are introduced. Using SCALE, we re-analyze the scRNA-seq data for 122 mouse blastocyst cells (93) and 104 human fibroblast cells (108). The mouse blastocyst study initially found abundant RME generated by independent and stochastic allelic transcription (93); the human fibroblast study reported that 76.4% of the heterozygous loci displayed patterns of ME (108). Through proper modeling of technical noise, our re-analysis of these two datasets brings forth new insights: While for 90% of the bursty genes, there are no significant deviations from the assumption of independent allelic bursting and shared bursting kinetics, the remaining bursty genes show differential burst frequency by a *cis*-effect and/or non-independent bursting with an enrichment in coordinated bursting. Collectively, we present a genome-wide approach to systematically analyze expression variation in an allele-specific manner with single-cell resolution. SCALE is an open-source R package available at https://github.com/yuchaojiang/SCALE.

3.2 Results

Here we propose SCALE, a statistical framework for systematic characterization of ASE using data generated from scRNA-seq experiments. Our approach allows us to profile allele-specific bursting kinetics while accounting for technical variability and cell size difference. For genes that are classified as biallelic bursty through a Bayes categorization framework, we further examine whether transcription of the paternal and maternal alleles are independent, and whether there are any kinetic differences, as represented by bursty frequency and burst size, between the two alleles. Our results on the re-analysis of Deng et al. (93) and Borel et al. (108) provide insights into the extent of differences, coordination, and repulsion between alleles in transcriptional bursting.

Figure 3.3 shows an overview of the analysis pipeline of SCALE. We start with allele-specific read counts of endogenous RNAs across all profiled single cells. An empirical Bayes method is adopted to classify expression of genes into monoallelic, biallelic, and silent states based on ASE data across cells. SCALE then estimates allele-specific transcriptional bursting parameters via a hierarchical Poisson-Beta model, while adjusting for technical variabilities and cell size differences. Statistical testing procedures are then performed to identify genes whose two alleles have different bursting parameters or burst non-independently. We describe each of these steps in turn.

3.2.1 Gene Classification by ASE Data across Cells

SCALE first determines for each gene whether its expression is silent, paternal/maternal monoallelic, or biallelic. Figure 3.1C outlines this categorization scheme. Briefly, for each gene, each cell is assigned to one of four categories corresponding to scenarios where both alleles are off ($\emptyset$), only A allele is expressed (A), only B allele is expressed (B), and both alleles are expressed (AB). An expectation-maximization (EM) algorithm is implemented for parameter estimation. This classification accounts for both sequencing depth variation and sequencing errors. The assignment of the *gene* is then determined based on the posterior assignments of all cells. For example, if all cells are assigned to $\{\emptyset\}$, the gene is silent; if all cells are assigned to either $\{\emptyset\}$ or $\{A\}$, the gene has ME of the A allele; if all cells are assigned to either $\{\emptyset\}$ or $\{B\}$, the gene has ME of the B allele; if both A and B allele are expressed in the cell pool, then the gene is biallelically expressed. Refer to 3.4.2 Empirical Bayes Method for Gene Categorization for detailed statistical method and the EM algorithm.

Through simulation studies (under section 3.2.8 Assessment of estimation accuracy and testing power), we show that bursting parameters can only be stably estimated for *bursty* genes, that is, genes that are silent in a non-zero proportion of cells. Therefore, for biallelic bursty genes, allele-specific transcriptional kinetics are modeled through a Poisson-Beta distribution with

adjustment of technical noise. For silent, monoallelically expressed, or constitutively expressed genes, there is no way nor need to estimate bursting kinetics for both alleles.

3.2.2 Allele-Specific Transcriptional Bursting

When studying ASE in single cells, it is critical to consider transcriptional bursting due to its pervasiveness in various organisms (111-115). We adopt a Poisson-Beta hierarchical model to quantify allele-specific transcriptional kinetics while accounting for dropout events and amplification and sequencing bias. Here, we start by reviewing the relevant literature with regard to transcriptional bursting at the single-cell level.

A two-state model for gene transcription is shown in Figure 3.1A, where genes switch between the "ON" and "OFF" states with activation and deactivation rates k_{on} and k_{off}. When the gene is at the "ON" state, DNA is transcribed into RNA at rate s while RNA decays at rate d. A Poisson-Beta stochastic model was firstly proposed by Kepler and Elston (125):

$$Y \sim \text{Poisson}(sp),$$

$$p \sim \text{Beta}(k_{on}, k_{off}),$$

where Y is the number of mRNA molecules and p is the fraction of time that the gene spends in the active state, the latter having mean $k_{on}/(k_{on}+k_{off})$. Under this model, $1/k_{on}$ and $1/k_{off}$ are the average waiting times in the inactive and active states, respectively. *Burst size,* defined as the average number of synthesized mRNA per burst episode, is given by s/k_{off}, and *burst frequency* is given by k_{on}. Kepler and Elston (125) gave detailed analytic solutions via differential equations. Raj et al. (114) offered empirical support for this model via single-molecule FISH experiment on reporter genes. Since the kinetic parameters are measured in units of time and only the stationary distribution is assumed to be observed (e.g., when cells are killed for sequencing and fixed for FISH experiment), the rate of decay d is set to one (106). This is equivalent to having three kinetic parameters $\{s, k_{on}, k_{off}\}$, each normalized by the decay rate d. Kim and Marioni (116) applied this Poisson-Beta model to total gene-level transcript counts from scRNA-seq data of mouse embryonic stem cells. While they found that the inferred kinetic

parameters are correlated with RNA polymerase II occupancy and histone modification (116), they didn't address the issue of technical noise, especially the dropout events, introduced by scRNA-seq. Failure of accounting for gene dropouts may lead to biased estimation of bursting kinetics.

Furthermore, since the transitions between active and inactive states occur separately for the two alleles, when allele-specific expression data are available, it seems more appropriate to model transcriptional bursting in an allele-specific manner. The fact that transcriptional bursting occurs independently for the two alleles has been supported by empirical evidence: Case studies based on imaging methods have suggested that the two alleles of genes are transcribed in an independent fashion (126, 127); using scRNA-seq data, Deng et al. (93) showed that the two alleles of most genes tend to fire independently with the assumption that both alleles share the same set of kinetic parameters. These findings, although limited in scale or relying on strong assumptions, emphasize the need to study transcriptional bursting in an allele-specific manner.

3.2.3 Technical Noise in scRNA-seq and Other Complicating Factors

Figure 3.2 outlines the major steps of the scRNA-seq protocols and the sources of bias that are introduced during library preparation and sequencing. After the cells are captured and lysed, exogenous spike-ins are added as internal controls, which have fixed and known concentration and can thus be used to convert the number of sequenced transcripts into actual abundances. During the reverse transcription, pre-amplification, and library preparation steps, lowly expressed transcripts might be lost, in which case they will not be detected during sequencing. This leads to the so-called "dropout" events. Since spike-ins undergo the same experimental procedure as endogenous RNAs in a cell, amplification and sequencing bias can be captured and estimated through the spike-in molecules. Here we adopt the statistical model in TASC (Toolkit for Analysis of Single Cell data, unpublished), which explicitly models the technical noise through spike-ins. TASC's model is based on the key observation that the probability of a gene being a "dropout" depends on its true expression in the cell, with lowly expressed gene more likely to drop out.

Specifically, let Q_{cg} and Y_{cg} be, respectively, the observed and true expression level of gene g in cell c. The hierarchical mixture model used to model dropout, amplification and sequencing bias is:

$$Q_{cg} \sim Z_{cg}\text{Poisson}\left(\alpha_c\left(Y_{cg}\right)^{\beta_c}\right),$$

$$Z_{cg} \sim \text{Bernoulli}\left(\pi_{cg}\right),$$

$$\pi_{cg} = \text{expit}\left(\kappa_c + \tau_c \log(Y_{cg})\right),$$

where Z_{cg} is a Bernoulli random variable indicating that gene g is detected in cell c, that is, a dropout event has not occurred. The success probability $\pi_{cg} = P(Z_{cg} = 1)$ depends on $\log(Y_{cg})$, the logarithm of the true underlying expression. Cell-specific parameters α_c models the capture and sequencing efficiency; β_c models the amplification bias; κ_c and τ_c characterize whether a transcript is successfully captured in the library. This model will later be used to adjust for technical noise in allele-specific expression.

As input to SCALE, we recommend scRNA-seq data from cells of the same type. Unwanted heterogeneity, however, still persists as the cells may differ in size or may be in different phases of the cell cycle. Through a series of single-cell FISH experiments, Padovan-Merhar et al. (128) showed how gene transcription depends on these exogenous factors: burst size is independent of cell cycle but is kept proportional to cell size by a *trans* mechanism; burst frequency is independent of cell size but is reduced approximately by half, through a *cis* mechanism, between G1 and G2 phase to compensate for the doubling of DNA content. Figure 3.4 gives an illustration on how burst size and burst frequency change with cell size and cell cycle phase. Note that, while the burst frequency from *each* DNA copy is halved when the amount of DNA is doubled, the total burst frequency remains roughly constant through the cell cycle. Thus, SCALE adjusts for variation in cell size through modulation of burst size, and does not adjust for variation in cell cycle phase. Details will be given below.

There are multiple ways to measure cell size. Padovan-Merhar et al. (128) proposed using the expression level of *GAPDH* as a cell size marker. When spike-ins are available, we use the ratio

of the total number of endogenous RNA reads over the total number of spike-in reads as a measure (Figure 3.4) of the total RNA volume, which was shown to be a good proxy for cell size (119). SCALE allows the user to input the cell sizes ϕ_c, if these are available through other means.

3.2.4 Modeling Transcriptional Bursting with Adjustment of Technical and Cell-Size Variation

We are now ready to formulate the allele-specific bursting model for scRNA-seq data. For genes that are categorized as biallelic bursty (with proportion of cells expressing each allele between 5% and 95% from the Bayes framework), SCALE proceeds to estimate the allele-specific bursting parameters using a hierarchical model:

$$Y_{cg}^A \sim \text{Poisson}\left(\phi_c s_g^A p_{cg}^A\right) \qquad Y_{cg}^B \sim \text{Poisson}\left(\phi_c s_g^B p_{cg}^B\right)$$

$$p_{cg}^A \sim \text{Beta}\left(k_{on,g}^A, k_{off,g}^A\right) \qquad p_{cg}^B \sim \text{Beta}\left(k_{on,g}^B, k_{off,g}^B\right),$$

where Y_{cg}^A and Y_{cg}^B are the true allele-specific expressions for gene g in cell c. The two alleles of each gene are modeled by separate Poisson-Beta distributions with kinetic parameters that are gene- and allele-specific. These two Poisson-Beta distributions share the same cell size factor ϕ_c, which affects burst size. The true allele-specific expressions Y_{cg}^A and Y_{cg}^B are not directly observable. The observed allele-specific read counts Q_{cg}^A and Q_{cg}^B are confounded with technical noise, and follow the Poisson mixture model outlined in the previous section:

$$Q_{cg}^A \sim Z_{cg}^A \text{Poisson}\left(\alpha_c \left(Y_{cg}^A\right)^{\beta_c}\right) \qquad Q_{cg}^B \sim Z_{cg}^B \text{Poisson}\left(\alpha_c \left(Y_{cg}^B\right)^{\beta_c}\right)$$

$$Z_{cg}^A \sim \text{Bernoulli}\left(\pi_{cg}^A\right) \qquad Z_{cg}^B \sim \text{Bernoulli}\left(\pi_{cg}^B\right)$$

$$\pi_{cg}^A = \text{expit}\left(\kappa_c + \tau_c \log\left(Y_{cg}^A\right)\right) \qquad \pi_{cg}^B = \text{expit}\left(\kappa_c + \tau_c \log\left(Y_{cg}^B\right)\right).$$

How to generate input to SCALE for both endogenous RNAs and exogenous spike-ins is included in 3.4.1 Input for Endogenous RNAs and Exogenous Spike-ins. For parameter estimation, we developed a new "histogram-repiling" method to obtain the distribution of Y_{cg} from the observed distribution of Q_{cg}. The bursting parameters are then derived from the distribution of

Y_{cg} by moment estimators. Standard errors and confidence intervals of the parameters are obtained using nonparametric bootstrap.

3.2.5 Hypothesis Testing

For biallelic bursty genes, we use nonparametric Bootstrap to test the null hypothesis that the burst frequency and burst size of the two alleles are the same ($k_{on}^{A} = k_{on}^{B},\ s^{A}/k_{off}^{A} = s^{B}/k_{off}^{B}$) against the alternative hypothesis that either or both parameters differ between alleles. For each gene, we also perform chi-square test to determine if the transcription of the two alleles are independent by comparing the observed proportions of cells from the gene categorization framework against the expected proportions under independence. For genes where the proportion of cells expressing both alleles is significantly higher than expected, we define their bursting as coordinated; for genes where the proportion of cells expressing only one allele is significantly higher than expected, we define their bursting as repulsed (Figure 3.3). We adopt false discovery rate (FDR) to adjust for multiple comparisons. Details of the testing procedures are outlined in 3.4.4 Hypothesis Testing Framework.

3.2.6 Analysis of scRNA-seq Dataset of Mouse Cells during Preimplantation Development

We re-analyze the scRNA-seq dataset of mouse blastocyst cells dissociated from *in vivo* F1 embryos (CAST/female x C57/male) from Deng et al. (93). Transcriptomic profiles of each individual cell was generated using the Smart-seq (129) protocol. For 22,958 genes, reads per kilo base per million reads (RPKM) and total number of read counts across all cells are available. Parental allele-specific read counts are also available at heterozygous loci. Principal component analysis (PCA) was performed on cells from oocyte to blastocyst stages of mouse preimplantation development and showed that the first three principal components well separate the early-stage cells from the blastocyst cells. The cluster of early-, mid-, and late-blastocyst cells are combined to gain sufficient sample size. In discussion, we give further insights on the potential effects of cell subtype confounding. Quality control (QC) procedure was adopted to remove outliers in library size, mean and standard deviation of allelic read counts/proportions. We

apply SCALE to this dataset of 122 mouse blastocyst cells, with a focus on addressing the issue of technical variability and modeling of transcriptional bursting.

Eight exogenous RNAs with known serial dilutions are added in late blastocyst cells and are used to estimate the technical-noise associated parameters (Figure 3.5A). We apply the Bayes gene classification framework to these cells to get the genome-wide distribution of gene categories. Specifically, out of the 22,958 genes profiled across all cells, ~43% are biallelically expressed (~33% of the total are biallelic bursty and ~10% of the total are biallelic non-bursty), ~7% are monoallelically expressed, and ~50% are silent. Our empirical Bayes categorization results show that, on the genome-wide scale, the two alleles of most biallelic bursty genes share the same bursting kinetics and burst independently (Figure 3.6A), as has been reported by Deng et al. (93).

For the 7,486 genes that are categorized as biallelic bursty, we apply SCALE to identify genes whose alleles have different bursting kinetic parameters by the Bootstrap-based hypothesis tests as previously described. After FDR control, we identify 425 genes whose two alleles have significant differential burst frequency (Figure 3.7A) and 2 genes whose two alleles have significant differential burst size (Figure 3.7B). Figure 3.8 shows the allelic read counts of a gene that has differential burst frequency (*Btf3l4*) and a gene that has differential burst size (*Fdps*). The two genes with significant differential allelic burst size, namely, gene *Fdps* and *Atp6ap2*, are also significant in having differential burst frequency between the two alleles. P-values from differential burst frequency testing have a spike below the significance level after FDR control (Figure 3.7A), while those from differential burst size testing are roughly uniformly distributed (Figure 3.7B).

