8-2020

Statistical Methods For Resolving Intratumor Heterogeneity With Single-Cell Dna Sequencing

Alexander Davis

STATISTICAL METHODS FOR RESOLVING INTRATUMOR HETEROGENEITY WITH SINGLE-CELL DNA SEQUENCING

by

Alexander Davis, BA

APPROVED:

_____
Nicholas E. Navin, Ph.D.
Advisory Professor

_____
Ken Chen, Ph.D.

_____
Wenyi Wang, Ph.D.

_____
Mary Edgerton, M.D., Ph.D.

_____
Luay Nakhleh, Ph.D.

APPROVED:

_____
Dean, The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

STATISTICAL METHODS FOR RESOLVING INTRATUMOR HETEROGENEITY WITH SINGLE-CELL DNA SEQUENCING

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Alexander Davis, BA

Houston, Texas

August, 2020

STATISTICAL METHODS FOR RESOLVING INTRATUMOR HETEROGENEITY WITH SINGLE-CELL DNA SEQUENCING

Alexander Davis, BA

Advisory Professor: Nicholas E. Navin, PhD

Tumor cells have heterogeneous genotypes, which drives progression and treatment resistance. Such genetic intratumor heterogeneity plays a role in the process of clonal evolution that underlies tumor progression and treatment resistance. Single-cell DNA sequencing is a promising experimental method for studying intratumor heterogeneity, but brings unique statistical challenges in interpreting the resulting data. Researchers lack methods to determine whether sufficiently many cells have been sampled from a tumor. In addition, there are no proven computational methods for determining the ploidy of a cell, a necessary step in the determination of copy number. In this work, software for calculating probabilities from a multinomial distribution was written to estimate the number of cells that must be sequenced (chapter 2). Two new methods were developed for predicting the number of mutations which would be discovered in additional single-cell sequencing of a tumor (chapter 3). Theoretical reasoning suggested that additional single-cell sequencing will always result in additional mutation discoveries, demonstrating the necessity a different approach to guide judgments of whether sufficiently many tumor cells were sequenced. To test computational methods for inferring ploidy from single-cell whole genome sequencing data, estimates were compared with fluorescence-based

measurements of DNA content (chapter 4). Previously proposed methods for quantum estimation were found to correctly infer ploidy from most cells, enabling inference of precise copy number in copy number aberrations. Additionally, a weighting procedure based on a probabilistic model of sequencing read counts (described in chapter 3) reduced the error rate of ploidy inference in high-ploidy samples. The lessons learned and methodology proposed in this work may be useful in research and clinical applications of single-cell DNA sequencing.

Acknowledgments

This research would not have been possible without my advisor and my advisory committee. I would also like to thank my friends who helped me plan the research and encouraged me to complete it, especially Lexy Plumer, Charissa Kim, Tessa Tsai, and Aislyn Schalck. I am also grateful to the staff of the Genetics Department and the GSBS for their hard work and patience.

Table of Contents

List of Illustrations

List of Tables

1. Introduction

A tumor consists of the cancer patient's own cells, and shares almost all of their DNA sequence. The median number of mutations in the protein-coding regions of a breast tumor genome is only 36, and even a lung adenocarcinoma, with its genome likely damaged by smoking, has a median of only 240 mutations in protein-coding regions, according to data from The Cancer Genome Atlas (Kandoth et al. 2013). This is much less than the difference between two individuals: an individual's genome has over twenty thousand variants in protein-coding regions which are not present in the reference genome, according to the 1000 Genomes Project data (Auton et al. 2015). The tumor cells are therefore recognizably the patient's own. Nevertheless, these mutations are enough to break down the systems maintaining homeostasis of cell number, and enable uncontrolled growth of the tumor.

In almost all tumors, the cancer cells are descended from a single initiating cell, and consequently share all the mutations which were present in that initiating cell. Evidence for the single-cell origin of tumors has come from multiregion sequencing studies, in which some mutations were found to be shared throughout the tumor in each of 50 breast cancers (Yates et al. 2015), 44 melanomas (Shi et al. 2014), and 100 lung tumors (Jamal-Hanjani et al. 2017). However, despite these shared mutations inherited from the tumor initiating cell, the tumor cells do not have exactly the same mutations. The mutations that differ between tumor cells constitute genetic intratumor heterogeneity, and the mutations that are not present in all cells are called "subclonal".

Subclonal mutations constitute the majority of somatic mutations in some tumors, and a small minority in others. However, just as it takes a relatively small number of mutations to make cells cancerous, even a few subclonal mutations may make the cells that carry them more malignant than the rest of the tumor, by causing more invasive behavior or resistance to chemotherapy or radiation treatment. After treatment, these resistant cells tend to be the ones that survive, and new tumors that form from them will not respond to the original treatment. In this process, called "clonal evolution" (Nowell 1976), the patient's disease changes over time due to the changes in frequency of subclonal mutations. In clonal evolution, intratumor heterogeneity plays a central role in the processes of tumor progression and treatment resistance. The work in this dissertation is aimed towards addressing challenges in the statistical analysis of sequencing data to obtain information about intratumor heterogeneity.

Techniques for studying genetic intratumor heterogeneity

Early research on intratumor heterogeneity used flow cytometry, which could detect cell subpopulations that differed in ploidy (Ewers et al. 1984; Hansen et al. 1980; Hedley et al. 1985), and karyotyping, which could detect subclones with large-scale chromosomal aberrations (Heppner 1984). Heterogeneity of copy number of individual genes could be detected using fluorescence *in situ* hybridization, a technique in which fluorescent probes bind to individual genes, and the number of copies of the gene in an individual cell is counted under a microscope. For example, an early FISH study detected, in some breast tumors, cell subpopulations with dozens of copies of ERBB2 (Kallioniemi et al. 1992). However, in a single-gene study, it is difficult to tell whether the cells with the ERBB2

amplifications constitute a subclone, since the amplification may have occurred multiple times independently. Resolving subclones and their interrelationships can be accomplished with multiplexed FISH (Chowdhury et al. 2013). For example, multiplexed FISH with probes for eight different genes was used to quantify subclonal diversity in breast cancer, and show that aneuploid tumors had more intratumor heterogeneity (Oltmann et al. 2018).

Instead of studying coarse cytogenetic features or a few target genes, comprehensive and high-resolution evaluation of intratumor heterogeneity is now possible using next-generation sequencing. One method to observe genetic intratumor heterogeneity is multiregion sequencing: cutting the tumors into pieces and preparing DNA sequencing libraries from each piece. For example, an early multiregion sequencing study of two pancreatic cancer patients found that metastatic tumors originated from region-specific subclones of the primary tumor (Yachida et al. 2010). In another early study of two kidney cancer cases (Gerlinger et al. 2012), evidence of convergent evolution was found: the same gene could have different loss-of-function mutations in different regions of the tumor, an observation which was confirmed in multiregion sequencing studies of larger cohorts of kidney cancer patients (Gerlinger et al. 2014; Turajlic et al. 2018). Since these early studies, much larger cohorts are being studied with multiregion sequencing. The TracerX lung clinical trial aims to apply multiregion sequencing to 824 lung cancer patients (Jamal-Hanjani et al. 2014), and the TracerX Renal trial aims to enroll 320 patients (Turajlic and Swanton 2017). These and other large multiregion sequencing studies promise to provide valuable information about intratumor heterogeneity and its correlation with clinical outcome. However, the method has the disadvantage that regions of the tumor are themselves heterogeneous, and naive analysis which treats genotypes of regions as if they

are genotypes of subclones produces misleading results (Alves et al. 2017). Another approach is deep sequencing of a sequencing library from a tumor, thereby detecting mutations which are present at low frequency in the library. Deep sequencing was first applied to single genes (Campbell et al. 2008), but later used to determine the frequencies of mutations throughout the exome (Shah et al. 2012). Deep sequencing requires clustering of mutation frequencies to determine which mutations are in the same cells as each other (Roth et al. 2014), making it difficult to resolve subclones which are present at similar frequencies. Though both multiregion sequencing and deep sequencing have disadvantages for observing intratumor heterogeneity, they are complementary. What appears to be one subclone when a sample is analyzed individually can be revealed as two separate subclones when sequencing data from two different regions are analyzed together (Sun et al. 2017).