At the whole genome level, these results show that allelic differences in the expression of bursty genes during embryo development is achieved through differential modulation of burst frequency rather than burst size. This seems to agree with intuition, since allelic differences must be caused by factors that act in *cis* to regulate gene expression, and *cis* factors are likely to

change burst frequency by affecting promoter accessibility (128, 130-132). On the contrary, while it is plausible for *cis* factors to affect allelic burst size through, for example, the efficiency of RNA Polymerase II recruitment or the speed of elongation, the few known cases of burst size modulation are controlled in *trans* (128). Furthermore, previous studies have shown that the kinetic parameter that varies the most – along the cell cycle (128), between different genes (133), between different growth conditions (134), or under regulation by a transcription factor (135) – is the probabilistic rate of switching to the active state k_{on}, while the rates of gene inactivation k_{off} and of transcription s vary much less.

Our analysis includes 107 male cells (X^AY) and 15 female cells (X^AX^B) and this allows us to use those bursty X-chromosome genes as positive controls. As a result of this gender mixture, there are more cells expressing the maternal X^A allele compared to the paternal X^B allele. As shown in Figure 3.7, SCALE successfully detects these bursty X-chromosome genes with significant difference in allelic burst frequency but not in allelic burst size. If we only keep the 107 male cells, these X-chromosome genes are correctly categorized as monoallelically expressed – the bursting kinetics for the paternal X^B allele are not estimable – and in this case there is no longer a cluster of significant X-chromosome genes separated from the autosomal genes (Figure 3.9).

For biallelic bursty genes, we also used a simple Binomial test to determine if the mean allelic coverage across cells is biased towards either allele. This is comparable to existing tests of allelic imbalance in bulk tissue, although the total coverage across cells in this dataset is much higher than standard bulk tissue RNA-seq data. After multiple hypothesis testing correction, we identify 417 genes with significant allelic imbalance, out of which 238 overlap with the significant genes from the testing of differential bursting kinetics (Figure 3.10A). Inspection of the estimated bursting kinetic parameters in Figure 3.10A shows that, when the burst size and burst frequency of the two alleles change in the same direction (e.g., gene *Gprc5a* in Figure 3.10B), testing of allelic imbalance can detect more significant genes with higher power. This is not unexpected – a

small insignificant increase in burst size adds on top of an insignificant increase in burst frequency resulting in a significant increase in overall expression levels between the two alleles. However, for genes in red in the top left and bottom right quadrants of Figure 3.10A, the test for differential bursting kinetics detects more genes than the allelic imbalance test. This is due to the fact that when burst size and burst frequency change in opposite directions (e.g., gene *Dhrs7* in Figure 3.10B), their effects cancel out when looking at the mean expression. Furthermore, even when the burst size does not change, if the change in burst frequency is small, by using a more specific model SCALE has higher power to detect it as compared to an analysis based on mean allelic imbalance. Overall, the allelic imbalance test and differential bursting test report overlapping but substantially different set of genes, with each test having its benefits. Compared to the allelic imbalance test, SCALE gives more detailed characterization of the nature of the difference by attributing the change in mean expression to a change in the burst frequency and/or burst size.

It is also noticeable that in Figure 3.10A the vertical axis, $\Delta freq$, has a 50% wider range than the horizontal axis, $\Delta size$. Therefore, while it is visually not obvious from this scatter plot, there are much more genes with large absolute $\Delta freq$ than with large absolute $\Delta size$. Although the standard errors of these estimated differences are not reflected in the plot, given our testing results, those genes with large estimated differences in $\Delta size$ also have large standard errors in their estimates, which is further confirmed via simulations.

Further chi-squared test of the null hypothesis of independence (Figure 3.8C) shows that there are 424 genes whose two alleles fire in a significantly non-independent fashion. We find that all significant genes have higher proportions of cells expressing both alleles than expected, indicating coordinated expression between the two alleles. In this dataset, there are no significant genes with repulsed bursting between the two alleles. Repulsed bursting, in the extreme case where at most one allele is expressed in any cell, is also referred to as stochastic ME (122). Our testing results indicate that, in mouse embryo development, all cases of stochastic ME (i.e.,

repulsion between the two alleles) can be explained by independent and infrequent stochastic bursting. The burst synchronization in the 424 significant genes is not unexpected and is possibly due to a shared *trans* factor between the two alleles (e.g., co-activation of both alleles by a shared enhancer). This result is concordant with the findings from a mouse embryonic stem cell scRNA-seq study by Kim et al. (122), which reported that the two alleles of a gene show correlated allelic expression across cells more often than expected by chance, potentially suggesting regulation by extrinsic factors (122). We further discuss the sharing of such extrinsic factors under the context of cell population admixtures in Discussion.

In summary, our results by SCALE suggest that: (i) The two alleles from 10% of the bursty genes show either significant deviations from independent firing or significant differences in bursting kinetic parameters, (ii) For genes whose alleles differ in their bursting kinetic parameters, the difference is found mostly in the burst frequency instead of the burst size, (iii) For genes whose alleles violate independence, their expression tends to be coordinated.

3.2.7 Analysis of scRNA-seq Dataset of Human Fibroblast Cells

To further examine our findings in a dataset without potential confounding of cell type admixtures, we apply SCALE to a scRNA-seq dataset of 104 cells from female human newborn primary fibroblast culture from Borel et al. (108). The cells were captured by Fluidigm C1 with 22 PCR cycles and were sequenced with on average 36 million reads (100 bp, paired end) per cell. Bulk-tissue whole genome sequencing was performed on two different lanes with 26-fold coverage on average and was used to identify heterozygous loci in coding regions. After QC procedures, 9016 heterozygous loci from 9016 genes were identified (if multiple loci coexist in the same gene, we pick the one with the highest mean depth of coverage). At each locus, we use SAMtools (36) mpileup to obtain allelic read counts in each single cell from scRNA-seq, which are further used as input for SCALE. 92 ERCC synthesized RNAs were added in the lysis buffer of 12 fibroblast cells with a final dilution of 1:40000. The true concentrations and the observed number of reads for all spike-ins are used as baselines to estimate technical variability (Figure 3.5B).

We apply the gene categorization framework by SCALE and find that out of the 9016 genes, the proportions of monoallelically expressed, biallelically expressed, and silent genes are 11.5%, 45.7%, and 42.8%, respectively. For the 2277 genes that are categorized as biallelic bursty, we estimate their allele-specific bursting kinetic parameters and find that the correlations between the estimated burst frequency and burst size between the two alleles are 0.859 and 0.692 (Figure 3.11). We then carry out hypothesis testing on differential allelic bursting kinetics. After FDR correction, we identified 26 genes with significant differential burst frequency between the two alleles (Figure 3.11A) and one gene *Nfx1* with significantly differential burst size between the two alleles, which is also significant in burst frequency testing (Figure 3.11B). We further carry out testing of non-independent bursting between the two alleles and identify 35 significant genes after FDR correction (Figure 3.6B). Out of the 35 significant genes, 27 showed patterns of coordinated bursting while the rest 8 showed repulsed patterns.

3.2.8 Assessment of estimation accuracy and testing power

First, we investigate the accuracy of the moment estimators for the bursting parameters under four different scenarios in the Poisson-Beta transcription model: (i) small k_{on} and small k_{off} , which we call bursty and leads to relatively few transitions between the "ON" and "OFF" state with a bimodal mRNA distribution across cells (Figure 3.12A); (ii) large k_{on} and small k_{off}, which leads to long durations in the "ON" state and resembling constitutive expression with the mRNA having a Poisson-like distribution (Figure 3.12B); (iii) small k_{on} and large k_{off}, which leads to most cells being silent (Figure 3.12C); (iv) and large k_{on} and large k_{off}, which leads to constitutive expression (Figure 3.12D).

We generate simulated data for 100 cells from the four cases above and start with no technical noise or cell size confounding. Within each case, we vary k_{on}, k_{off}, and s and use relative absolute error $\left|\hat{\theta} - \theta\right|/\theta$ as a measurement of accuracy (Figure 3.13). Our results show that genes with large k_{on} and small k_{off} (shown as the black curves in Figure 3.13) have the largest estimation errors of the bursting parameters. Statistically it is hard to distinguish these

constitutively expressed genes from genes with large k_{on} and large k_{off} and thus the kinetic parameters in this case cannot be accurately estimated, which has been previously reported (116, 136). Furthermore, the estimation errors are large for genes with small k_{on}, large k_{off}, and small s (shown as red curves in Figure 3.13) due to lack of cells with nonzero expression. The standard errors and confidence intervals of the estimated kinetics from bootstrap resampling further confirm the underperformance for the above two classes (Table 3.1). This emphasizes the need to adopt the Bayes categorization framework as a first step so that kinetic parameters are stably estimated only for genes whose both alleles are bursty. For genes whose alleles are perpetually silent or constitutively expressed across cells, there is no good method, nor any need, to estimate their bursting parameters.

Importantly, we see that the estimation bias in transcription rate s and deactivation rate k_{off} cancel – over/under estimation of s is compensated by over/under estimation of k_{off} – and as a consequence the burst size s/k_{off} can be more stably estimated than either parameter alone, especially when $k_{on} \ll k_{off}$ (shown as red curves in Figure 3.13). This is further confirmed by empirical results that allelic burst size has much higher correlation (0.746 from the mouse blastocyst dataset and 0.692 from the human fibroblast dataset) than allelic transcription and deactivation rate (0.464 and 0.265 for mouse blastocyst, and 0.458 and 0.33 for human fibroblast) (Figure 3.14). For this reason, all of our results on real data are based on s/k_{off} and we do not consider s and k_{off} separately.

We further carry out power analysis on the testing of differential burst frequency and burst size between the two alleles. The null hypothesis is both alleles sharing the same bursting kinetics ($k_{on}^{A} = k_{on}^{B} = 0.2, k_{off}^{A} = k_{off}^{B} = 0.2, s^{A} = s^{B} = 50$), while the alternative hypotheses with differential burst frequency or burst size are shown in the legends in Figure 3.15. The detailed setup of the simulation procedures are as follows. (i) Simulate the true allele-specific read counts Y^{A} and Y^{B} across 100 cells from the Poisson-Beta model under the alternative hypothesis. Technical noise is then added based on the noise model described earlier with technical noise

parameters $\{\alpha, \beta, \kappa, \tau\}$ estimated from the mouse blastocyst cell dataset. (ii) Apply SCALE to the observed expression level Q^A and Q^B, which returns p-value for testing differential burst size or burst frequency. If the p-value is less than the significance level, we reject the null hypothesis. (iii) Repeat (i) and (ii) N times with the power estimated as $\frac{\text{Number of } p-\text{values} \leq 0.05}{N}$. Our results indicate that the testing of burst frequency and burst size have similar power overall with relatively reduced power if the difference in allelic burst size is due to difference in the deactivation rate k_{off}.

We then simulate allele-specific counts from the full model including technical noise as well as variations in cell size with the ground truth $k_{on}^A = k_{on}^B = k_{off}^A = k_{off}^B = 0.2, s^A = s^B = 100$ (bursty with small activation and deactivation rate). For parameters quantifying the degree of technical noise, we use the estimates from the mouse blastocyst cells (Figure 3.5A) as well as the human fibroblast cells (Figure 3.5B). Cell sizes are simulated from a normal distribution with mean 0 and standard deviation 0.1 and 0.01. We run SCALE under four different settings: (i) in its default setting, (ii) without accounting for cell size, (iii) without adjusting for technical variability, (iv) not in an allele-specific fashion but using total coverage as input. Each is repeated 5000 times with a sample size of 100 and 400 cells, respectively. Relative estimation errors of burst size and burst frequency are summarized across all simulation runs. Our results show that SCALE in its default setting has the smallest estimation errors for both burst size and burst frequency (Figure 3.16, Figure 3.17). Not surprisingly, cell size has larger effect on burst size estimation than burst frequency estimation, while technical variability leads to biased estimation of both burst frequency and burst size. The estimates taking total expression instead of ASE as input are completely off. Furthermore, the estimation accuracy improved as the number of cells increased. These results indicate the necessity to profile transcriptional kinetics in an allele-specific fashion with adjustment of technical variability and cell size.

3.3 Discussion

We propose SCALE, a statistical framework to study ASE using scRNA-seq data. The input data to SCALE are allele-specific read counts at heterozygous loci across all cells. In the two datasets that we analyzed, we use the F1 mouse crossing and the bulk-tissue sequencing to profile the true heterozygous loci. When these are not available, scRNA-seq itself can be used to retrieve allele-specific expression and more specifically haplotype, as illustrated in Edsgard et al. (137). SCALE estimates parameters that characterize allele-specific transcriptional bursting, after accounting for technical biases in scRNA-seq and size differences between cells. This allows us to detect genes that exhibit allelic differences in burst frequency and burst size, and genes whose alleles show coordinated or repulsed bursting patterns. Differences in mean expression between the two alleles have long been observed in bulk RNA-seq. By scRNA-seq, we now move beyond the mean and characterize the difference in expression distributions between the two alleles, specifically in terms of their transcriptional bursting parameters.

Transcriptional bursting is a fundamental property of gene expression, yet its global patterns in the genome has not been well characterized, and most studies consider bursting at the gene level by ignoring the allelic origin of transcription. In this paper, we reanalyzed the Deng et al. (93) and Borel et al. (108) data. We confirmed the findings from Levesque and Raj (123) and Deng et al. (93) that for most genes across the genome there is no sufficient evidence against the assumption of independent bursting with shared bursting kinetics between the two alleles. For genes where significant deviations are observed, SCALE allows us to attribute the deviation to differential bursting kinetics and/or non-independent bursting between the two alleles.

More specifically, for genes that are transcribed in a “bursty” fashion, we compared the burst frequency and burst size, between their two alleles. For both scRNA-seq datasets, we identify significant number of genes whose allele-specific burstings differ in the burst frequency but not in the burst size. Our findings provide evidence that burst frequency, which represents the rate of gene activation, is modified in *cis*, and that burst size, which represents the ratio of

transcription rate to gene inactivation rate, is less likely to be modulated in *cis*. Although our testing framework may have slightly reduced power in detecting differential deactivation rate (Figure 3.15), the regulation in burst size can either result from a global *trans* factor or extrinsic factors that acts upon both alleles. Similar findings have been previously reported, from different perspectives and on different scales, using various technologies, platforms, and model organisms (122, 128, 133-135).

It is worth noting that the estimated bursting parameters by SCALE are normalized by the decay rate, where the inverse $1/d$ denotes the average life time of an mRNA molecule. Here we implicitly make the assumptions that for each allele, the gene-specific decay rates (d_g^A and d_g^B) are constant, and thus the estimated allelic burst frequencies are the ratio of true burst frequency over decay rate (that is $k_{on,g}^A/d_g^A$ and $k_{on,g}^B/d_g^B$). The decay rates, however, cancel out in the numerator and denominator in the allelic burst sizes, $s_g^A/k_{off,g}^A$ and $s_g^B/k_{off,g}^B$. Therefore, the differences that we observe in the allelic burst frequencies can also potentially be due to differential decay rates between the two alleles, which has been previously reported to be regulated by microRNAs (138).

It is also important to note that 44% of the genes found to be significant for differential burst frequency are not significant in the allelic imbalance test based on mean expression across cells. This suggests that expression quantitative trait loci (eQTL) affecting gene expression through modulation of bursting kinetics is likely to escape detection in existing eQTL studies by bulk sequencing, especially when burst size and burst frequency change in different directions. This is further underscored by the study of Wills et al. (139), which measured the expression of 92 genes affected by Wnt signaling in 1,440 single cells from 15 individuals, and then correlated SNPs with various gene-expression phenotypes. They found bursting kinetics as characterized by burst size and burst frequency to be heritable, thus suggesting the existence of bursting-QTLs. Taken together, these results should further motivate more large scale genome-wide studies to systematically characterize the impact of eQTLs on various aspects of transcriptional bursting.

Kim et al. (122) described a statistical framework to quantify the extent of stochastic ASE in scRNA-seq data by using of spike-ins, where stochastic ASE is defined as excessive variability in the ratio of the expression level of the paternal (or maternal) allele between cells after controlling for mean allelic expression levels. While they attributed 18% of the stochastic ASE to biological variability, they did not examine what biological factors lead to these stochastic ASE. In this paper, we attribute the observed stochastic ASE to difference in allelic bursting kinetics. By studying bursting kinetics in an allele-specific manner, we can compare the transcriptional differences between the two alleles at a finer scale.