Single-cell DNA sequencing is a technique which reveals the genotypes of individual cells. Single-cell exome sequencing enables detection of single-nucleotide variants (SNVs) and indels in individual cells, but the studies to date have sampled only a few patients per study, and sequenced less than one hundred tumor cells per patient (Bryant et al. 2018; Hou et al. 2012; Li et al. 2017; Peng et al. 2019; Wang et al. 2014b; Wu et al. 2016; Yu et al. 2014). Single-cell whole genome sequencing (scWGS) has been scaled up much more than whole exome, and recent scWGS datasets include thousands of cells from individual samples (Andor et al. 2020; Conterno Minussi et al. n.d.; Laks et al. 2019). These high-throughput scWGS data have low coverage per cell, with only a few percent of the genome being covered, and are therefore better suited to detection of large copy number aberrations (CNAs) than to detection of SNVs. Single-cell DNA sequencing removes the problems of interpretation associated with clustering mutation frequencies, but introduces

problems of its own, due to the limited number of cells which can be sequenced from each sample, and the technical difficulties of sequencing the extremely small amount of DNA present in a single cell. Improving the interpretation of single-cell DNA sequencing data is the aim of the methods proposed in this dissertation.

Statistical analysis of single-cell DNA sequencing data

Analyzing single-cell DNA sequencing data is in many respects just like analyzing any other sequencing data. The same tools for alignment to the human reference genome can be used, and calling copy number aberrations involves segmentation tools which were previously used for microarray data (Baslan et al. 2012).

However, single-cell DNA sequencing also brings unique challenges and opportunities. Besides the opportunity to resolve intratumor heterogeneity, it also removes confounding factors presents in bulk sequencing. One confounding factor is imperfect tumor purity due to the presence of noncancerous cells in the tumor stroma. Another confounding factor is subclonal heterogeneity (Van Loo and Campbell 2012). Although the point of single-cell sequencing is to study tumor heterogeneity, a sequencing library from a single-cell does not contain a mixture of subpopulations from multiple cells the way a bulk sequencing library does.

Single-cell DNA sequencing data also brings unique challenges. One challenge is coverage nonuniformity: differences in sequencing coverage of different parts of the genome, which are exacerbated by the methods used to amplify the minute amounts of DNA present in individual cells (Zhang et al. 2015). Another challenge is incomplete sampling of the tumor population. Some subclones may not be represented due randomly

being left out of the sample of cells that was sequenced, and others may not be represented because they were not in the region of the tumor which was sampled. Both kinds of omission complicate interpretation of single-cell DNA sequencing studies of intratumor heterogeneity.

Completeness of sampling

In a single-cell DNA sequencing study of intratumor heterogeneity, the conclusion drawn may depend on the number of cells sequenced. For example, a previous single-cell sequencing study has addressed whether chemoresistant subclones observed in a tumor sample after chemotherapy were also present prior to chemotherapy (Kim et al. 2018). Whether the chemoresistant subclones are detected in the pretreatment sample may depend on the number of cells from it which are sequenced. Other studies of intratumor heterogeneity, using multiregion sequencing, have quantified the degree of diversity in a tumor, and tested for correlations between diversity and patient survival (Negrao et al. 2018; Turajlic et al. 2018). If single-cell sequencing is used instead of multiregion sequencing for such a study, the degree of diversity reported may depend on the number of cells sequenced. These issues necessitate methodology for deciding whether sufficiently many cells have been sequenced from a sample.

Single-cell sequencing requires a view of a tumor as a population of cells, and the cells which have been sequenced as a sample of individuals from that population. From that perspective, relevant methods from outside of cancer biology may help determine the number of cells which must be sequenced. If cells are viewed as belonging to mutually exclusive subclones, then the relevant methodology is that used in ecology when sampling

organisms belonging to mutually exclusive species. On the other hand, if the goal is to observe subclonal mutations, then the relevant methodology is that used in population genetics, when studying single nucleotide polymorphisms.

If it is assumed that sampling of cells is random, that there are a specific number k of subclones, and specific numbers for the frequency of each subclone are assumed, then the probability of sampling all of them can be calculated from a multinomial distribution. These are strong simplifying assumptions, and should make the problem easy, but there was a surprising lack of software for this simple calculation. For example, R has a function "pbinom", which can be used to calculate the probability of observing a single subclone, but there's no function "pmultinom" which would answer similar questions about multiple subclones. It seems that the only scientific software which implements the relevant multinomial probability calculation is Mathematica, but it is too slow for practical use. For example, for 50 cells and 7 equally frequent clones, calculating the probability of sampling at least 2 cells from each clone takes over twelve hours (using Mathematica 12.1.1.0 on a 1.6 GHz processor). Several algorithms have been proposed which scale linearly with the sample size (Ewens and Wilf 2007; Sobel and Frankowski 2004). Another algorithm for multinomial probabilities was proposed by Levin (1981), using a series expansion which should quickly converge when the number of cells sequenced is large. Levin's algorithm had apparently never been implemented in software, except a partial implementation which would calculate only the first four terms of the series (Macrae 2018). A fast implementation of Levin's algorithm, plus a GUI for using it, is the subject of chapter 2.

Such multinomial calculations are useful but planning experiments, but are of limited use in judging the completeness of an experiment which has been performed. The issue is that the multinomial calculation requires an assumption about the number of subclones, and thus cannot answer the natural question of whether, given a sample, there are additional unobserved subclones which were not represented in the sample. To answer this question, the observed sample must be used to estimate the amount of unobserved diversity. Methods to infer unobserved diversity have been developed in population genetics and ecology. For example, Carothers (1973) reports an experiment in which researchers waited at fixed spots in the city of Edinburgh and wrote down the registration numbers of taxicabs which passed by. 172 different taxicabs were observed, 116 of which were observed only once, and 48 of which were observed twice. These data were analyzed by Chao (1984), who estimated a lower bound of $172 + 116^2/(2 \times 48) \approx 312$ on the number of taxicabs in the city. In fact, the city of Edinburgh had 435 taxicabs, according to police records. Though Chao's lower bound was not tight, it still shows that data on the observed taxicabs enabled inference of at least 140 unobserved taxicabs.

Analogously, it may be possible to estimate the number of unobserved subclones, from a sample of single cells. However, a more important question is whether there are unobserved subclones which are present at sufficiently high frequency to be observed in a second sample. The reason this question is more important is that the answer can inform the decision about whether to sample additional cells. Similar questions have been addressed in population genetics. Ionita-Laza et al. (2009) attempted to estimate how many single nucleotide polymorphisms (SNPs) would be observed in the 1000 Genomes Project, on the basis of previous sequencing studies. They assumed a probability model in

which the frequency of SNP is drawn from a beta distribution, and estimated the parameters of the beta distribution based on past data. The beta distribution is justified by population genetic theory in the case of a constant population size and no selection. But in fact, the human population has been exponentially increasing, and taking this into account results in a much larger prediction of the number of unseen variants (Gravel et al. 2011).

Such mistakes due to model misspecification suggest the use of a non-parametric method which is not dependent upon a specific model. A non-parametric method which was first proposed by Good and Toulmin (1956) results in the following prediction formula:

$$\text{Predicted number of new mutations}$$

$$= \text{number of mutations observed once} \times \frac{\text{size of new sample}}{\text{size of old sample}}$$

$$- \text{number of mutations observed twice} \times \left(\frac{\text{size of new sample}}{\text{size of old sample}}\right)^2$$

$$+ \text{number of mutations observed three times} \times \left(\frac{\text{size of new sample}}{\text{size of old sample}}\right)^3 - \cdots$$

According to this formula, a mutation observed only once adds to the prediction, whereas a mutation observed twice subtracts from the prediction. This is analogous to how in the estimate of Chao (1984) of the number of taxicabs in Edinburgh, the number of taxicabs observed once was in the numerator, and the number of taxicabs observed twice was in the denominator. In both cases, the observations seen only once–the ones that just barely made it into the sample–provide the evidence for unobserved diversity. The importance of such singleton observations presents a problem for using the method with

single-cell sequencing data, since in single-cell exome sequencing, mutations observed once are mostly false positives (Wang et al. 2014b).

My efforts to adapt these methods to single-cell exome sequencing data, and use them to decide whether sequencing additional cells is required, are the subject of chapter 3, in which I propose a model-based prediction using a population genetic model appropriate for cancer cell populations (Durrett 2013), and a non-parametric prediction which does not use low-frequency mutations.

Quantum estimation for obtaining integer copy number

Detection of CNAs from scWGS data proceeds by counting the number of reads which align to small regions of the reference genome called "bins". The number of reads in a bin, called the bincount, is a sum of contributions from each copy of the corresponding portion of the genome which is present in the cell. Mathematically,

$$N_j = rC_j + \epsilon_j \qquad (1)$$

where C_j is the copy number, r is the average contribution of a single copy to the bincount, and N_j is the bincount.