Kim and Marioni (116) described a procedure to estimate bursting kinetic parameters using scRNA-seq data. Our method differs from Kim and Marioni (116) in several ways. First, our model is an allele-specific model that infers kinetic parameters for each allele separately, thus allowing comparisons between alleles. Second, we infer kinetic parameters based on the distribution of "true expression" rather than the distribution of observed expression. We are able to do this through the use of a simple and novel deconvolution approach, which allows us to eliminate the impact of technical noise when making inference on the kinetic parameters. Appropriate modeling of technical noise, in particular, gene dropouts, is critical in this context, as failing to do so could lead to the overestimation of k_{off}. Third, we employ a gene categorization procedure prior to fitting the bursting model. This is important because the bursting parameters can only be reliably estimated for genes that have sufficient expression and that are bursty.

As a by-product, SCALE also allows us to rigorously test, for scRNA-seq data, whether the paternal and maternal alleles of a gene are independently expressed. In both scRNA-seq datasets we analyzed, we identified more genes whose allele-specific burstings are in a coordinated fashion than those in a repulsed fashion. The tendency towards coordination is not surprising, since the two alleles of a gene share the same nuclear environment and thus the same ensemble of transcription factors. We are aware that this degree of coordination can also arise from the mixture of non-homogeneous cell populations, e.g., different lineages of cells

during mouse embryonic development, as we combine the early-, mid-, and late-blastocyst cells to gain a large enough sample size. While it is possible that this might lead to false positives in identifying coordinated bursting events, it will result in a decrease in power for the testing of differential bursting kinetics. Given the amount of stochasticity that is observed in the allele-specific expression data, how to define cell sub-types and how to quantify between-cell heterogeneity need further investigation.

3.4 Methods

3.4.1 Input for Endogenous RNAs and Exogenous Spike-ins

For endogenous RNAs, SCALE takes as input the observed allele-specific read counts at heterozygous locus Q_{cg}^{A} and Q_{cg}^{B}, with adjustment by library size factor:

$$\eta_c = \underset{g}{\text{median}} \frac{Q_{cg}^{A} + Q_{cg}^{B}}{\left[\prod_{c^*=1}^{C}\left(Q_{c^*g}^{A} + Q_{c^*g}^{B}\right)\right]^{1/C}}.$$

In addition, for spike-ins, SCALE takes as input the true concentrations of the spike-in molecules, the lengths of the molecules, as well as the depths of coverage for each spike-in sequence across all cells. The true concentration of each spike-in molecule is calculated according to the known concentration (denoted as C attomoles/uL) and the dilution factor (x40000):

$$\frac{C \times 10^{-18}\ \text{moles/uL} \times 6.02214 \times 10^{23}\text{mole}^{-1}\ (\text{Avogadro constant})}{40000\ (\text{dilution factor})}.$$

The observed number of reads for each spike-in is calculated by adjusting for the library size factor, the read length, and the length of the spike-in RNA. The bioinformatic pipeline to generate the input for SCALE can be found at https://github.com/yuchaojiang/SCALE.

3.4.2 Empirical Bayes Method for Gene Categorization

We propose an empirical Bayes method that categorizes gene expressions across cells into silent, monoallelic, biallelic states based on their ASE data. Without loss of generality, we focus on one gene here with the goal of determining the most likely gene category based on its ASE pattern. Let n_c^A and n_c^B be the allele-specific read counts in cell c for allele A and B, respectively. For *each* cell, there are four different categories based on its ASE – $\{\emptyset, A, B, AB\}$ corresponding to

scenarios where both alleles are off, only A allele is expressed, only B allele is expressed, and both alleles are expressed, respectively. Let $k \in \{1,2,3,4\}$ represent this cell-specific category. The log-likelihood for the gene across all cells can be written as:

$$\log\big(\mathcal{L}(\Theta|n^A, n^B)\big) = \log\prod_c f(n_c^A, n_c^B|\Theta) = \sum_c \log\Big[\sum_{k=1}^4 \varphi_k f_k(n_c^A, n_c^B|\epsilon, a, b)\Big],$$

where the parameters are $\Theta = \{\varphi_1, \dots, \varphi_4, \epsilon, a, b\}$ with $\sum_{k=1}^4 \varphi_k = 1$ and each f_k is a density function parameterized by ϵ, a, b. ϵ is the per-base sequencing error rate, and a and b are hyper-parameters for a Beta distribution, where $\theta_c \sim \mathrm{Beta}(a, b)$ corresponds to the relative expression of A allele when both alleles are expressed. It is easy to show that

$$f_1(n_c^A, n_c^B|\epsilon, a, b) \propto \epsilon^{n_c^A + n_c^B},$$
$$f_2(n_c^A, n_c^B|\epsilon, a, b) \propto (1-\epsilon)^{n_c^A}\epsilon^{n_c^B},$$
$$f_3(n_c^A, n_c^B|\epsilon, a, b) \propto \epsilon^{n_c^A}(1-\epsilon)^{n_c^B},$$
$$f_4(n_c^A, n_c^B|\epsilon, a, b) \propto \int_0^1 [\theta_c(1-\epsilon) + (1-\theta_c)\epsilon]^{n_c^A}[\theta_c\epsilon + (1-\theta_c)(1-\epsilon)]^{n_c^B} \frac{{\theta_c}^{a-1}(1-\theta_c)^{b-1}}{B(a,b)} d\theta_c.$$

ϵ can be estimated using sex chromosome mismatching or be prefixed at the default value, 0.001. We require $a = b \geq 3$ in the prior on θ_c so that the AB state is distinguishable from the A and B states. This is a reasonable assumption in that most genes have balanced ASE on average and the use of Beta distribution allows variability of allelic ratio across cells. We adopt an EM algorithm for estimation, with Z being the missing variables:

$$Z_{ck} = \begin{cases} 1 & \text{if cell } c \text{ belongs to category } k \\ 0 & \text{otherwise} \end{cases}.$$

The complete-data log-likelihood is given as

$$\log\big(\mathcal{L}(\Theta|n^A, n^B, Z)\big) = \log\Big[\sum_c \prod_{k=1}^4 f_k(n_c^A, n_c^B|\epsilon, a, b)^{Z_{ck}} {\varphi_k}^{Z_{ck}}\Big]$$
$$= \sum_c \sum_{k=1}^4 Z_{ck}\log(\varphi_k) + \sum_c \sum_{k=1}^4 Z_{ck}\log[f_k(n_c^A, n_c^B|\epsilon, a, b)].$$

For each cell, we assign the state that has the maximum posterior probability and only keep a cell if its maximum posterior probability is greater than 0.8. Let $N_\emptyset$, N_A, N_B, and N_{AB} be the number of cells in state $\{\emptyset\}$, $\{A\}$, $\{B\}$, and $\{AB\}$, respectively. We then assign a gene to be: (i) silent if $N_A =$

$N_B = N_{AB} = 0$; (ii) A-allele monoallelic if $N_A > 0, N_B = N_{AB} = 0$; (iii) B-allele monoallelic if $N_B > 0, N_A = N_{AB} = 0$; (iv) biallelic otherwise (more specifically, biallelic bursty if $0.05 \leq (N_A + N_{AB})/(N_\emptyset + N_A + N_B + N_{AB}) \leq 0.95$ and $0.05 \leq (N_B + N_{AB})/(N_\emptyset + N_A + N_B + N_{AB}) \leq 0.95$).

3.4.3 Parameter Estimation for Poisson-Beta Hierarchical Model

Since exogenous spike-ins are added in a fixed amount and don't undergo transcriptional bursting, they can be used to directly estimate the technical-variability-associated parameters $\{\alpha, \beta, \kappa, \tau\}$ that are shared across all cells from the same sequencing batch. Specifically, we use non-zero read counts to estimate α and β through log-linear regression:

$$Q_{cg} \sim \text{Poisson}\left(\alpha\left(Y_{cg}\right)^{\beta}\right),$$

where $Q_{cg} > 0$, capture and sequencing efficiencies are confounded in α and amplification bias is modeled by β. We then use the Nelder-Mead simplex algorithm to jointly optimize κ and τ, which models the probability of non-dropout, using the likelihood function:

$$\log\left(\mathcal{L}\left(\kappa, \tau \middle| Q, Y, \hat{\alpha}, \hat{\beta}\right)\right) = \prod_c \prod_g \log\left\{\text{pPoisson}\left(Q_{cg}, \hat{\alpha}\left(Y_{cg}\right)^{\hat{\beta}}\right)\text{expit}\left(\kappa + \tau \log Y_{cg}\right) + \left(1 - \text{expit}\left(\kappa + \tau \log Y_{cg}\right)\right)\mathbb{1}(Q_{cg} = 0)\right\},$$

where $\text{pPoisson}(x, y)$ specifies the Poisson likelihood of getting x from a Poisson distribution with mean y. This log-likelihood function together with the estimated parameters decomposes the zero read counts ($Q_{cg} = 0$) into being from the dropout events or from being sampled as zero from the Poisson sampling during sequencing.

The allele-specific kinetic parameters are estimated via the moment estimator methods, which is more computational efficient than the Gibbs sampler method adopted by Kim and Marioni (116). For each gene, the distribution moments of the A allele given true expression levels Y_c^A and Y_c^B are:

$$m_1^A \equiv \frac{E[\sum_c Y_c^A]}{\sum_c \phi_c} = \frac{k_{on}^A s^A}{k_{on}^A + k_{off}^A}$$

$$m_2^A \equiv \frac{E[\sum_c Y_c^A (Y_c^A - 1)]}{\sum_c \phi_c^2} = \frac{k_{on}^A (k_{on}^A + 1)(s^A)^2}{\left(k_{on}^A + k_{off}^A\right)\left(k_{on}^A + k_{off}^A + 1\right)}$$

$$m_3^A \equiv \frac{E[\sum_c Y_c^A (Y_c^A - 1)(Y_c^A - 2)]}{\sum_c \phi_c^3} = \frac{k_{on}^A (k_{on}^A + 1)(k_{on}^A + 2)(s^A)^3}{\left(k_{on}^A + k_{off}^A\right)\left(k_{on}^A + k_{off}^A + 1\right)\left(k_{on}^A + k_{off}^A + 2\right)}.$$

Solving this system of three equations, we have:

$$\hat{k}_{on}^A = \frac{-2(-m_1^A (m_2^A)^2 + (m_1^A)^2 m_3^A)}{-m_1^A (m_2^A)^2 + 2(m_1^A)^2 m_3^A - m_2^A m_3^A}$$

$$\hat{k}_{off}^A = \frac{2((m_1^A)^2 - m_2^A)(m_1^A m_2^A - m_3^A)(m_1^A m_3^A - (m_2^A)^2)}{((m_1^A)^2 m_2^A - 2(m_2^A)^2 + m_1^A m_3^A)(2(m_1^A)^2 m_3^A - m_1^A (m_2^A)^2 - m_2^A m_3^A)}$$

$$\hat{s}^A = \frac{-m_1^A (m_2^A)^2 + 2(m_1^A)^2 m_3^A - m_2^A m_3^A}{(m_1^A)^2 m_2^A - 2(m_2^A)^2 + m_1^A m_3^A}.$$

Substituting A with B we get the kinetic parameters for the B allele. To get the sample moments, we propose a novel histogram repiling method that gives the sample distribution and sample moment estimates of the true expression from the distribution of the observed expression (Figure 3.18). Specifically, for each gene we denote $c(Q)$ as the number of cells with observed expression Q and $n(Y)$ as the number of cells with the corresponding true expression Y. $c(Q)$ follows a Binomial distribution indexed at $n(Y)$ with probability of no dropout:

$$c(Q) \sim \text{Binomial}\big(n(Y), \text{expit}(\hat{\kappa} + \hat{\tau} \log Y)\big).$$

Then,

$$\hat{n}(Y) = \frac{c(Q)}{\text{expit}(\hat{\kappa} + \hat{\tau} \log Y)} = \frac{c(Q)}{\text{expit}\left(\hat{\kappa} + \frac{\hat{\tau}}{\hat{\beta}} \log \frac{Q}{\hat{\alpha}}\right)}.$$

These moment estimates of the kinetic parameters are sometimes negative as is pointed out by Kim and Marioni (116). By *in silico* simulation studies, we investigate the estimation accuracy and robustness under different settings.

3.4.4 Hypothesis Testing Framework

We carry out a nonparametric bootstrap hypothesis testing procedure with the null hypothesis that the two alleles of a gene share the same kinetic parameters. The procedures are as follow.

(i) For gene g, let $\{Q_{1g}^{A}, Q_{2g}^{A}, \dots, Q_{ng}^{A}\}$ and $\{Q_{1g}^{B}, Q_{2g}^{B}, \dots, Q_{ng}^{B}\}$ be the observed allele-specific read counts. Estimate allele-specific kinetic parameters with adjustment of technical variability:

$$\hat{\theta}^{A} = \{\hat{k}_{on,g}^{A}, \hat{k}_{off,g}^{A}, \hat{s}_{g}^{A}\};\ \hat{\theta}^{B} = \{\hat{k}_{on,g}^{B}, \hat{k}_{off,g}^{B}, \hat{s}_{g}^{B}\}.$$

(ii) Combine the $2n$ observed allelic measurements and draw samples of size $2n$ from the combined pool with replacement. Assign the first n with their corresponding cell sizes to allele A as $\{Q_{1g}^{A*}, Q_{2g}^{A*}, \dots, Q_{ng}^{A*}\}$, the next n to allele B $\{Q_{1g}^{B*}, Q_{2g}^{B*}, \dots, Q_{ng}^{B*}\}$. Estimate kinetic parameters with adjustment of technical variability from the bootstrap samples:

$$\theta^{A*} = \{k_{on,g}^{A*}, k_{off,g}^{A*}, s_{g}^{A*}\};\ \theta^{B*} = \{k_{on,g}^{B*}, k_{off,g}^{B*}, s_{g}^{B*}\}.$$

Iterate this N times.

(iii) Compute the *p*-values:

$$p = \frac{\sum \mathbb{1}\left(|\theta^{A*} - \theta^{B*}| \geq \left|\hat{\theta}^{A} - \hat{\theta}^{B}\right|\right)}{N}.$$

We adopt a Binomial test of allelic imbalance with the null hypothesis that the allelic ratio of the mean expression across all cells is 0.5. Chi-square test of independence is further performed to test whether the two alleles of a gene fire independently. The observed number of cells is from the direct output of the Bayes gene categorization framework. For all hypothesis testing, we adopt FDR to adjust for multiple comparisons.

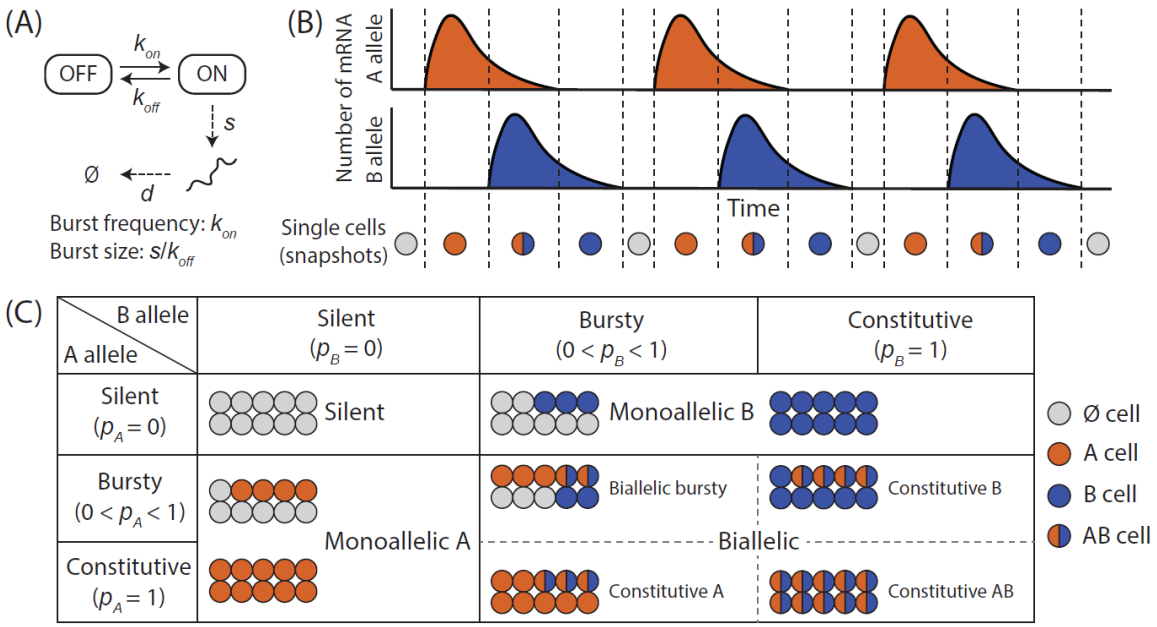


Figure 3.1: ***Allele-specific transcriptional bursting and gene categorization by single-cell ASE.*** (A) Transcription from DNA to RNA occurs in bursts, where genes switch between the "ON" and the "OFF" states. k_{on}, k_{off}, s, and d are activation, deactivation, transcription, and mRNA decay rate in the kinetic model respectively. (B) Transcriptional bursting of the two alleles of a gene give rise to cells expressing neither, one, or both alleles of a gene, sampled as vertical snapshots along the time axis. Partially adapted from Reinius and Sandberg (97). (C) Empirical Bayes framework that categorizes each gene as silent, monoallelic and biallelic (biallelic bursty, one-allele constitutive, and both-alleles constitutive) based on ASE data with single-cell resolution.

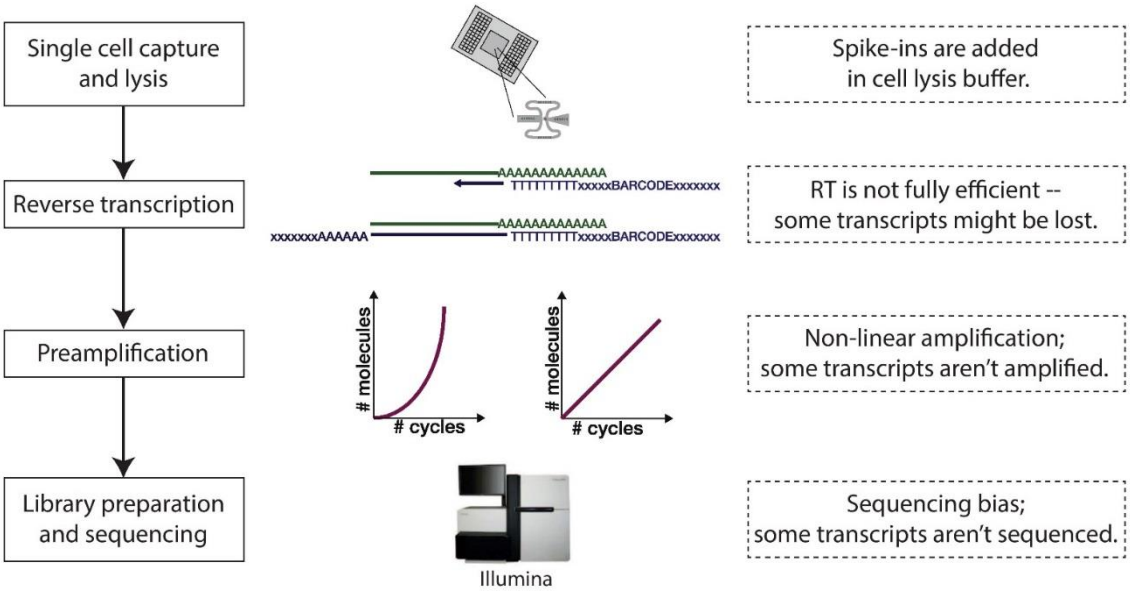


Figure 3.2: ***scRNA-seq protocol and technical variability.*** *Dropouts and amplification and sequencing bias are introduced in library preparation and sequencing. These technical variability needs to be adjusted for accurate and unbiased downstream analysis.*

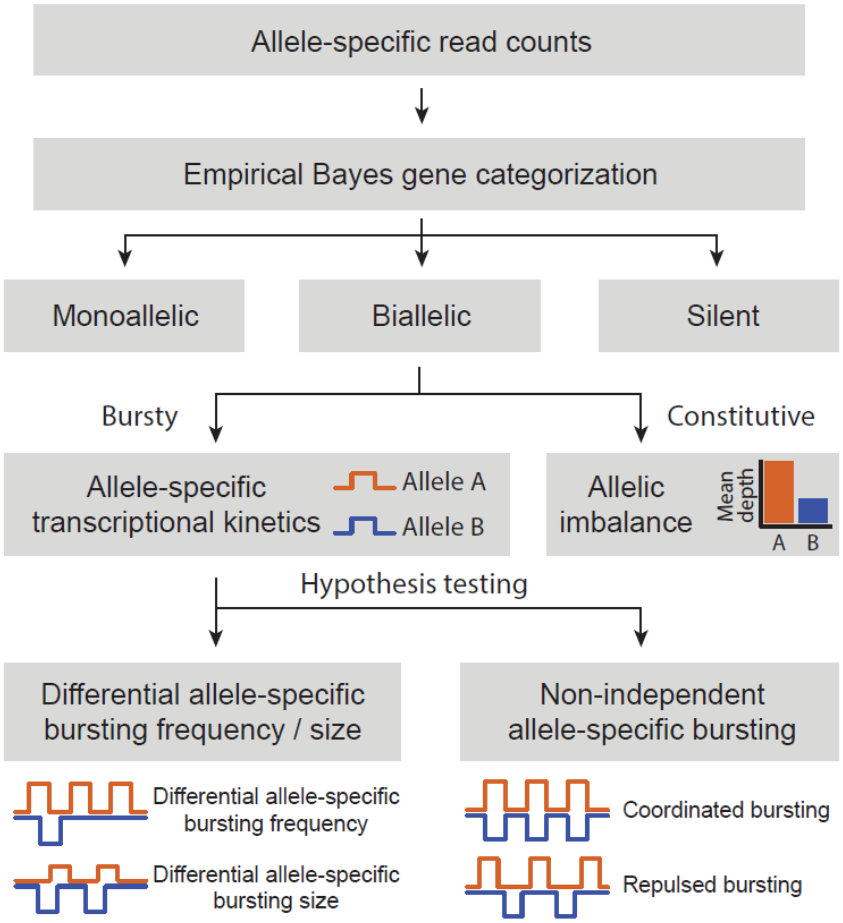


Figure 3.3: ***Overview of analysis pipeline of SCALE.*** *SCALE takes as input allele-specific read counts at heterozygous loci and carries out three major steps: (i) an empirical Bayes method for gene classification, (ii) a Poisson-Beta hierarchical model to estimate allele-specific transcriptional kinetics with adjustment of technical variability and cell size, (iii) a hypothesis testing framework to test the two alleles of a gene have differential bursting kinetics and/or non-independent firing.*

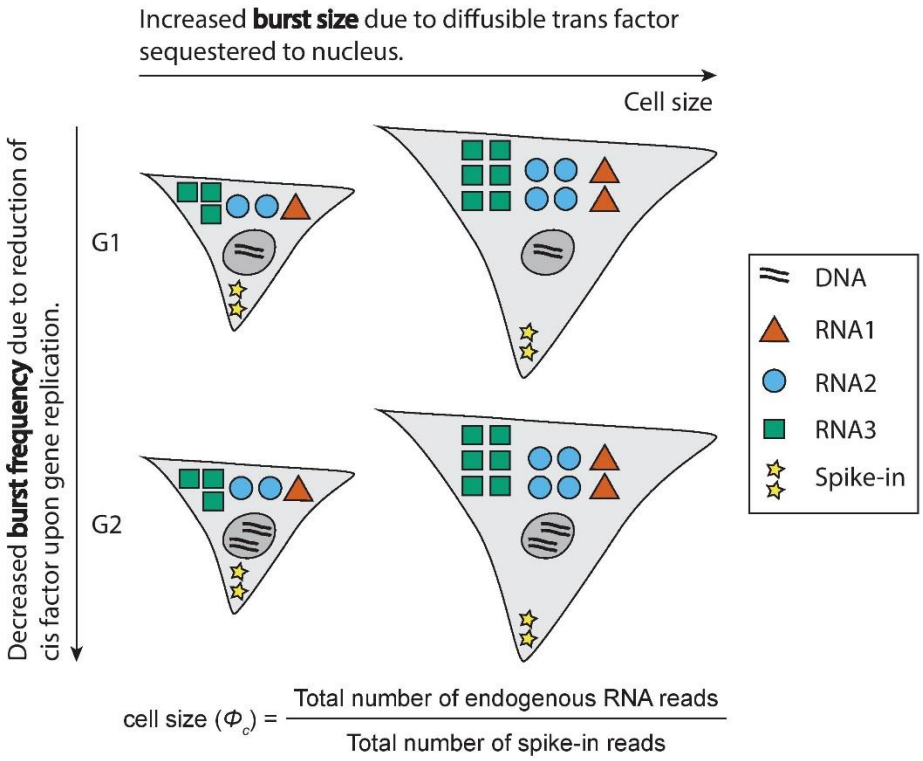


Figure 3.4: ***Cell size and cell cycle affects transcriptional bursting.*** *Large cell size leads to large burst size due to trans-effect whereas cells with duplicated DNAs in G2 phase have decreased burst frequency due to cis-effect. Spike-ins are added as internal controls. Plot is partially adapted from Padovan-Merhar et al. (128).*

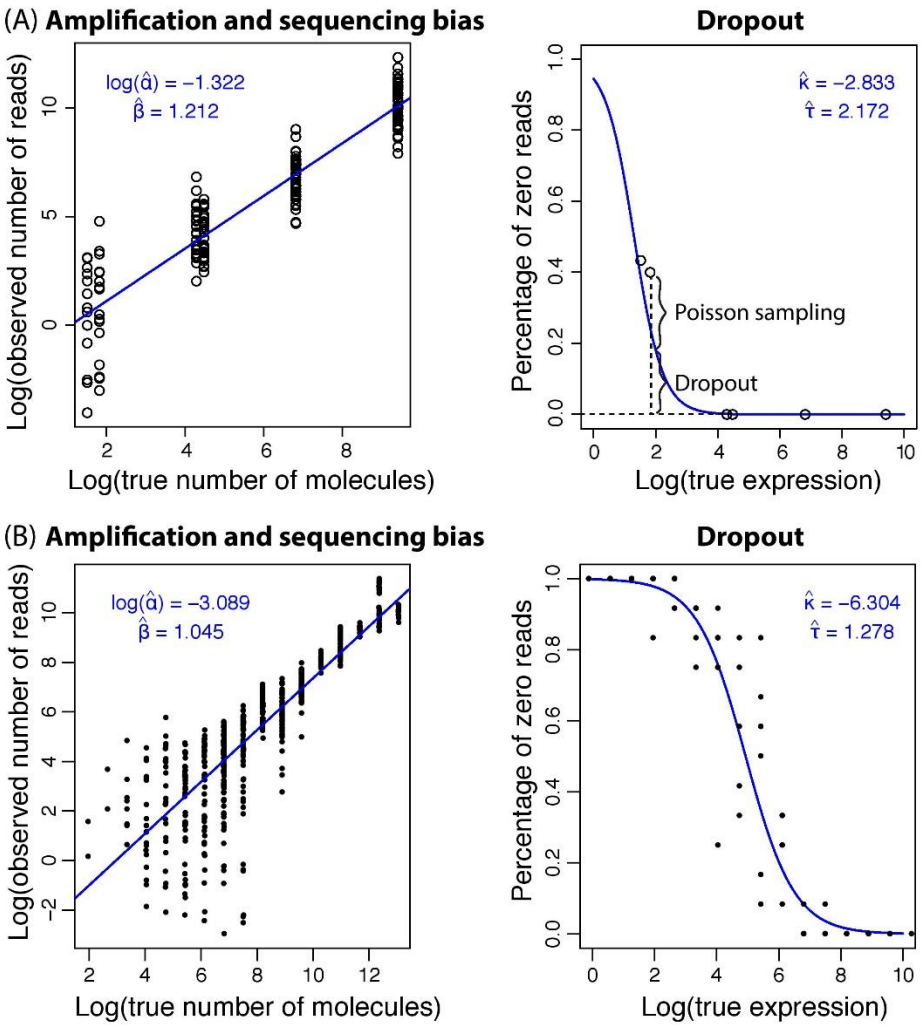


Figure 3.5: ***Modeling of technical variability and parameter estimation.*** *Amplification and sequencing bias are modeled and captured by parameter α and β. Estimation is carried out by log-linear regression. Probability of dropout is modeled by κ and τ and depends on the logarithm of the true expression. Estimation is carried out by the Nelder-Mead simplex algorithm. (A) Estimation results from 8 spike-ins from mouse blastocyst cells (93). The percentage of zero read counts are decomposed into those from Poisson sampling and those from dropout (spike-ins are non-bursty). (B) Estimation results from 92 ERCC spike-ins from human fibroblast cells (108).*

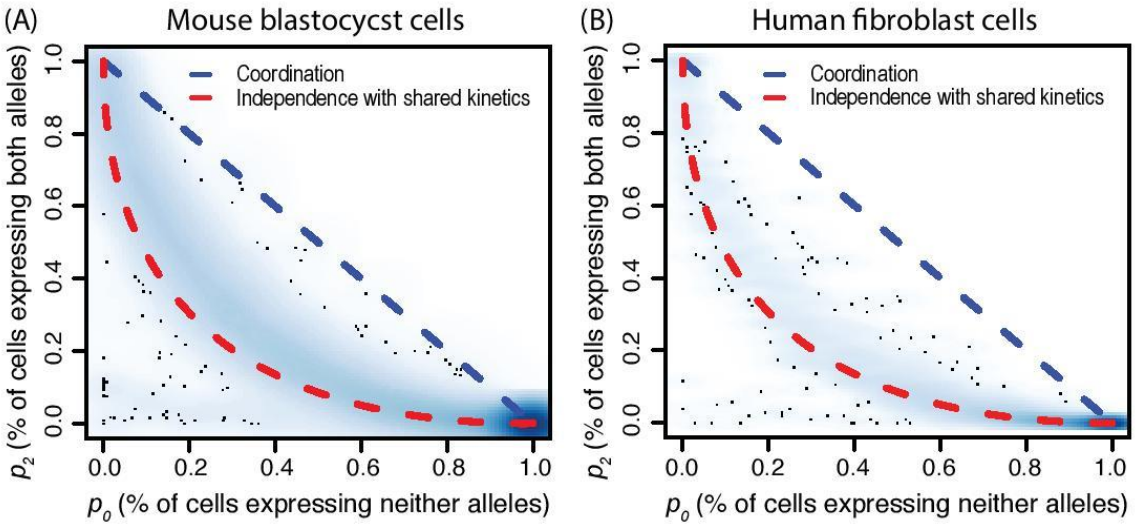


Figure 3.6: ***Gene categorization results on scRNA-seq dataset of mouse blastocyst and human fibroblast cells.*** *For each gene, the proportion of cells expressing neither, one, or both alleles, estimated through the Bayes procedure are denoted as* p_0, p_1, *and* p_2. *The smoothed scatterplot of* p_2 *against* p_0 *across all genes is shown. If the two alleles of a gene are expressed in a coordinated fashion, then there is no monoallelic expression and thus* $p_0 + p_2 = 1$, *which corresponds to the diagonal line. If the two alleles fire independently and share the same bursting kinetics, let* $p = p_A = p_B$ *be the proportion of cells expressing each allele, then we have* $p_0 = (1-p)^2$, $p_1 = 2p(1-p)$, *and* $p_2 = p^2$. *This corresponds to the red curve, where* $p_2 = \left(\sqrt{p_0} - 1\right)^2$. *The observed data, on the genome-wide scale, generally don't show significant deviations from this red curve, providing visual evidence that for most genes the assumption of shared bursting kinetics and independent bursting between the two alleles is reasonable. Smooth scatterplot is plotted by smoothScatter function in R. For genes that are significantly deviated, hypothesis testing is carried out to determine whether it is due to differential bursting kinetics and/or non-independent bursting between the two alleles. (A) Results from mouse blastocyst cells. (B) Results from human fibroblast cells.*