To obtain estimates of the copy numbers from the observed bincounts, the bincounts are multiplied by a constant such that the average of the resulting values is equal to the average copy number, and then each value is rounded to the nearest integer. This procedure requires an estimate of average copy number, which can be obtained using measurements of DAPI fluorescence. Fluorescence measurements are available whenever

the single cells were separated using fluorescence activated cell sorting. However, when fluorescence measurements are unavailable, average copy number is unknown.

Estimating copy number when the average is not known is a statistical problem of quantum estimation. A quantum is an unknown unit, of which the data are multiples, corrupted by noise. Quantum estimation is mostly used in archaeology. For example, Hewson (1980) reported a collection of small objects from Ghana, which may have been used as weights. Their weights of the objects were found to be approximately integer multiples of 17.5 grams, suggesting that this may have been a unit of measurement used hundreds of years ago in Ghana.

In the analysis of scWGS data, the quantum is the number of reads contributed to a bincount by a single copy, which is the parameter r in (1). (1) is a linear model, and if C_j were known estimating the quantum would be a simple univariate reression problem, but since C_j are unknown it is more analogous to a mixture model with the unusual property that the mixture components are equally spaced (Broadbent 1955). DG Kendall proposed a method of quantum inference, the cosine quantogram, which he describes in an Encyclopedia of Statistical Sciences article which provides a good overview of the subject of quantum estimation (Kendall 1986). The cosine quantogram is the following functional statistic:

$$\phi(s) = \sqrt{\frac{2}{n}}\,\mathrm{Re}\left(\sum_j e^{2\pi s X_j}\right)$$

where n is the sample size, and X_j are the observed data. Re denotes the real part of a complex number. In more familiar statistical terminology, if $\psi(t)$ is the empirical characteristic function, then $\phi(s) = \sqrt{\frac{2}{n}}\mathrm{Re}(\psi(2\pi s))$. The empirical characteristic function has many other applications (Feuerverger and Mureika 1977), and can be thought of as the Fourier transform of the empirical distribution. In fact, it can be computed by applying the fast Fourier transform to a histogram.

Kendall shows that

$$E[\phi(r)] \approx \sqrt{2n}e^{-2\pi^2\sigma^2/r^2}$$

whereas, for other values s which are not close to r or integer multiples of r,

$$E[\phi(r)^2] \approx 1$$

Therefore, the quantum r can be recognized as a peak in this cosine quantogram, against a background level of 1.

Adapting these ideas to single-cell sequencing requires some consideration of the varying reliability of segment means, as well as the nature of count distributions. I describe the requisite statistical modeling of copy number data in chapter 4, and the application to inferring copy number in chapter 5.

2. Sample size calculations for single-cell sequencing experiments from a multinomial distribution

This chapter consists of the text of the following paper:

Davis, A., Gao, R., & Navin, N. E. (2019). SCOPIT: sample size calculations for single-cell sequencing experiments. BMC bioinformatics, 20(1), 566.

Background

Biological tissues consist of a heterogeneous mixture of cells, including a variety of cell types in normal tissue or subclones in tumor tissue. This heterogeneity can be resolved using single-cell DNA or RNA sequencing methods (Navin 2015,Baran-Gale2018). Single-cell sequencing studies require sufficiently many cells to be sampled so that normal cell types or cancer subclones of interest (both hereafter referred to as "subpopulations") are represented in the sample. In most studies, however, the total number of cells is determined arbitrarily by the limits of an instrumentation run, or by budget constraints, which may result in the sampling of too few or too many cells. Here, we have developed an interactive web tool, called SCOPIT (Single-Cell One-sided Probability Interactive Tool), which provides assistance for planning experiments, using calculations from a multinomial distribution.

Implementation

The first fact used for calculating multinomial probabilities is the well-known equivalence between the probability mass function of a multinomial distribution and conditional probabilities of a Poisson distribution. This equivalence was first noted, to our knowledge, by Fisher (1922).

Theorem 1: Assume that

$$N \sim \text{Multinomial}(p, n)$$

where N and p are length k vectors, and $\sum_{i=1}^{k} p_i = 1$. Also assume that

$$X_i \sim \text{Poisson}(\lambda_i)$$

for $i = 1$ to k, where $\lambda_i = \alpha p_i$ for some α. Furthermore, assume that $X_1 \ldots X_k$ are independent. Let E be a set of possible values of a random vector. Then for any event E,

$$P(N \in E) = P\left(X \in E \left| \sum_{i=1}^{k} X_i = n \right.\right)$$

The second fact is a relationship between conditional Poisson probabilities, and an expression involving the sum of truncated Poisson random variables. The following is a slight variant of a theorem due to Levin (Levin 1981).

Theorem 2: Let $X_i^{(a_i,b_i)}$ be a truncated Poisson random variable, with probability mass function

$$P(X_i^{(a_i,b_i)} = x) = P(X_i = x | a_i < X_i \leq b_i)$$

where X_i is a Poisson random variable with rate λ_i. For a vector a and b, let $X^{(a,b)}$ be the vector containing all of these truncated Poisson random variables. Let E be the set of vectors x such that $a < x_i \leq b$.

$$P\left(X \in E \left| \sum_{i=1}^{k} X_i = n \right.\right) = P\left(\sum_{i=1}^{k} X_i^{(a_i,b_i)} = n\right) \frac{\prod_{i=1}^{k} P\left(a_i < X_i \leq b_i\right)}{P\left(\sum_{i=1}^{k} X_i = n\right)}$$

Proof. By Bayes' theorem,

$$P\left(X \in E \left| \sum_{i=1}^{k} X_i = n \right.\right) = P\left(\sum_{i=1}^{k} X_i = n \middle| X \in E\right) \frac{P(X \in E)}{P\left(\sum_{i=1}^{k} X_i = n\right)}$$

Substituting $P\left(\sum_{i=1}^{k} X_i^{(a_i,b_i)} = n\right)$ for $P\left(\sum_{i=1}^{k} X_i = n \middle| X \in E\right)$ and $\prod_{i=1}^{k} P\left(a_i < X_i \leq b_i\right)$ for $P(X \in E)$ yields the theorem.

This theorem enables a fast calculation of the multinomial probability. The rate-limiting step is calculation of the probability distribution of $\sum_{i=1}^{k} X_i^{(a_i,b_i)}$. Levin (Levin 1981) provided two suggestions for computing this probability distribution: the first by convolution of the distributions of each $X_i^{(a_i,b_i)}$, and the second using an Edgeworth expansion of the probability distribution of $\sum_{i=1}^{k} X_i^{(a_i,b_i)}$. We implemented both suggestions, which are used for different values of n. For small values of n, convolution is performed,

using The Fastest Fourier Transform In The West (Frigo and Johnson 2005). For large values of n, an Edgeworth expansion is used. However, whereas Levin (Levin 1981) used the first four terms in the expansion, we continue adding terms until the last term added is sufficiently small.

SCOPIT also computes Bayesian posterior probability distributions for the multinomial probabilities. The multinomial probabilities described above are a function of the population frequencies. When the true population frequencies are not known, but observed frequencies from a previous experiment are available, SCOPIT computes a posterior distribution for the frequencies. The prior used for the frequencies is Dirichlet(0, … ,0), following Jaynes and Bretthorst (2003) for an experiment in which the possible outcomes are not known in advance. The resulting posterior is Dirichlet($n_1, \dots, n_k$), where n_i is the number of cells observed from population i. Possible frequency vectors are randomly drawn from this posterior using the R package rBeta2009 (Cheng et al. 2012; Hung et al. 2011). Then, the desired multinomial probability is calculated from each sampled frequency vector, resulting in samples from the posterior distribution of possible multinomial probabilities. A posterior distribution over the number of cells required is calculated in the same way.

Results

Estimating required sample size using the multinomial distribution.

We make the simplifying assumption that a successful experiment requires sampling a sufficient number of representatives from each subpopulation of interest in the tissue.

Defining c as the required number of representatives from each subpopulation, N_i as the number of cells of subpopulation i which are sampled, and k as the number of subpopulations of interest, then the probability of meeting this condition is

$$P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c)$$

Assuming that a fixed number of cells are chosen at random from the population, the distribution of $N_1, \ldots, N_k$ is multinomial. To calculate this probability, we created an R implementation of a previously described algorithm (Levin 1981), described further in the Implementation section. Our implementation is available for R scripting in the package "pmultinom", available from CRAN (Table 1).

Table 1: R functions in the package "pmultinom" for calculating multinomial probabilities

Function	Arguments	Description
pmultinom	lower, upper, size, probs, method	Probability that a multinomial random vector is elementwise greater than "lower" and elementwise less than or equal to "upper". "size" and "probs" specify the parameters of the multinomial distribution. Either "lower" or "upper" may be left unspecified.
invert.pmultinom	lower, upper, probs, target.prob, method	Returns the "size" parameter required for pmultinom to reach the target probability "target.prob".

Our web tool, SCOPIT, provides an interactive interface for multinomial calculations. SCOPIT provides both prospective and retrospective calculations, described below.

Prospective calculations.

SCOPIT's prospective mode is intended to estimate the number of cells that must be sampled in a single-cell sequencing experiment. Ideally, the number of cells can be decided by finding a number of cells, n^*, such that the above multinomial probability is above a specified success probability, p^*. Such a calculation would require specifying the frequency of each subpopulation of cells in the tissue, but the precise subpopulation frequencies are usually unknown before performing the experiment.

The strategy implemented in the prospective mode is to specify the frequency of the rarest subpopulations that the researcher intends to find, as well as k, the number of populations with approximately this frequency. Both numbers are relevant, since it is harder to find, for example, 10 subpopulations with frequency 1%, than it is to find only one.

The required number of cells is defined as follows:

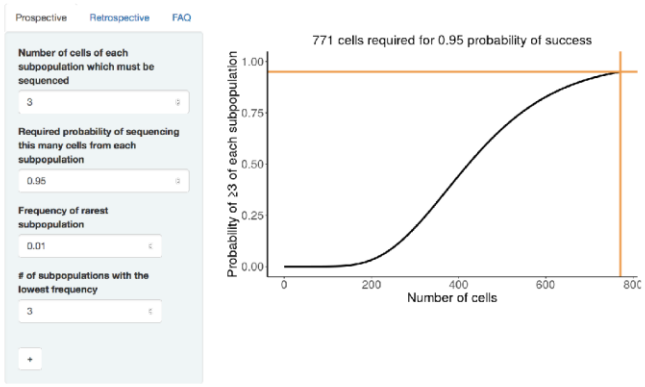
$$n^* = \min\{n \mid P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c) \geq p^*\}$$

SCOPIT reports n^* along with a plot of the probability as a function of the number of cells sequenced (Figure 1A).

This mode requires only one subpopulation frequency to be specified: the minimum frequency among all subpopulations of interest. The SCOPIT interface does enable the user to add additional subpopulations with higher frequencies, but the user will find that these

additional subpopulations have negligible effects on n^*, unless they are very close in frequency to the rarest subpopulations. This phenomenon justifies specifying only the lowest frequency.

A



B

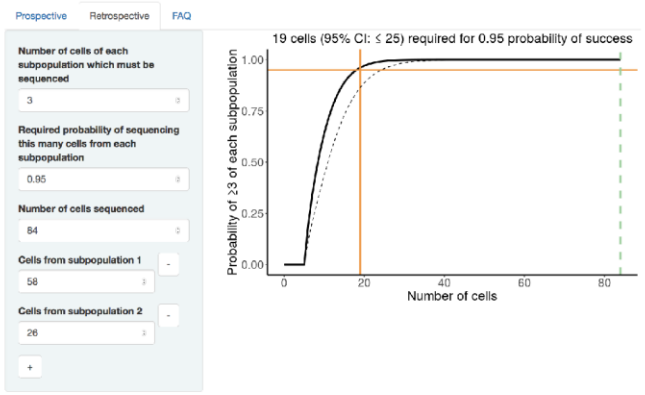


Figure 1: ***SCOPIT interface. A.*** *Interface for prospective calculations. Orange lines identify the number of cells required and the target probability of detecting a specified number of each subpopulation.* ***b.*** *Interface for retrospective calculations. The number of cells which were sequenced is entered, and is marked on the plot with a dotted green line. In this example, the orange line is far to the left of the dotted green line, suggesting that more cells were*

sequenced than required to detect these three subpopulations. To quantify confidence in the results, a dotted black line is plotted that shows the lower end of a 95% credible interval for the probability. The plot title states the upper end of a 95% credible interval for the number of cells required

Retrospective calculations

After an experiment has been performed, estimates of the subpopulation frequencies are available as input parameters. It is then possible to use SCOPIT in retrospective mode to estimate how many cells would be required, in a hypothetical replicate experiment, to detect all k observed subpopulations, with c representatives from each. In retrospective mode, the information required from the user consists of the total number of cells sequenced in a previous experiment, and the number of cells observed from each subpopulation. With this information, SCOPIT will calculate, for each number of cells n, the probability $P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c)$, assuming the true subpopulation frequencies are equal to the empirically observed ones. For example, in Figure 1B, we use single cell DNA data from a triple-negative breast tumor (Gao et al. 2016) in which the authors sequenced $N = 84$ single cells and detected two major clonal subpopulations. Using SCOPIT we estimated that only 19 cells were required to detect the two subpopulations with a 0.95 probability, suggesting that this study sequenced about 4 times the number of cells that were necessary.

Because the retrospective analysis involves uncertainty about the true frequencies of each population, SCOPIT provides measures of uncertainty using Bayesian credible intervals at a 95% confidence level. For the number of cells required, SCOPIT reports the

upper end of a one-sided credible interval, which is interpretable as the highest number of cells consistent with the data. For the probability of obtaining a sufficient number of cells from each population, SCOPIT plots the lower end of a one-sided credible interval, interpretable as the lowest probability consistent with the data. In the example described above, the credible interval boundaries were close to the estimated values, indicating that the estimated values were strongly supported by the data provided.

The retrospective tool is useful for planning a second experiment, assuming that all the subpopulations of interest were observed in the first experiment, and that the underlying subpopulation frequencies are consistent in both experiments. Although the exact subpopulation frequencies are not known, overconfident conclusions on the basis of limited information can be avoided using the credible intervals provided by the retrospective tool.

Comparison with independence approximation

Another previous software tool for estimating single cell sample sizes is an unpublished web application (https://satijalab.org/howmanycells). The previous tool is based upon two simplifying assumptions: that the subpopulations have equal frequencies, and that the observed frequencies of each subpopulation are statistically independent. Under these assumptions:

$$P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c) = P(N \geq c)^k$$

where N represents the number of cells sampled from an arbitrary subpopulation. To compare the independence approximation method to SCOPIT, the required number of cells

was calculated with and without the independence assumption (Table 2). The calculations performed under the independence assumption underestimated the required number of cells by at most 1 cell and were highly similar. These data suggests that using independence approximation is an alternative approach that can also be used for estimating single cell sample sizes.

Table 2: ***Comparison of Independent Approximation and Exact Calculations.*** *The number of cells required to achieve a 95% certainty of sampling sufficiently many cells from each subpopulation. The number of cells was calculated in two ways: by an exact calculation, and by an approximate calculation in which the counts of different subpopulations were assumed to be independent*

Subpopulation frequency	# of subpopulations	Cells required (exact)	Cells required (approx.)
0.1	6	186	186
0.2	3	85	85
0.3	2	53	53
0.1	8	191	191
0.2	4	87	87
0.4	2	39	39
0.1	9	193	193
0.3	3	55	55
0.1	10	195	194
0.2	5	89	89
0.5	2	30	30

Discussion

SCOPIT's function is to calculate the number of cells that must be sampled in a single-cell sequencing experiment, on the basis of input subpopulation frequencies, and under the assumption of random sampling. To achieve this goal, we implemented a fast multinomial probability calculation approach that is provided as open access software through the R package 'pmultinom'. This method enables calculations at speeds sufficient for interactive plotting. The retrospective sample size calculation performed by SCOPIT is distinct from estimation of the number of undiscovered subpopulations (Gotelli and Colwell 2011) or the number likely to be discovered in further sampling (Shen et al. 2003), and can instead be interpreted as the required sample size of a replicate experiment which would detect the same subpopulations as the original experiment.

To determine the number of cells required, SCOPIT calculates the probability of sampling sufficiently many representatives of each subpopulation. The probability calculated by SCOPIT is relevant to a wide variety of analyses and technologies, but specific technologies introduce additional experimental design considerations. For example, in single-cell differential expression analysis, it is important not only to sample sufficiently many cells, but also to sample sufficiently many transcripts from each cell. Other tools have been developed to calculate the probability of detecting a specific transcript (Svensson et al. 2017), to calculate the power to detect differential expression (Jenkins et al. 2018), and to determine the number of cells and reads required to find accurate low-dimensional representations of single-cell RNA sequencing data (Svensson et al. 2019). Accommodating

the unique aspects of other technologies and analyses is an important topic for future research in the design of single-cell sequencing experiments.