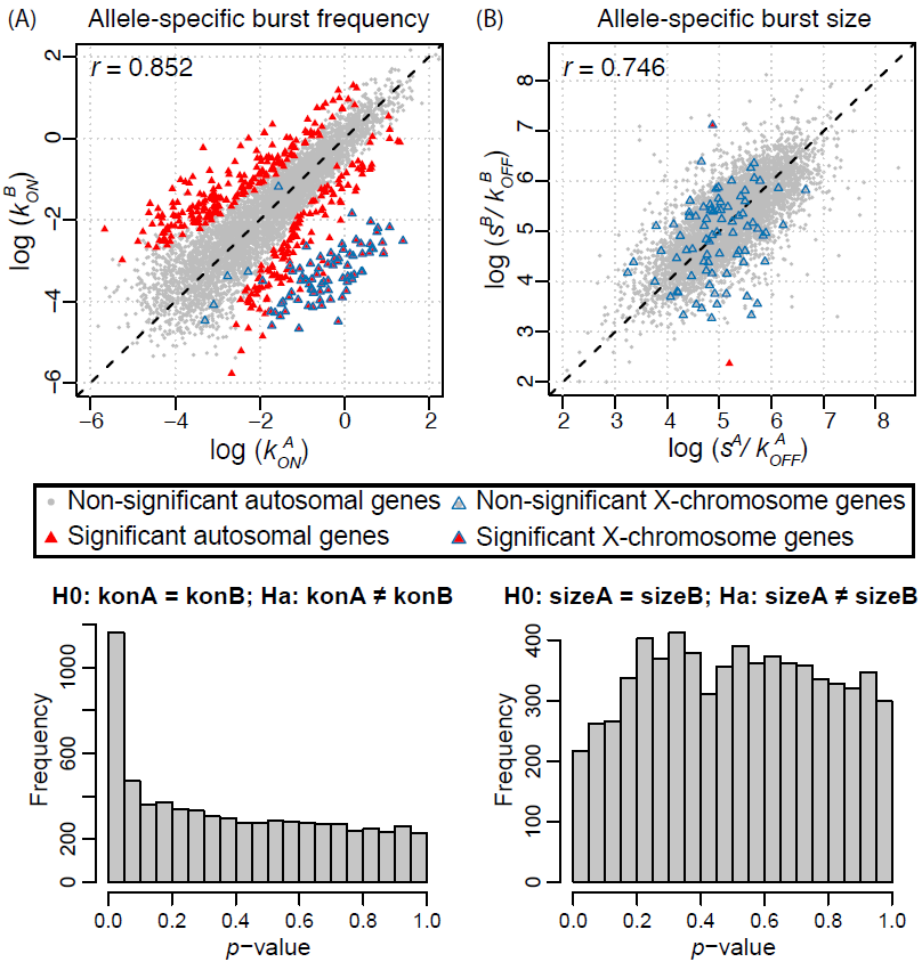


Figure 3.7: ***Allele-specific transcriptional kinetics of 7486 genes from 122 mouse blastocyst cells.*** *(A) Burst frequency of the two alleles has a correlation of 0.852. 425 genes show significant allelic difference in burst frequency after FDR control. (B) Burst size of the two alleles has a correlation of 0.746. Two genes show significant allelic difference in burst size. X-chromosome genes as positive controls show significant higher burst frequencies of the maternal alleles than those of the paternal alleles. The p-values for allelic burst size difference (bottom right panels) are uniformly distributed as expected under the null, whereas those for allelic burst frequency difference (bottom left panels) have a spike below significance level after FDR control.*

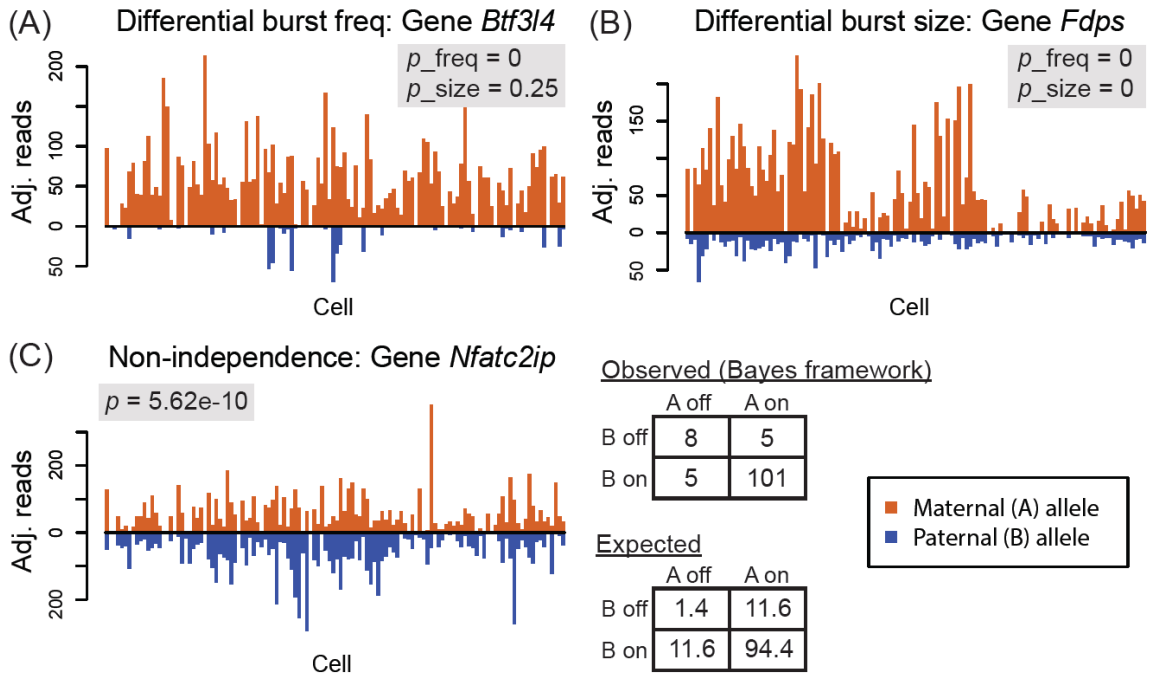


Figure 3.8: ***Examples of significant genes from hypothesis testing.*** *(A) The two alleles of the gene have significantly differential burst frequency from the bootstrap-based testing. (B) The two alleles of the gene have significantly differential burst size and burst frequency. (C) The two alleles of the gene fire non-independently from the chi-square test of independence.*

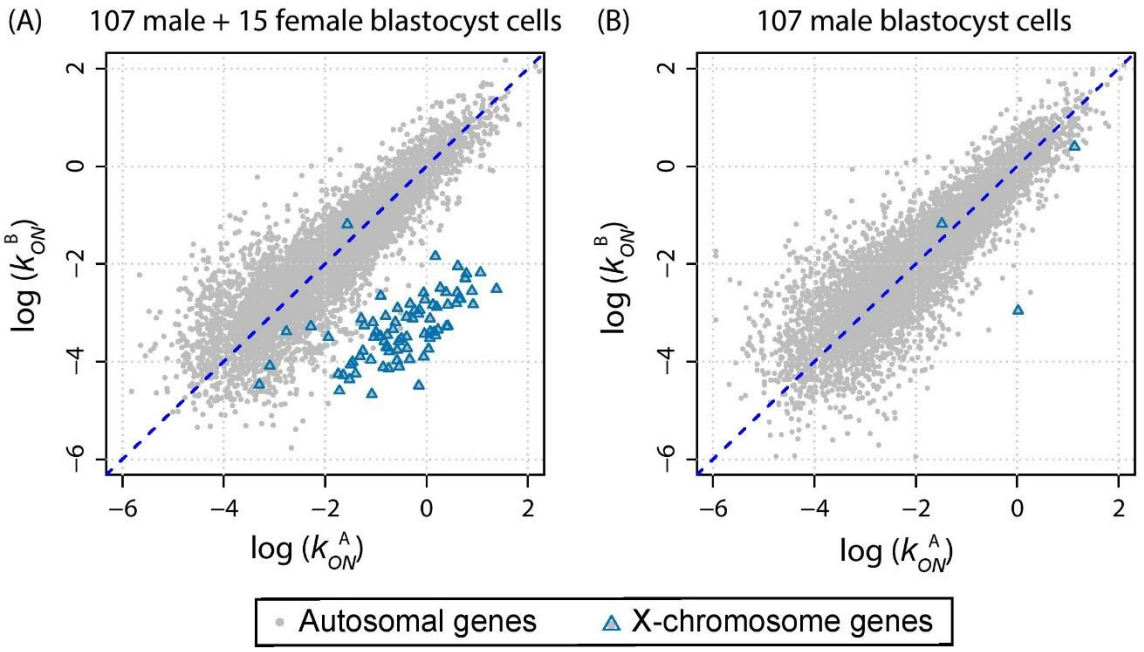


Figure 3.9: ***Allele-specific kinetic parameter estimation using bursty X-chromosome genes as positive controls.*** *When the sample pool is mixed with male (X^AY) and female (X^AX^B) cells, the maternal A allele has significantly higher burst frequency than the paternal B allele while the burst size difference remains insignificant. When the sample pool consists of male (X^AY) cells only, the bursty X-chromosome genes are categorized as maternal monoallelic A expression, whose allelic kinetic parameters for the paternal B allele are not estimable. X-chromosome genes serve as a positive control and a sanity check, which shows that SCALE estimates the allele-specific kinetics as is expected.*

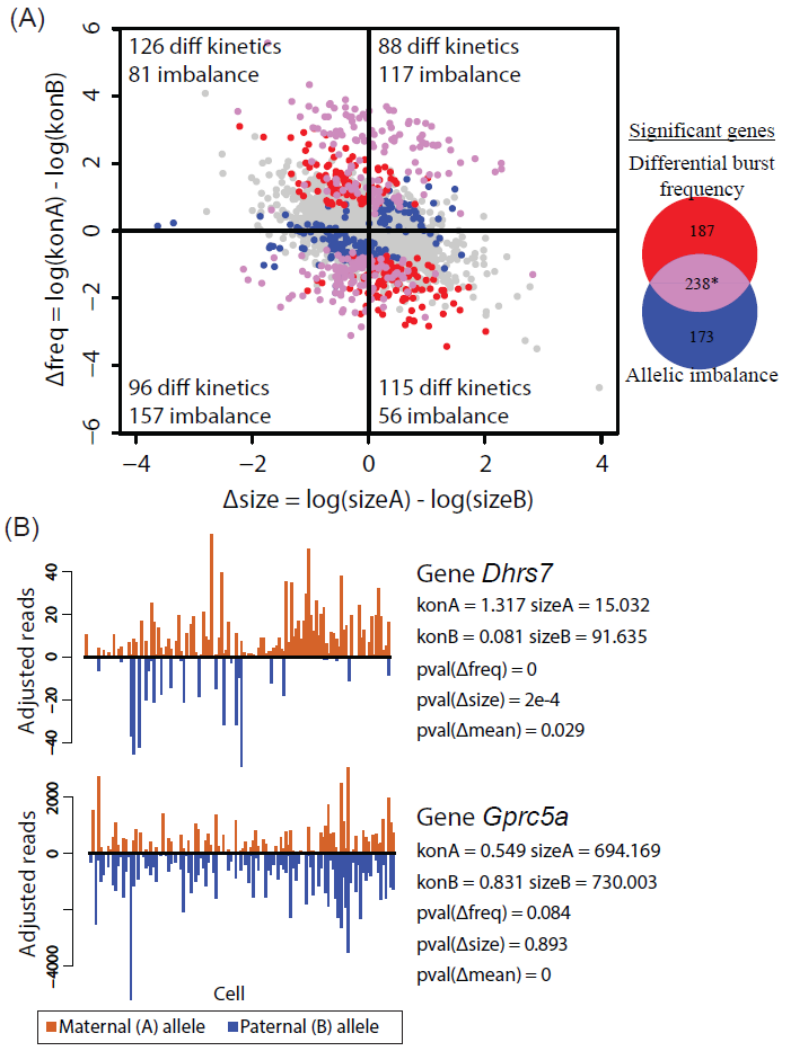


Figure 3.10: ***Testing of bursting kinetics by scRNA-seq and testing mean difference by bulk-tissue sequencing.*** *(A) Venn diagram of genes that are significant from testing of shared burst frequency and allelic imbalance. *Also includes the two genes that are significant from testing of shared burst size. Change in burst frequency and burst size in the same direction leads to higher detection power of allelic imbalance; change in different direction leads to allelic imbalance testing being underpowered. (B) Gene Dhrs7 whose two alleles have bursting kinetics in different direction and gene Gprc5a whose two alleles have bursting kinetics in the same direction. Dhrs7 is significant from testing of differential allelic bursting kinetics; Gprc5a is significant from the testing of mean difference between the two alleles.*

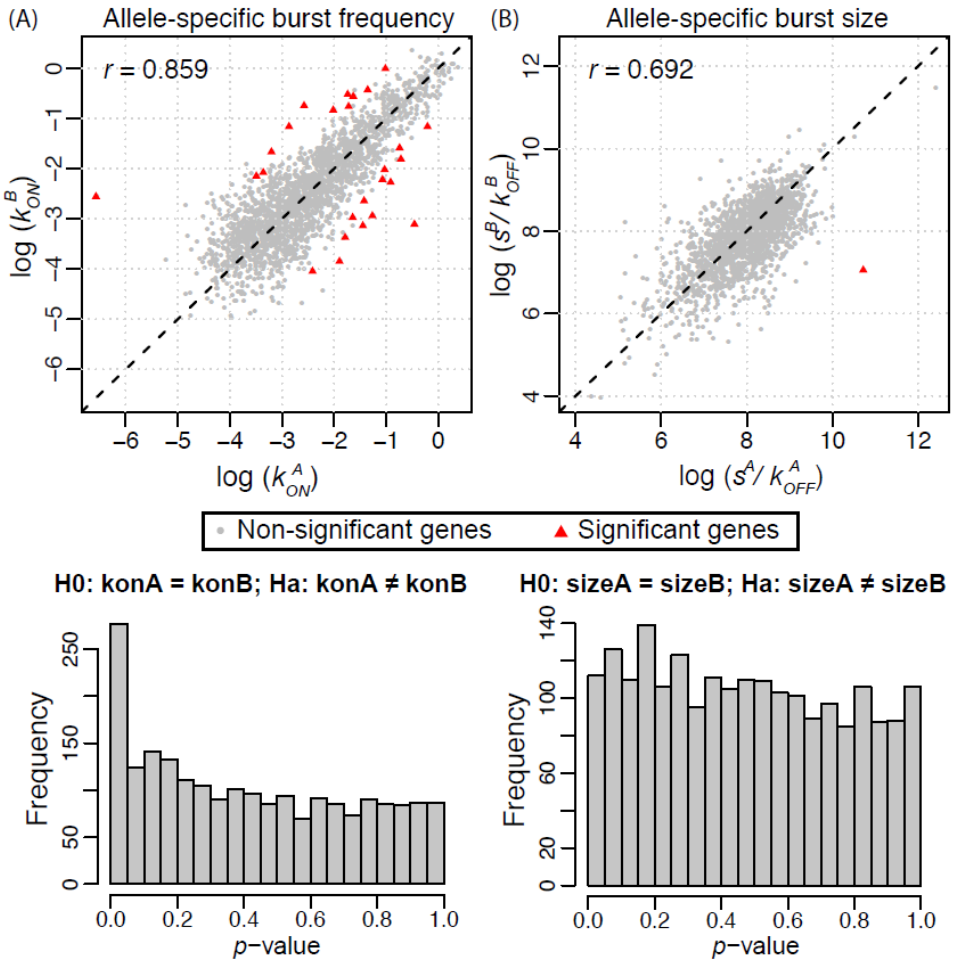


Figure 3.11: ***Allele-specific transcriptional kinetics of 2277 genes from 104 human fibroblast cells.*** *(A) Burst frequency of the two alleles has a correlation of 0.859. 26 genes show significant allelic difference in burst frequency after FDR. (B) Burst size of the two alleles has a correlation of 0.692. One gene has significant allelic difference in burst size. The results are concordant with the findings from the mouse embryonic development study.*