A previous tool is available for calculating the number of cells to sequence (https://satijalab.org/howmanycells) and a direct comparison to SCOPIT shows that it generates results that are highly similar to SCOPIT, despite using independent approximations instead of exact probabilities. However SCOPIT offers several additional features, including the ability to enter multiple cell type frequencies, and interfaces to perform both prospective estimates of the sample sizes for planning experiments and retrospective calculations which include measures of confidence in the result.

While SCOPIT can be used to decide how many cells to sample from a tissue, another important question is how many spatial regions to sample to capture the diversity of the population. In the case of sampling from tumor tissue, the question of how widely to sample can be addressed by simulating the generation of intratumor heterogeneity (Sun et al. 2017), followed by simulating sampling. However, simpler statistical calculations which avoid detailed simulations are currently not available and represent an important future direction.

Conclusions

This study reports a useful tool for estimating sample size calculations for planning single cell sequencing experiments prospectively and retrospectively. We expect that SCOPIT will have applications in many diverse areas of biology, and for planning experiments on a variety of single cell technologies (scDNA, scRNA and scATAC-seq).

3. Predicting Mutation Discoveries

Introduction

In single-cell sequencing of cancer, a sample of cells are taken from a tumor to detect the subclonal mutations in each individual cell. In analysis of single-cell exome sequencing data, mutations will be removed from the analysis if such mutations are present in too few cells. The cutoff to include a mutation has varied from two cells (Wang et al. 2014b) to five cells(Hou et al. 2012) in the literature. The practice of removing mutations which do not meet the cutoff is called "consensus filtering." Consensus filtering is necessary because many apparent mutations in single-cell exome sequencing data are actually caused by inaccuracies in DNA amplification (Wang et al. 2014b).

Many single-cell sequencing protocols begin with preparing a suspension of cell nuclei from the tumor. Once single-cell sequencing data has been obtained, the option remains of sequencing additional cells from the same nuclear suspension. Whether expanding the sample size in this way is worth it depends on the number of new subclonal mutations that would be discovered in the second sample. Therefore it is important to be able to predict the number of such discoveries, using the data from the first sample. Let c, for "cutoff", be the number of observations of a mutation required for it to pass consensus filtering and be included in data analysis. A "discovery" is a mutation which is in fewer than c cells in the first sample, but c or more cells in the expanded sample, where c is the chosen cutoff. Mathematically, what must be predicted is

$$\Delta_c = |\{j: X_j < c \text{ and } X_j + Y_j \geq c\}| \qquad (2)$$

where Δ_c is the number of discoveries, j is the index of a mutation, X_j is the number of cells with the mutation in the first sample, and Y_j is the number of cells with the mutation in the second sample.

The problem of predicting discoveries has been addressed before, for other applications. The classic applications are related to quantifying biodiversity (Fisher et al. 1943) and vocabulary size. For example, Efron and Thisted (1976) predicted the number of words which would be found in a hypothetical undiscovered work of Shakespeare, but which are not present in his known works. Methods for the problem can be broadly divided into parametric and non-parametric methods. The early work of Fisher et al. (1943) assumed a parametric model of species frequencies. The first non-parametric method was proposed by Good and Toulmin (1956), whose prediction of the number of discoveries is a polynomial function of the size of the future sample. This polynomial served as the starting point for two other non-parametric methods: Orlitsky et al. (2016) weighted the terms of the polynomial to obtain a more stable estimate, and Daley and Smith (2013) constructed a rational function whose Taylor coefficients match the non-zero polynomial terms. An independent approach to non-parametric prediction was proposed by Shen et al. (2003), who built on existing non-parametric estimators of species coverage and species richness. A Bayesian non-parametric method has also been proposed, which can be considered to be in a third category between parametric and non-parametric, since it does not assume a specific distribution of species frequencies, but does assume that these frequencies are generated by a specific stochastic process, namely a two-parameter Poisson-Dirichlet process (Favaro et al. 2009).

These methods for predicting the number of discoveries make the assumption that a discovery requires only a single observation. In the present notation, this is the $c = 1$ case. However, methods for predicting discoveries when consensus filtering is applied (the $c > 1$ case) are needed for predicting mutation discoveries in single-cell DNA sequencing. Very recently, Deng et al. (2020) addressed this challenge with a generalization of the rational function method of Daley and Smith (2013).

Two new prediction formulas for the $c > 1$ case are presented below. The first prediction formula is a polynomial which generalizes the non-parametric method of Good and Toulmin (1956). The other is based on a population genetic model of the cells in the tumor, which implies a specific distribution of mutation frequencies. Since these formulas accommodate consensus filtering, they can be used for predicting the number of mutations which will be discovered in additional single-cell sequencing from a previously sampled tumor.

The non-parametric prediction

In the $c = 1$ case, it is possible to predict discoveries without any assumptions about the unknown mutation frequencies, using a formula derived by Good and Toulmin (1956). The method of Good and Toulmin (1956) is unbiased under a Poisson approximation, regardless of the distribution of mutation frequencies (Efron and Thisted 1976). However, no method has been reported which is unbiased when a cutoff is used, and $c > 1$. Mathematically, the Poisson approximation is that

$$X_j \sim \text{Poisson}(\rho_j)$$

$$Y_j \sim \text{Poisson}(t\rho_j)$$

and that they are independent. ρ_j is the expected number of mutations of type j in the first sample, and t is the ratio of the size of the second sample to the first.

In this work, a new prediction formula was derived which is unbiased under the Poisson approximation in the $c > 1$ case. It is equivalent to Good and Toulmin's formula when $c = 1$. Therefore, it can be regarded as an extension of Good and Toulmin's method which accommodates consensus filtering. The formula for the prediction will follow directly from the following theorem:

Theorem 3: Under the assumptions above, and defining Δ_c as in (2),

$$\mathbb{E}[\Delta_c] = -\sum_{k=c}^{\infty} \mathbb{E}\,[\eta_k] I_{-t}(k-c+1, c) \qquad (3)$$

Proof. The expected number of discoveries is

$$\sum_j \sum_{x=0}^{c-1} \mathbb{P}\left[X_j = x\right] \mathbb{P}\left[Y_j \geq c - x\right]$$

Using the Poisson assumption, replace $\mathbb{P}\left[X_j = x\right]$ with $e^{-\rho_j}\rho_j^x/x!$ and the other probability with the following power series:

$$\mathbb{P}[Y_k \geq c - x] = \sum_{k \geq c} \binom{k-x-1}{c-x-1} (-1)^{k-c} \frac{(t\rho_j)^{k-x}}{(k-x)!}$$

After combining the ρ_j terms and interchanging the order of summation, eliminate the sum over j using

$$E[\eta_k] = \sum_j e^{-\rho_j} \frac{\rho_j^k}{k!}$$

Then, use the relationship between the beta function and the gamma function to obtain

$$\sum_{k \geq c} \mathbb{E}\,[\eta_k](-1)^{k-c} \frac{1}{B(k-c+1,c)} \sum_{x=0}^{c-1} \binom{c-1}{x} \frac{t^{k-x}}{k-x} \qquad (4)$$

To prove that the coefficient on $\mathbb{E}[\eta_k]$ is the same as in (3), begin with the definition of the incomplete beta function:

$$I_{-t}(k-c+1,c) = \frac{\int_0^{-t} w^{k-c}\,(1-w)^{c-1} dw}{B(k-c+1,c)} \qquad (5)$$

The values at $t = 0$ are zero in both (4) and (5). Expanding (5) with the binomial theorem reveals that the derivatives with respect to t are equal as well. QED

Replacing $\mathbb{E}[\eta_k]$ with η_k in (3) yields the following unbiased estimate of $\mathbb{E}[\Delta_c]$:

$$\hat{\Delta}_c = -\sum_{k=c}^{\infty} \eta_k\, I_{-t}(k-c+1,c) \qquad (6)$$

The estimate $\hat{\Delta}_c$ defined in (6) is the proposed prediction of Δ_c. In the case when there is no cutoff and one observation is sufficient to detect a mutation, the prediction is

$$\hat{\Delta}_1 = -\sum_{k=1}^{\infty} \eta_k\,(-t)^k$$

This is exactly the prediction proposed by Good and Toulmin (1956).