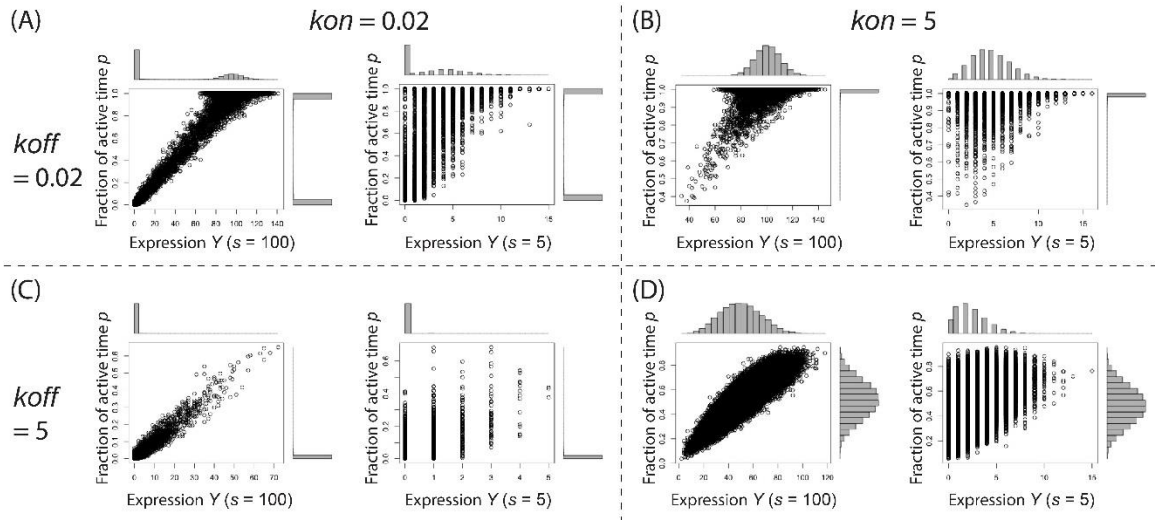
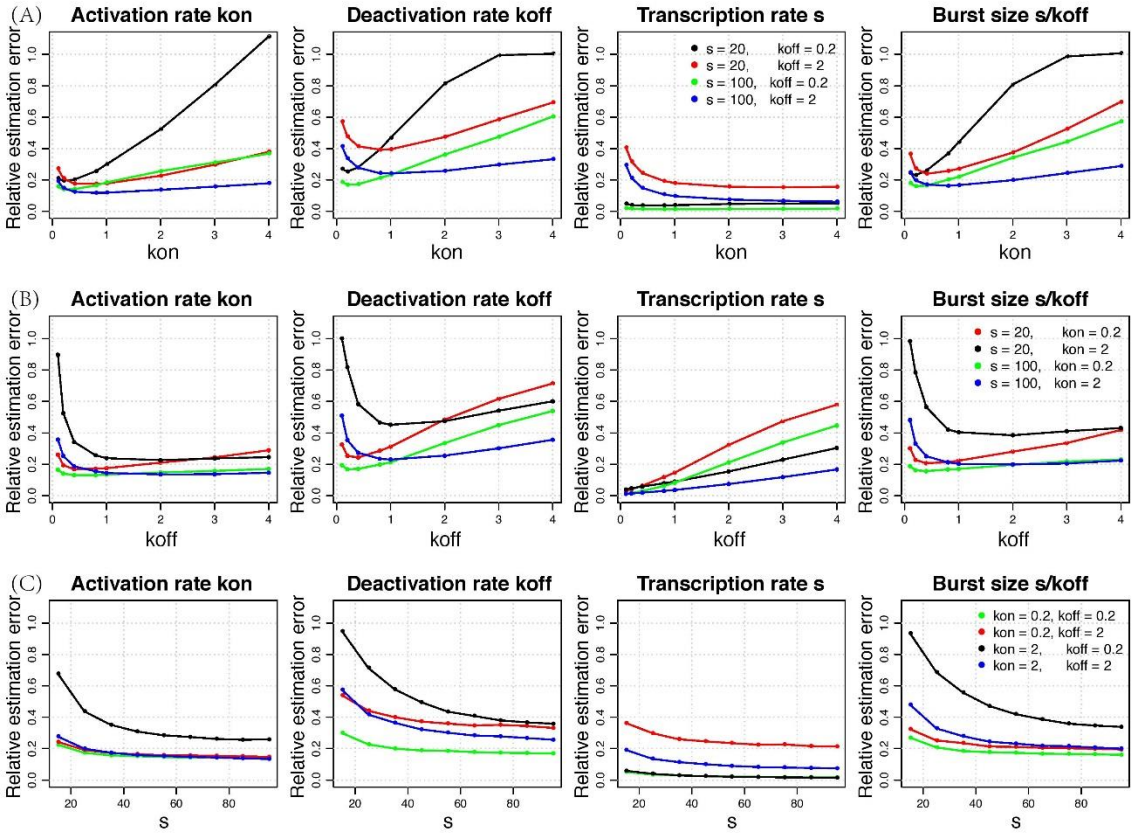


Figure 3.12: ***Three classes of Poisson-Beta transcription model.*** *Each dot corresponds to a cell, whose read count is generated in silico with underlying true parameters shown in each panel. (A) Genes with small k_{on} and small k_{off} are bursty, whose bursting kinetic parameters are identifiable. (B) Genes with large k_{on} and small k_{off} are typically highly expressed – the system collapses down to a constitutive expression model, resulting in a Poisson or negative-Binomial-like distribution. (C) Genes with small k_{on} and large k_{off} have low expression in most cells and high expression in a small number of cells, shown as a long exponential tail. (D) Genes with large k_{on} and large k_{off} are statistically hard to be distinguished from genes shown in (B).*


(A)
Activation rate kon
Deactivation rate koff
Transcription rate s
Burst size s/koff
Relative estimation error
kon
s = 20, koff = 0.2
s = 20, koff = 2
s = 100, koff = 0.2
s = 100, koff = 2
(B)
koff
s = 20, kon = 0.2
s = 20, kon = 2
s = 100, kon = 0.2
s = 100, kon = 2
(C)
s
kon = 0.2, koff = 0.2
kon = 0.2, koff = 2
kon = 2, koff = 0.2
kon = 2, koff = 2

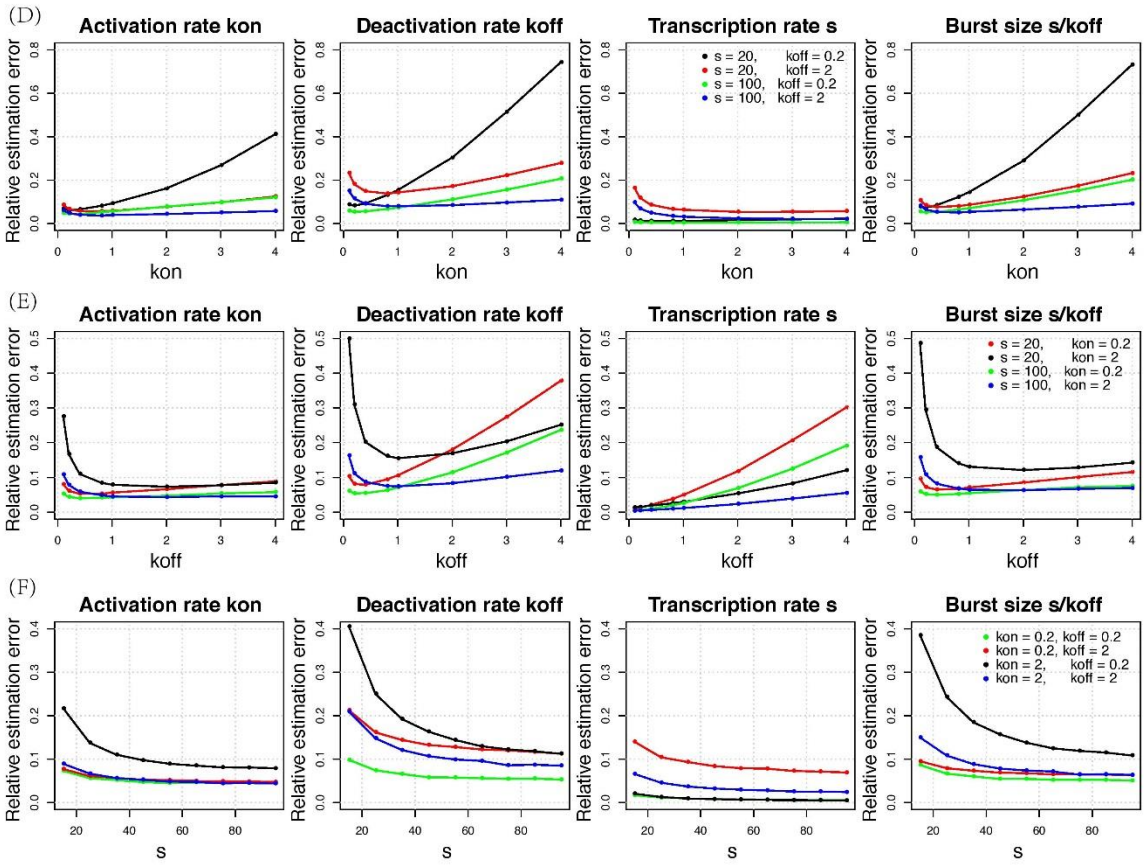


Figure 3.13: ***Assessment of moment estimators by simulations studies.*** *Estimation accuracy is measured by relative estimation error* $|\hat{\theta} - \theta|/\theta$ *for* k_{on}, k_{off}, s, *and* s/k_{off}. *Simulation is carried out with different underlying true parameters across 100 and 1000 cells: (A) varied* k_{on} *with fixed* s *and* k_{off} *acorss 100 cells; (B) varied* k_{off} *with fixed* s *and* k_{on} *across 100 cells; (C) varied* s *with fixed* k_{on} *and* k_{off} *across 100 cells; (D) varied* k_{on} *with fixed* s *and* k_{off} *acorss 1000 cells; (E) varied* k_{off} *with fixed* s *and* k_{on} *across 1000 cells; (F) varied* s *with fixed* k_{on} *and* k_{off} *across 1000 cells. Cases where* $k_{on} \ll k_{off}$ *(silence) and* $k_{on} \gg k_{off}$ *(constitutive expression), shown as red and black curves, have high estimation errors.* k_{off} *has higher estimation uncertainty than* s *in burst size.*

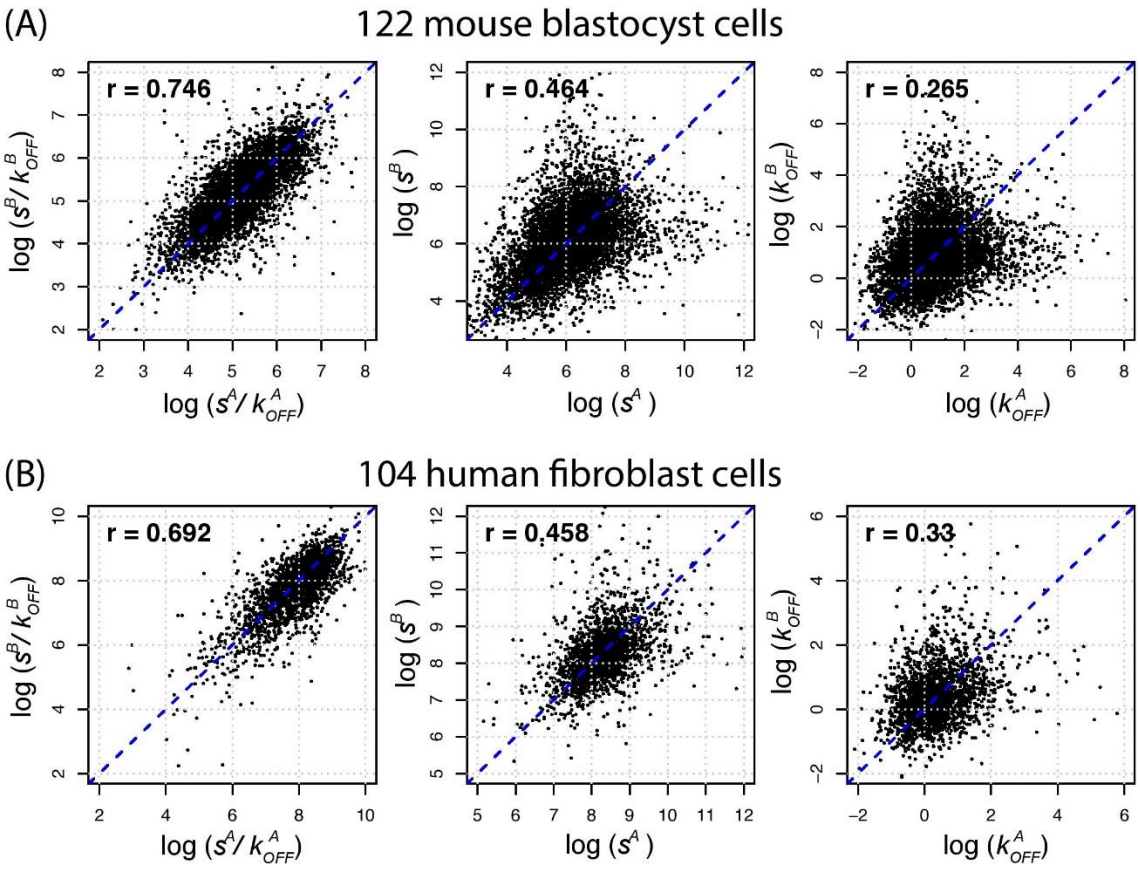


Figure 3.14: ***Correlation between allele-specific burst size s/k_{off}, transcription rate s, and deactivation rate k_{off}.*** *Over/under estimation of s is compensated by over/under estimation of k_{off}, resulting in the ratio burst size (s/k_{off}) having higher correlation between the two alleles. Each point is a biallelic bursty gene, whose kinetic parameters are estimated from real dataset of (A) 122 mouse blastocyst cells (93) and (B) 104 human fibroblast cells (108).*