Some basic properties of the resulting predictor are given by the following theorem:

Theorem 4: Under the definitions and assumptions above,

$$\mathbb{E}\left[\hat{\Delta}_c\right] = \mathbb{E}[\Delta_c] \qquad (7)$$

$$\mathrm{Var}(\hat{\Delta}_c) \leq \sum_{k=c}^{\infty} \mathbb{E}\left[\eta_k\right](I_{-t}(k-c+1,c))^2 \qquad (8)$$

Proof. (7) follows immediately from (6) using linearity of expectation. To obtain the inequality (8), use the fact that counts of a fixed number of elements in mutually exclusive subsets are negatively correlated. QED

This theorem shows that regardless of the distribution of frequencies of mutation types, the prediction formula is an unbiased estimator of the expected number of discoveries. Furthermore, it shows that the variance of the predictor is related to the magnitude of the coefficients of η_k in (6).

The model-based prediction

In human population genetics, the number of genetic variants which will be discovered in further sampling has been predicted from assumptions about human demographic history (Ionita-Laza et al. 2009). The method used cannot be directly applied to tumor cell populations, both because of consensus filtering, and because the assumptions about demographic history are not appropriate for tumor cell populations. In this work, a new formula was derived for predicting mutation discoveries from tumor cell populations, using a model of tumor growth. The assumptions of the model are that the tumor population grew exponentially from a single cell, that cells randomly acquire

mutations according to an unchanging mutation rate, and that mutations have no effect on the rates of cell division or death.

Under these assumptions, Durrett (2013) derived that

$$E[\eta_k] \approx n\frac{\theta}{k(k-1)} \qquad (9)$$

where n is the number of cells in the sample and θ is an unknown parameter that depends upon the rates of cell division, cell death, and mutation in the tumor. Summing (9) from $k = c$ to infinity yields a formula for the expected number of mutation discoveries in the second sample:

$$E[\Delta_c] \approx n\frac{\theta}{c-1}$$

Predicting with the formula requires estimating θ. θ was estimated by minimizing the sum of absolute deviations between the predicted and observed values of η_k. Mathematically,

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=c}^{\infty} |\eta_k - E[\eta_k]|$$

where $E[\eta_k]$ is understood to be a function of θ, as in (9), and

$$\hat{\Delta}_c = n\frac{\hat{\theta}}{c-1} \qquad (10)$$

Using this formula, it is possible to calculate a prediction of the number of mutations which will be discovered when a sample of single cells from a tumor is expanded, by estimating θ using the data from the original sample.

Example application

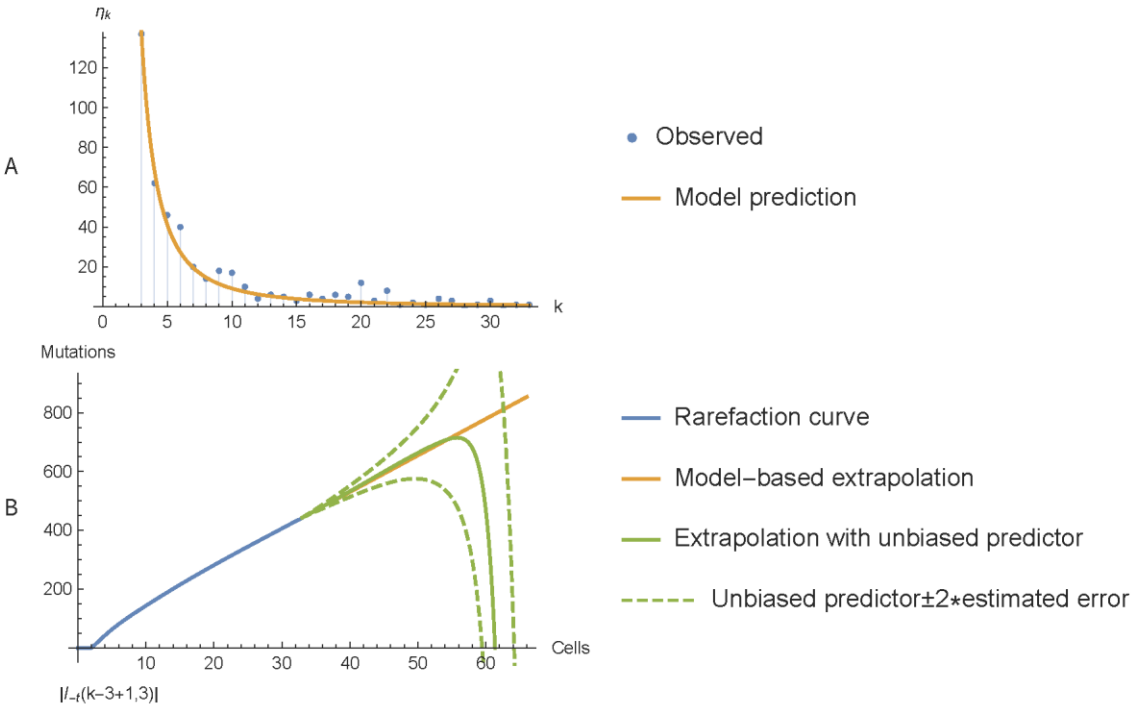


Figure 2: ***Example application.*** *A. Frequency spectrum of mutations observed in individual kidney cells, with best fit from a model of tumor growth. B. Rarefaction curve for these mutations, extrapolated using the model, and using the proposed prediction method.*

To illustrate the non-parametric and model-based predictions, single-cell sequencing data from a kidney tumor (Li et al. 2017) were analyzed. Mutations have been detected in a sample of individual cells, and the goal is to predict the number of mutations which would be detected in an expanded sample of cells. The number of mutations observed each number of times is shown in figure 2A. Mutations were only considered if they were present 3 or more times. The cutoff of 3 was chosen by the authors of the original study in order to eliminate false positives, which are abundant in single-cell sequencing data.

Using (9) it is possible to check whether the data is consistent with the assumptions. The agreement between the model fit and the data are shown in figure 2A.

2B shows rarefaction curves (Sanders 1968), extrapolated with predictions from each method, plotted as a function of the size of the expanded sample. The standard error of the non-parametric prediction, estimated with (8), begins to grow very large when the sample is expanded by 40% (47 cells total), at which point it loses log-concavity. Since this breakdown point is predictable, the prediction can still be used with confidence for expanded samples of less than 47 cells. When the original sample is expanded by 50% of its original magnitude (50 cells total), the predictions from each method are still within 1% of each other. Therefore, the biological knowledge embodied in the model was superfluous, if all that is needed is to predict the effect of supplementing the current sample with a small additional batch. However, the model is required for obtaining predictions when the second sample of cells sequenced is the same size as the first.

Discussion

The formula (6) predicts the number of mutation discoveries in an expanded sample, without relying on a model for the mutation frequencies. Predictions can therefore be made in a consistent way even for populations with different selection pressures and demographic histories, and without inferring the parameters of complex population genetic models. In all such applications the unbiasedness property (7) holds, under a Poisson approximation for mutation counts. This approximate unbiasedness property was already achieved by Good and Toulmin (1956). The novelty of the prediction formula (6) is to accommodate the common practice of "consensus filtering", ignoring mutations

observed in too few cells to be trusted. Whereas Good and Toulmin's method provided unbiased prediction of the number of mutations present in the sample, the formula (6) instead predicts the number of mutation present *c* or more times. The unbiased predictor (6) is very well suited to situations where a large sample is extended repeatedly by small batches. What counts as "small" in a given application can be determined using (8). However, the prediction fails for large expanded samples, and in fact cannot accurately predict the number of mutation discoveries when the second batch is the same size as the first.

More promising for long-term extrapolation was the model-based method, based upon a model of neutral evolution and exponential growth. However, the results show that the method cannot provide a rule for deciding whether to sequence another batch. Such a decision rule would specify that, if the number of predicted mutation discoveries is above some threshold, an additional batch should be sequenced. The reason the model-based method cannot provide such a decision rule is the linear form of the prediction function, as shown in equation (10) and figure 2B. This linear increase implies that each additional batch of cells sequenced is predicted to reveal the same number of subclonal mutations as the last. Therefore, as more cells are sequenced, the prediction cannot be expected to drop below the decision threshold.

The neutral evolution and exponential growth model, therefore, suggests that a different approach is required to obtain a decision rule. It suggests that the decision about whether to sequence additional cells cannot be based on the amount of diversity which

would be observed, but must instead be based on the value of this observing this diversity, in light of the goals of the study.

4. Probability Model for Read Counts

Introduction

Single cell whole genome sequencing using direct tagmentation is a new technique. Using these data to detect copy number aberrations requires an understanding of the probability distribution of the read counts. The results below support the modeling assumption that variance is proportional to the copy number. During quantum inference, segments are given different weights which depend on their standard error, estimated using this modeling assumption.

The constant index of dispersion model

As an illustration, bincounts from six cells are shown in Figure 3. Three are from a xenograft, with sequencing libraries prepared using a tagmentation-based protocol developed by Zahn et al. (2017). The other three are from a breast tumor, with libraries prepared by Hanghui Ye using a more recent tagmentation-based protocol (Conterno Minussi et al. n.d.).

Large segments of different bincounts are visible, corresponding to copy number aberrations. The segments appear thinner in Ye's data, showing that he achieved better coverage uniformity, although it is not a fair comparison, since Zahn's data are from his pioneering early experiments several years earlier. In all cells, the segments seem to get thicker higher up, showing that the variance of the counts increases as the copy number increases.

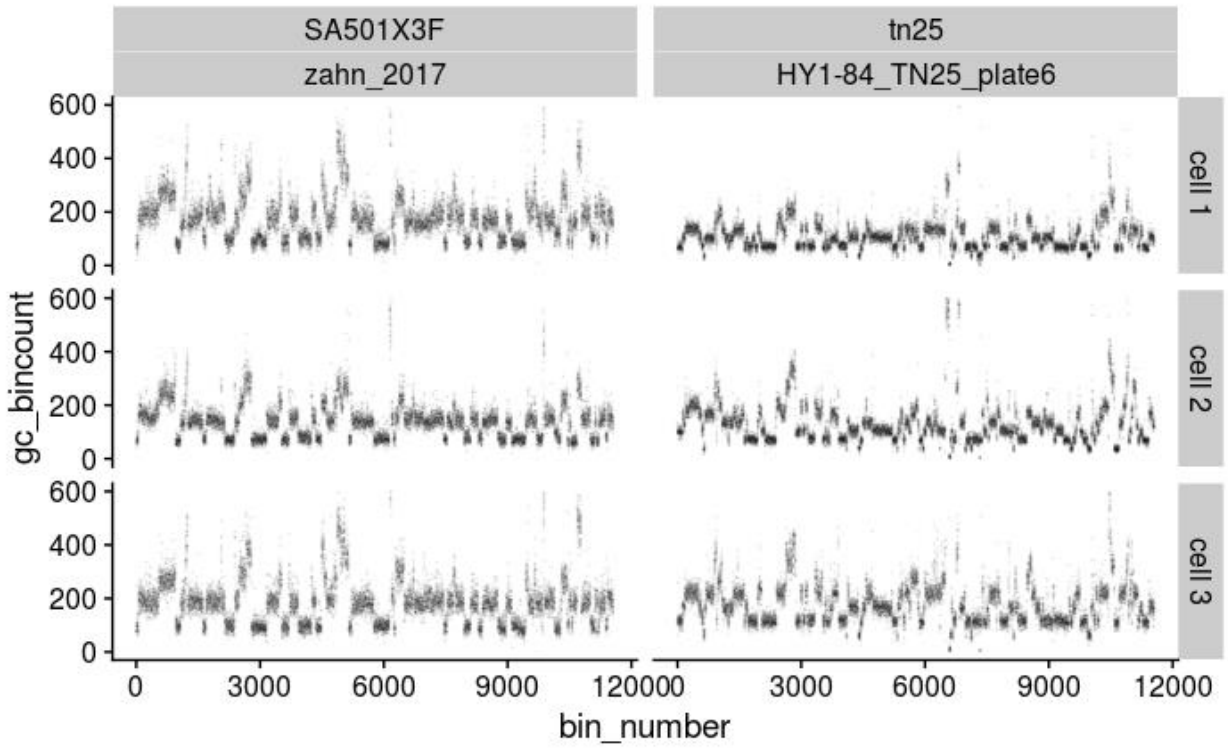


Figure 3: ***Bincounts from single cells. Left:*** *three cells from a breast cancer xenograft, sequenced by Zahn et al. (2017).* ***Right:*** *three cells from a breast tumor, sequenced by Hanghui Ye.*

Each bin represents a portion of the human reference genome, and the cell sequenced contains a certain number of copies of this portion of the reference genome in its nucleus. The increasing relationship between variance and copy number is expected, because the read count in a bin is the sum of contributions from each copy. Mathematically,

$$N_i = \sum_{k=1}^{c_i} N_{ik} \qquad (11)$$

where N_i is the number of reads in bin i, c_i is the number of copies of that bin in the cell, and N_{ik} is the number of reads aligning to that bin which originated from a specific copy k. Assuming that the contributions N_{ik} are independent and identically distributed,

$$\mathrm{Var}(N_i) = \sum_{k=1}^{c_i} \mathrm{Var}\,(N_{ik}) = c_i \sigma_1^2$$

where σ_1^2 is the variance of the contribution of a single copy. Therefore, segment variance should linearly increase with segment mean, and this relationship is observed for the six example cells (Figure 4).

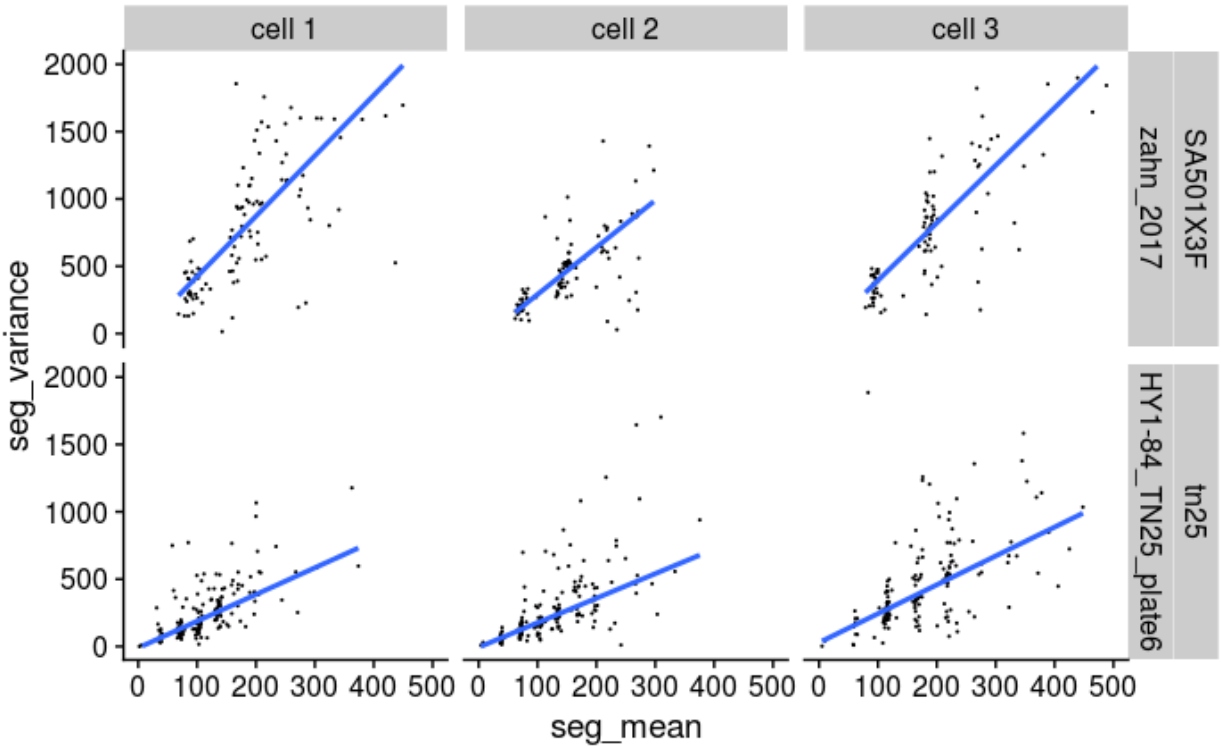


Figure 4: ***Linear relationship between segment mean and segment variance in example cells.***

The slope is higher in the cells from Zahn's early experiments, in concordance with the observation that the segments look wider in Figure 3. The slope is the ratio of the variance of the distribution to its mean, which is called the index of dispersion. The linear relationship between variance and mean suggests that the index of dispersion is a constant for a given cell, and therefore may be an effective measure of coverage nonuniformity.