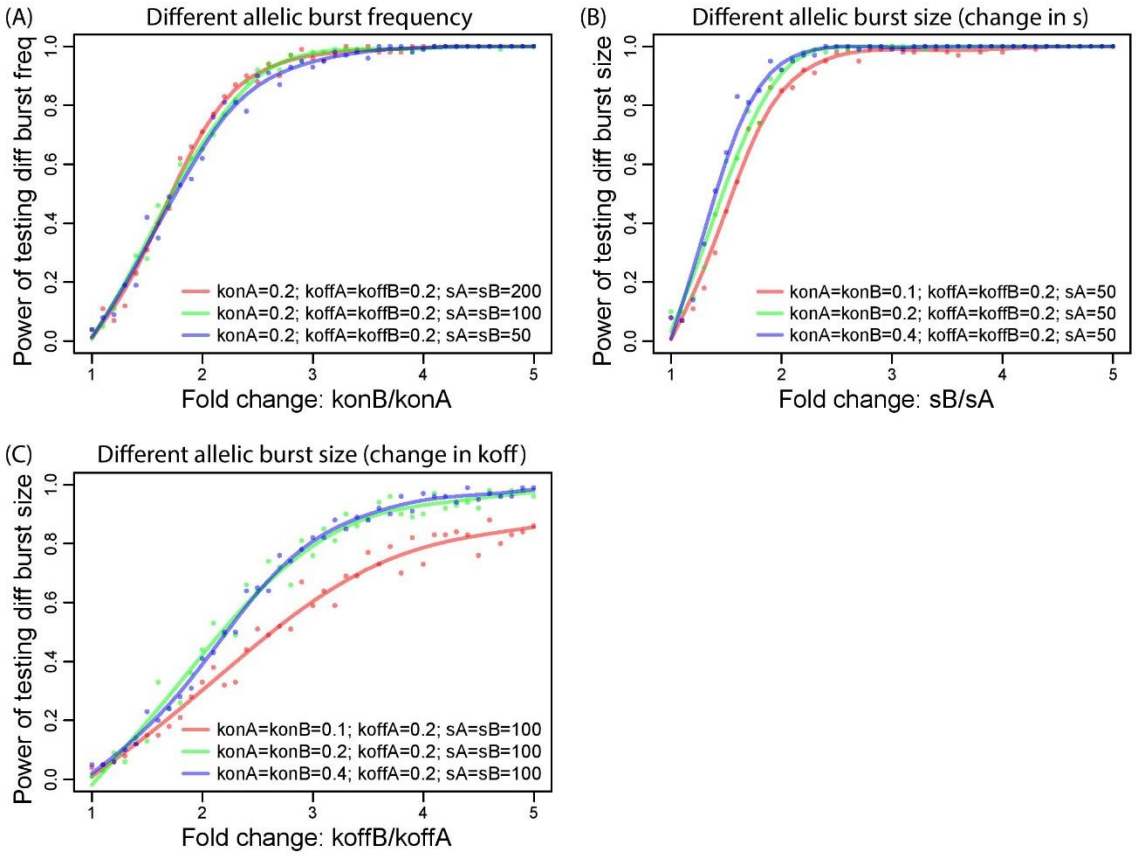
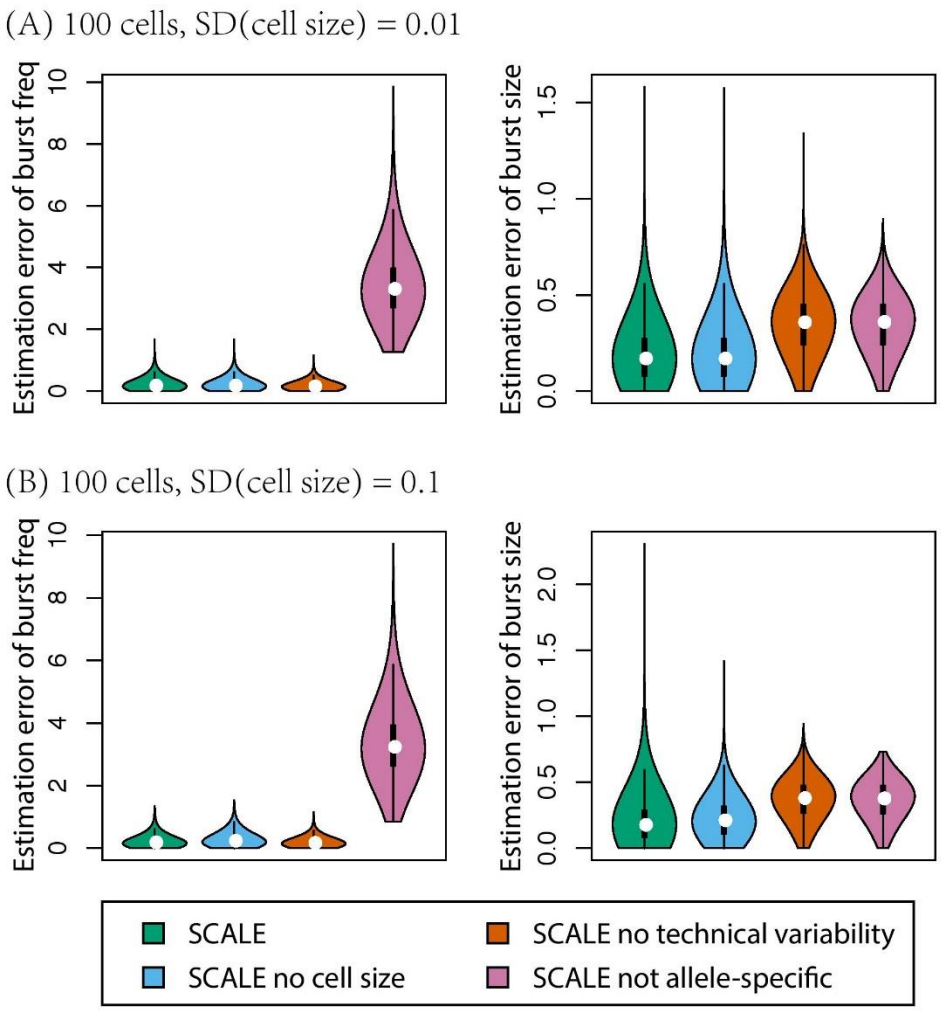


Figure 3.15: ***Power analysis for hypothesis testing of differential burst frequency and burst size between the two alleles.*** *The null hypothesis is both alleles sharing the same bursting kinetics* ($k_{on}^{A} = k_{on}^{B} = 0.2, k_{off}^{A} = k_{off}^{B} = 0.2, s^{A} = s^{B} = 50$)*. Different alternative hypotheses are included in the figure legends: (A) differential burst frequency; (B) differential burst size due to change in* s*; and (C) differential burst size due to change in* k_{off}*. Overall, the testing of burst frequency and burst size have similar power with relatively low power if the allelic difference in burst size is due to difference in the deactivation rate* k_{off}*. Power is evaluated at 0.05 significance level, suggesting a reduced power if a more stringent* p*-value cutoff is adopted.*


(A) 100 cells, SD(cell size) = 0.01
Estimation error of burst freq
0
2
4
6
8
10
Estimation error of burst size
0.0
0.5
1.0
1.5
(B) 100 cells, SD(cell size) = 0.1
Estimation error of burst freq
0
2
4
6
8
10
Estimation error of burst size
0.0
0.5
1.0
1.5
2.0
SCALE
SCALE no cell size
SCALE no technical variability
SCALE not allele-specific

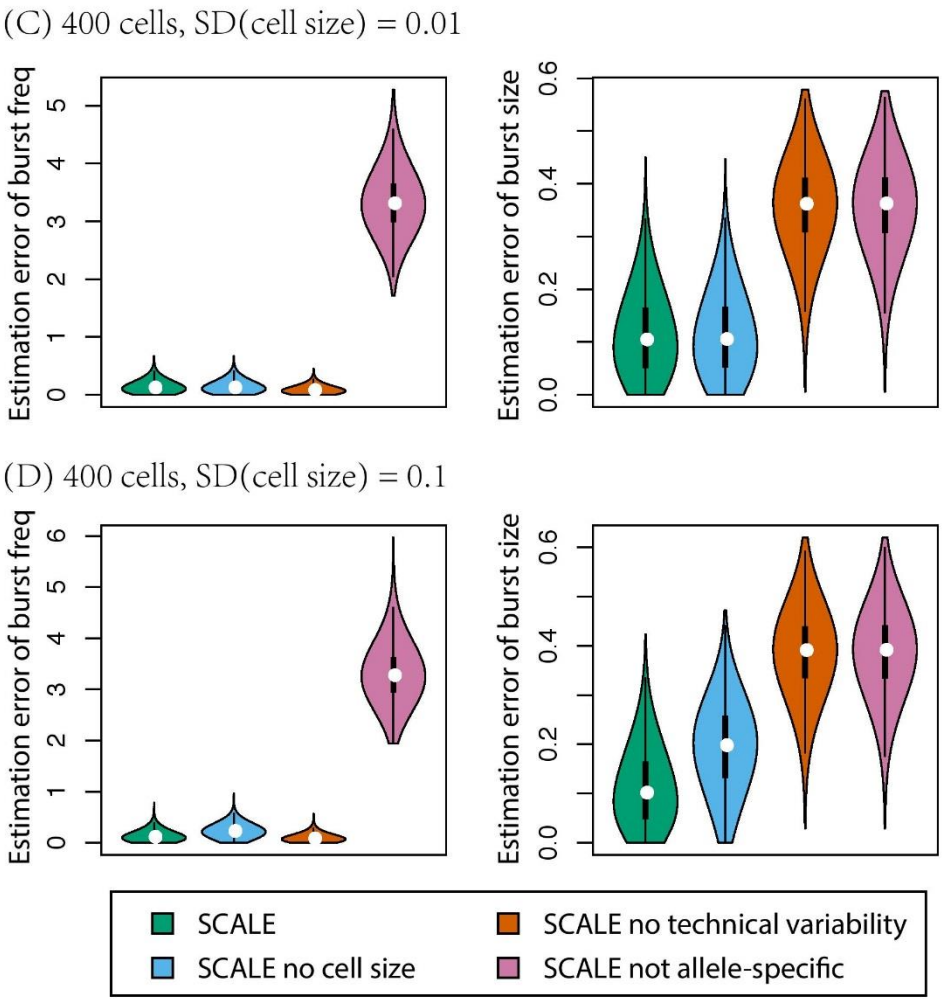
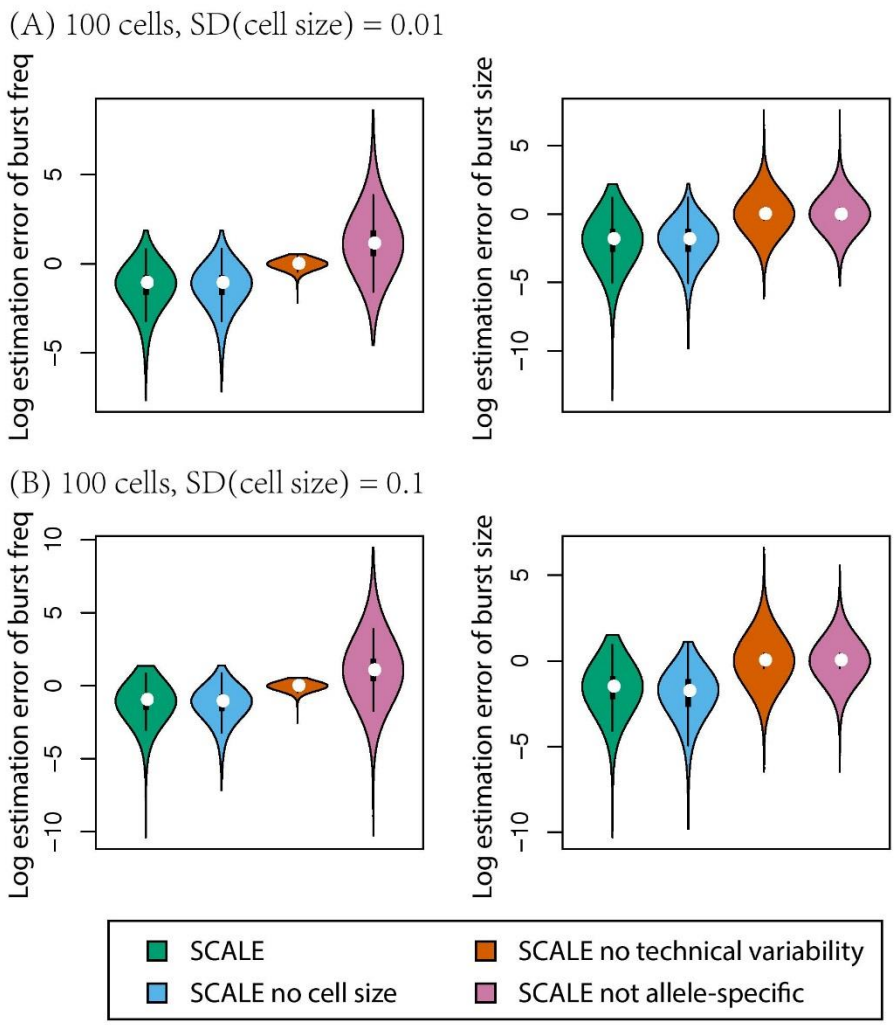


Figure 3.16: ***Adjustment of cell size and technical variability leads to more accurate estimation of allelic bursting kinetics.*** *Relative estimation error of burst frequency and burst size are measured through 5000 simulations across 100 and 400 cells respectively with fixed underlying true allele-specific kinetics ($k_{on}^{A} = k_{on}^{B} = k_{off}^{A} = k_{off}^{B} = 0.2, s^{A} = s^{B} = 100$). Technical variability is simulated with the estimated parameters from the mouse blastocyst dataset (Figure S5A). Cell size is simulated from a normal distribution with mean 1 and standard deviation 0.1 and 0.01 respectively. SCALE is applied in its default setting, without accounting for cell size, without adjustment of technical variability, and not in an allele-specific manner (using total coverage as input). SCALE in its default setting has the smallest relative estimation error across all four parallel runs. The estimation accuracy improves as the number of cells increases.*


(A) 100 cells, SD(cell size) = 0.01
Log estimation error of burst freq
Log estimation error of burst size
(B) 100 cells, SD(cell size) = 0.1
Log estimation error of burst freq
Log estimation error of burst size
SCALE
SCALE no cell size
SCALE no technical variability
SCALE not allele-specific

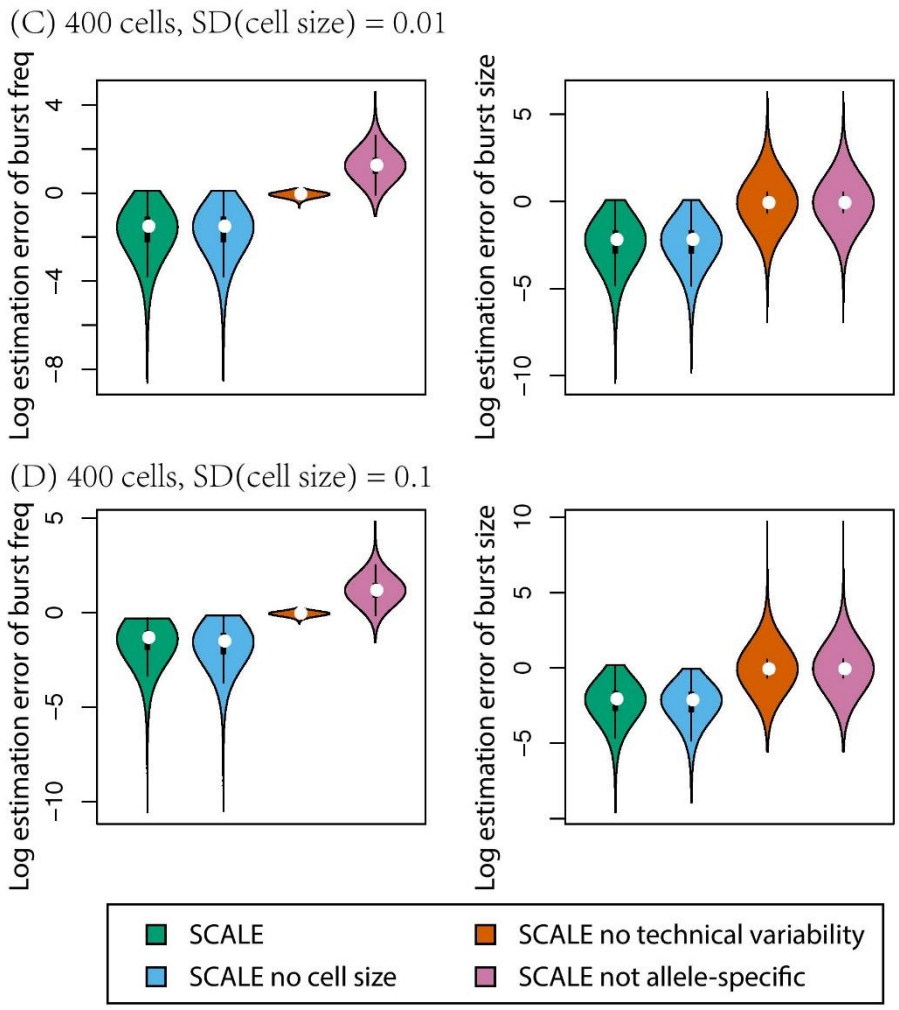


Figure 3.17: ***Adjustment of cell size and technical variability leads to more accurate estimation of allelic bursting kinetics.*** *Relative estimation error of burst frequency and burst size are measured through 5000 simulations across 100 and 400 cells respectively with fixed underlying true allele-specific kinetics ($k_{on}^{A} = k_{on}^{B} = k_{off}^{A} = k_{off}^{B} = 0.2, s^{A} = s^{B} = 100$). Technical variability is simulated with the estimated parameters from the human fibroblast dataset (Figure S5B). Cell size is simulated from a normal distribution with mean 1 and standard deviation 0.1 and 0.01 respectively. SCALE is applied in its default setting, without accounting for cell size, without adjustment of technical variability, and not in an allele-specific manner (using total coverage as input). SCALE in its default setting has the smallest relative estimation error across all four parallel runs. The estimation accuracy improves as the number of cells increases. Logarithm of the estimation error is shown as the Y-axis due to the completely-off estimation using total instead of allele-specific expression.*

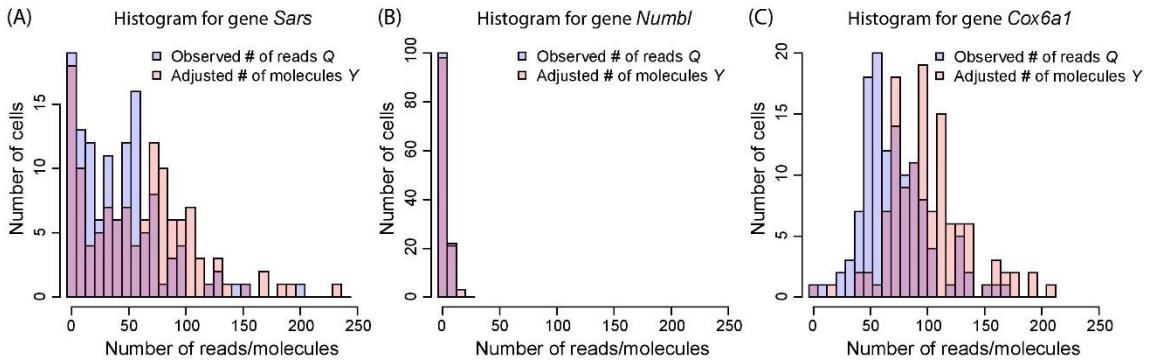


Figure 3.18: ***Histogram repiling method for kinetic parameter estimation with adjustment of technical variability.*** *Histogram of number of cells with observed read counts Q is shown in light blue; histogram of number of cells with true number of molecules Y is shown in light red. Three example genes are plotted: (A) Partial cells with zero read counts of a bursty gene are due to dropout events and are recovered to non-zero true number of molecules; (B) Gene is off in most cells; (C) Gene is constitutively expressed with expression levels adjusted for sequencing and amplification bias.*

Classes		Small k_{on} Small k_{off}	Small k_{on} Large k_{off}	Large k_{on} Small k_{off}	Large k_{on} Large k_{off}
True value	k_{on}	0.02	0.02	2	2
	k_{off}	0.02	2	0.02	2
	s	100	100	100	100
Estimate **Standard error** **Confidence Interval**	k_{on}	0.021 0.013 (-0.005, 0.047)	0.023 0.007 (0.01, 0.036)	1.973 616.977 (-17.216, 19.312)	1.964 0.324 (1.46, 2.721)
	k_{off}	0.022 0.015 (-0.005, 0.053)	1.662 13.185 (0.328, 5.51)	0.024 635.842 (0, 2.804)	2.13 0.921 (1.136, 4.546)
	s	99.1 1.537 (95.98, 102.01)	98.43 362.895 (24.48, 198.69)	100.49 30.995 (96.2, 105.03)	102.23 11.944 (86.48, 131.9)

Table 3.1: ***Standard errors and confidence intervals of estimated kinetic parameters.*** *Simulated dataset are generated from the Poisson-Beta transcriptional model with true underlying parameters shown in first row. Bootstrap resampling gives standard errors and confidence intervals of the moment estimates. Standard errors are large with unstable moment estimates for genes with $k_{on} \ll k_{off}$ (silence) and $k_{on} \gg k_{off}$ (constitutive expression).*

BIBLIOGRAPHY

1. McCarroll SA, Bradner JE, Turpeinen H, Volin L, Martin PJ, Chilewski SD, et al. Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. Nat Genet. 2009;41(12):1341-4.
2. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81.
3. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
4. Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, et al. Genome assembly comparison identifies structural variants in the human genome. Nat Genet. 2006;38(12):1413-8.
5. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. Copy number variation at 1q21.1 associated with neuroblastoma. Nature. 2009;459(7249):987-91.
6. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 2009;459(7246):569-73.
7. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet. 2008;40(9):1107-12.
8. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;20(2):207-11.
9. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.
10. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007;39(7 Suppl):S16-21.
11. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182-90.
12. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6(1):99-103.
13. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586-92.

14. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol. 2009;10(2):R23.
15. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-84.
16. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42(1):30-5.
17. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-9.
18. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011;43(4):309-15.
19. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41.
20. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 2015;43(6):e39.
21. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27(19):2648-54.
22. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76.
23. Amarasinghe KC, Li J, Halgamuge SK. CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC bioinformatics. 2013;14 Suppl 2:S2.
24. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012;28(21):2747-54.
25. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol. 2013;14(10):R120.

26. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22(8):1525-32.
27. Coin LJ, Cao D, Ren J, Zuo X, Sun L, Yang S, et al. An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. Bioinformatics. 2012;28(18):i370-i4.
28. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91(4):597-607.
29. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
30. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
31. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40(10):1166-74.
32. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
33. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, et al. The genetic landscape of high-risk neuroblastoma. Nat Genet. 2013;45(3):279-84.
34. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature. 2012;483(7391):589-93.
35. Cheung NK, Zhang J, Lu C, Parker M, Bahrami A, Tickoo SK, et al. Association of age at diagnosis and genetic mutations in patients with neuroblastoma. JAMA. 2012;307(10):1062-71.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
37. Lee S, Chugh PE, Shen H, Eberle R, Dittmer DP. Poisson factor models with applications to non-normalized microRNA profiling. Bioinformatics. 2013;29(9):1105-11.
38. Zhang NR, Siegmund DO, Ji HL, Li JZ. Detecting simultaneous changepoints in multiple sequences. Biometrika. 2010;97(3):631-45.

39. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72.
40. Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics. 2007;63(1):22-32.
41. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research. 2012;40(10).
42. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54.
43. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64.
44. Leek JT, Storey JD. A general framework for multiple testing dependence. Proc Natl Acad Sci U S A. 2008;105(48):18718-23.
45. Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J Am Stat Assoc. 2008;103(484):1438-56.
46. Friguet C, Kloareg M, Causeur D. A Factor Model Approach to Multiple Testing Under Dependence. J Am Stat Assoc. 2009;104(488):1406-15.
47. Sun YT, Zhang NR, Owen AB. Multiple Hypothesis Testing Adjusted for Latent Variables, with an Application to the Agemap Gene Expression Data. Ann Appl Stat. 2012;6(4):1664-88.
48. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.
49. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
50. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends in genetics : TIG. 1993;9(4):138-41.
51. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
52. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993-8.
53. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90-4.

54. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506-10.
55. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. The Journal of pathology. 2013;231(1):21-34.
56. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet. 2014;46(3):225-33.
57. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2015;518(7539):422-6.
58. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat Genet. 2015;47(3):209-16.
59. Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E, et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat Genet. 2015;47(7):736-45.
60. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30(5):413-21.
61. Andor N, Harness JV, Muller S, Mewes HW, Petritsch C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics. 2014;30(1):50-60.
62. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11(4):396-8.
63. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10(8):e1003665.
64. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24(11):1881-93.
65. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013;14(7):R80.
66. Li B, Li JZ. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. Genome Biol. 2014;15(9):473.

67. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
68. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512(7513):155-60.
69. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148(5):873-85.
70. Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. Genome Biol. 2015;16(1):91.
71. Niknafs N, Beleva-Guthrie V, Naiman DQ, Karchin R. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. PLoS Comput Biol. 2015;11(10):e1004416.
72. Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, et al. Inferring clonal composition from multiple sections of a breast cancer. PLoS Comput Biol. 2014;10(7):e1003703.
73. Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biol. 2015;16:36.
74. El-Kebir M, Satas G, Oesper L, Raphael BJ. Multi-State Perfect Phylogeny Mixture Deconvolution and Applications to Cancer Sequencing. arXiv preprint arXiv:160402605. 2016.
75. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113(37):E5528-37.
76. Chen H, Bell JM, Zavala NA, Ji HP, Zhang NR. Allele-specific copy number profiling by next-generation DNA sequencing. Nucleic Acids Res. 2015;43(4):e23.
77. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2015;26(1):64-70.
78. Lonnstedt IM, Caramia F, Li J, Fumagalli D, Salgado R, Rowan A, et al. Deciphering clonality in aneuploid breast tumors using SNP array and sequencing data. Genome Biol. 2014;15(9):470.
79. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC bioinformatics. 2014;15:35.

80. Gusfield D. Efficient Algorithms for Inferring Evolutionary Trees. Networks. 1991;21(1):19-28.
81. Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. 1969;61(4):893-903.
82. Minn AJ, Kang Y, Serganova I, Gupta GP, Giri DD, Doubrovin M, et al. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. The Journal of clinical investigation. 2005;115(1):44-55.
83. Jacob LS, Vanharanta S, Obenauf AC, Pirun M, Viale A, Socci ND, et al. Metastatic Competence Can Emerge with Selection of Preexisting Oncogenic Alleles without a Need of New Mutations. Cancer research. 2015;75(18):3713-9.
84. Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D, et al. Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci U S A. 2007;104(16):6740-5.
85. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2013;11(1110):11 0 1- 0 33.
86. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
87. Wagle N, Emery C, Berger MF, Davis MJ, Sawyer A, Pochanard P, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2011;29(22):3085-96.
88. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.
89. Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordon-Cardo C, et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell. 2003;3(6):537-49.
90. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, et al. Genes that mediate breast cancer metastasis to lung. Nature. 2005;436(7050):518-24.
91. Li SY, Pearl DK, Doss H. Phylogenetic tree construction using Markov chain Monte Carlo. J Am Stat Assoc. 2000;95(450):493-508.
92. Buckland PR. Allele-specific gene expression differences in humans. Hum Mol Genet. 2004;13 Spec No 2:R255-60.

93. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343(6167):193-6.
94. Gendrel AV, Attia M, Chen CJ, Diabangouaya P, Servant N, Barillot E, et al. Developmental dynamics and disease potential of random monoallelic gene expression. Dev Cell. 2014;28(4):366-80.
95. Eckersley-Maslin MA, Spector DL. Random monoallelic expression: regulating gene expression one allele at a time. Trends Genet. 2014;30(6):237-44.
96. Eckersley-Maslin MA, Thybert D, Bergmann JH, Marioni JC, Flicek P, Spector DL. Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev Cell. 2014;28(4):351-65.
97. Reinius B, Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nat Rev Genet. 2015;16(11):653-64.
98. Reinius B, Mold JE, Ramskold D, Deng Q, Johnsson P, Michaelsson J, et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet. 2016;48(11):1430-5.
99. Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, et al. SNP-specific array-based allele-specific expression analysis. Genome Res. 2008;18(5):771-9.
100. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res. 2011;21(10):1728-37.
101. Leon-Novelo LG, McIntyre LM, Fear JM, Graze RM. A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. BMC Genomics. 2014;15:920.
102. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 2015;16:195.
103. Knight JC. Allele-specific gene expression uncovered. Trends in genetics : TIG. 2004;20(3):113-6.
104. Bell CG, Beck S. Advances in the identification and analysis of allele-specific expression. Genome Med. 2009;1(5):56.
105. de la Chapelle A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. Oncogene. 2009;28(38):3345-8.
106. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature reviews Genetics. 2015;16(3):133-45.
107. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58(4):610-20.

108. Borel C, Ferreira PG, Santoni F, Delaneau O, Fort A, Popadin KY, et al. Biased allelic expression in human primary fibroblast single cells. Am J Hum Genet. 2015;96(1):70-80.
109. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. Curr Biol. 2006;16(10):1018-25.
110. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135(2):216-26.
111. Chong S, Chen C, Ge H, Xie XS. Mechanism of transcriptional bursting in bacteria. Cell. 2014;158(2):314-26.
112. Blake WJ, Balazsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, et al. Phenotypic consequences of promoter-mediated transcriptional noise. Mol Cell. 2006;24(6):853-65.
113. Fukaya T, Lim B, Levine M. Enhancer Control of Transcriptional Bursting. Cell. 2016;166(2):358-68.
114. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 2006;4(10):e309.
115. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332(6028):472-4.
116. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. 2013;14(1):R7.
117. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093-5.
118. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241.
119. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comput Biol. 2015;11(6):e1004333.
120. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, et al. Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics. 2015;31(13):2225-7.
121. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017.
122. Kim JK, Kolodziejczyk AA, Illicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun. 2015;6:8687.
123. Levesque MJ, Raj A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat Methods. 2013;10(3):246-8.

124. Jiang Y, Zhang NR, Li M. Modeling allele-specific gene expression by single-cell RNA sequencing. bioRxiv. 2017.
125. Kepler TB, Elston TC. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys J. 2001;81(6):3116-36.
126. Bix M, Locksley RM. Independent and epigenetic regulation of the interleukin-4 alleles in CD4+ T cells. Science. 1998;281(5381):1352-4.
127. Levesque MJ, Ginart P, Wei Y, Raj A. Visualizing SNVs to quantify allele-specific expression in single cells. Nat Methods. 2013;10(9):865-7.
128. Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol Cell. 2015;58(2):339-52.
129. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777-82.
130. Dadiani M, van Dijk D, Segal B, Field Y, Ben-Artzi G, Raveh-Sadka T, et al. Two DNA-encoded strategies for increasing expression with opposing effects on promoter dynamics and transcriptional noise. Genome Res. 2013;23(6):966-76.
131. Bartman CR, Hsu SC, Hsiung CC, Raj A, Blobel GA. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. Mol Cell. 2016;62(2):237-47.
132. Sepulveda LA, Xu H, Zhang J, Wang M, Golding I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. Science. 2016;351(6278):1218-22.
133. Skinner SO, Xu H, Nagarkar-Jaiswal S, Freire PR, Zwaka TP, Golding I. Single-cell analysis of transcription kinetics across the cell cycle. Elife. 2016;5.
134. Ochiai H, Sugawara T, Sakuma T, Yamamoto T. Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. Sci Rep. 2014;4:7125.
135. Xu H, Sepulveda LA, Figard L, Sokac AM, Golding I. Combining protein and mRNA quantification to decipher transcriptional regulation. Nat Methods. 2015;12(8):739-42.
136. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012;336(6078):183-7.
137. Edsgard D, Reinius B, Sandberg R. scphaser: haplotype inference using single-cell RNA-seq data. Bioinformatics. 2016;32(19):3038-40.
138. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev. 2006;20(5):515-24.

139. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013;31(8):748-52.