Zahn's cells have higher index of dispersion, not only in these example cells, but in most cells from these samples (Figure 5).

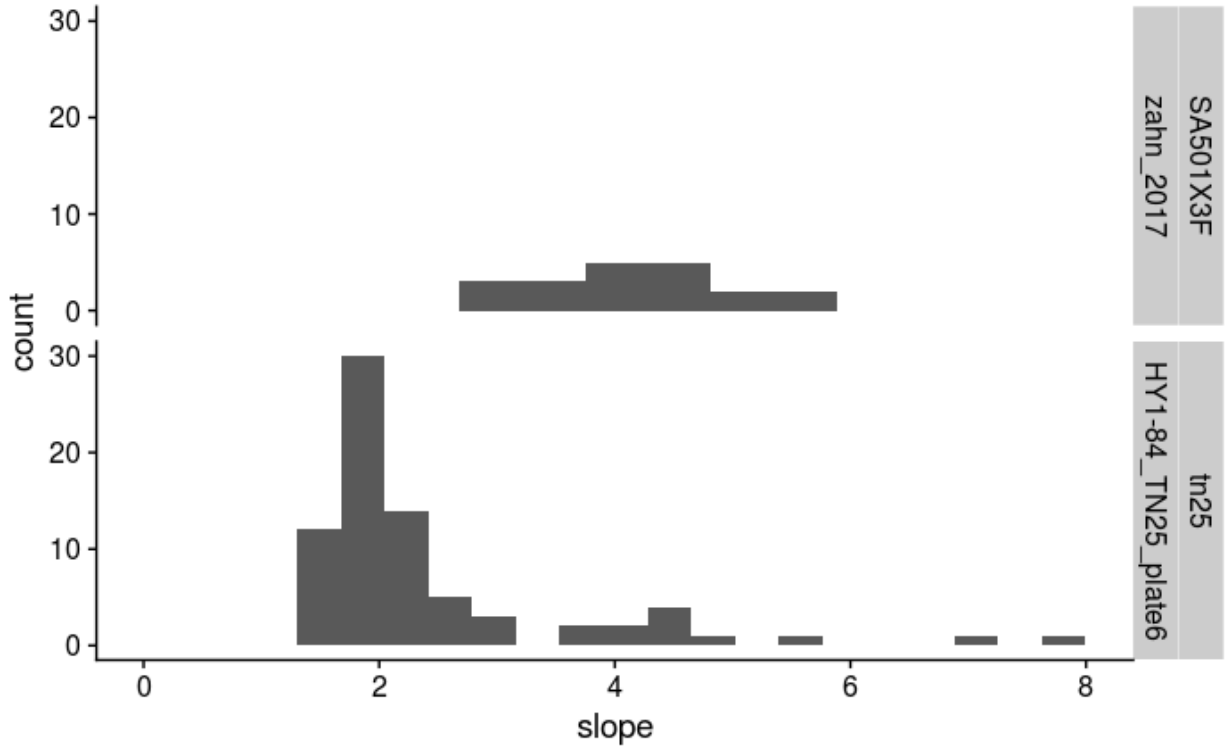


Figure 5: ***Index of dispersion for cells in two samples with different coverage nonuniformity.***

In the above, index of dispersion was estimated after segmentation, by calculating the slope of segment variance as a function of segment mean. However, in practice I estimate index of dispersion from unsegmented bincounts, prior to segmentation.

To do this, define

$$B_i = \frac{N_{i+1} - N_i}{N_{i+1} + N_i}$$

where N_i is the count in bin i. Since the mean of B_i is approximately 0 and the variance is approximately the index of dispersion except at a segment boundary, the index of dispersion can be recovered with a robust estimator of standard deviation. To obtain a robust estimate of the standard deviation of the B_i values, I use L2E to fit a normal distribution (Scott 2001).

Weighted quantogram

The above observations about the distribution of the data also have implications for quantum estimation.

Define the reads per copy, r, as the mean of the contribution of a single copy to the bincount. In the notation used in (11),

$$E[N_{ik}] = r$$

r is a quantum, and the quantum model in this case is:

$$N_i = rc_i + \epsilon_i$$

where c_j is the copy number at bin j. However, it will be more convenient to work with the normalized bincount, obtained by dividing each bincount by the overall mean bincount of the cell. Furthermore, the individual observations will be segment mean bincounts, not read counts in individual bins.

Let X_k be the normalized average bincount of a segment, μ the average bincount across bins in the cell.

By definition:

$$X_j = \overline{N}_j / \mu l$$

where $\overline{N}_j$ is the mean of the bincounts for bins in segment j. Now the quantum model is

$$X_j = mc_j + \epsilon_j$$

where $m = r/\mu$, which now serves as the quantum.

To estimate a quantum, Kendall (1986) proposed a method based on the empirical characteristic function $\widehat{\Psi}(t)$, essentially the Fourier transform of the sample distribution, defined as

$$\widehat{\Psi}(t) = \sum_j e^{itX_j}$$

Kendall's estimate of the quantum m is equivalent to $1/\hat{m} = \arg\max_{a \le s \le b} \mathrm{Re}(\widehat{\Psi}(2\pi s))$, where $[a, b]$ is a restricted search range defined based on domain knowledge. The function being maximized, $\mathrm{Re}(\widehat{\Psi}(2\pi s))$, is equivalent up to a constant with a functional statistic that

Kendall calls a "cosine quantogram". Therefore, I refer to this as the cosine quantogram (CQ) method.

With this background, it is possible to adapt the CQ method to inferring copy numbers from single cell whole genome sequencing (scWGS) data.

First of all, since $E[X_j] = mc_j$ and $\overline{X_j} = 1$, the ploidy, defined as average copy number, is $1/m$. Inference of the quantum m can be reframed as inference of the ploidy $1/m$, which helps to define the search range. Since the search range should be the range of plausible ploidies, I made the choice to use 1 to 8, which contains the ploidy values observed in flow cytometry studies.

Kendall's method needs to be modified to use our knowledge of the different reliability of the different segment means. Taking another look at the example cells in Figure 3, it is clear that the segment means from small amplifications will not be accurate and cannot be expected to assist estimation. However, Kendall was considering cases where the observations could be considered to all have the same standard deviation. Kendall's method is based on the empirical characteristic function, which is a linear combination of random variables. I use a different linear combination, in which the weight of a segment depends on its variance. Since the variance will decrease with segment size and increase with segment mean, this weighting will have the effect of downweighting unreliable segments, such as small amplifications.

The variance of the terms of the empirical characteristic function can be derived using the observation that each cell has a characteristic index of dispersion, along with

some basic facts about the empirical characteristic function, described by Csörgö (1981) among other sources.

Define $U_j(s)$ as contribution of a single observation to the empirical characteristic function $\widehat{\Psi}(2\pi s)$:

$$U_j(s) = \exp(i2\pi s X_j)$$

And its mean is given by Ψ_{X_j}, the characteristic function of X_j:

$$E[U_j(s)] = E[e^{i2\pi s X_j}] = \Psi(2\pi s)$$

Then, assuming Gaussian noise,

$$\mathrm{Var}(U_j(s)) = 1 - |\Psi(2\pi s)|^2 = 1 - e^{-4\pi^2\sigma_j^2 s^2}$$

The weighted version of $\widehat{\Psi}(2\pi s)$ would be

$$G(s) = \sum_j w_j\,(s) U_j(s)$$

Using weights proportional to $E[U_j]/\mathrm{Var}(U_j)$ yields

$$w_j(s) = \frac{1/(1 - e^{-4\pi^2\sigma_j^2 s^2})}{\sum_{j'} 1\,/(1 - e^{-4\pi^2\sigma_{j'}^2 s^2})}$$

According to the constant index of dispersion model proposed above, the variance of a segment mean is

$$\sigma_j^2 = \frac{\alpha}{l\mu} E[X_j]$$

where α is the index of dispersion and l is the number of bins in the segment.

Now, the ploidy estimate can be defined as $1/\hat{m} = \arg\max_{s} \mathrm{Re}(G(s))$.

I call this weighted analogue of the CQ method the weighted cosine quantogram (WCQ) method. The functional statistic $G(s)$ will be referred to as the weighted cosine quantogram.

5. Absolute Copy Number Inference from Single-Cell DNA Sequencing Data from Human Tumors

Introduction

Instability of genetic copy number is a common feature of human cancers, and likely plays a role in enabling tumor progression (Hanahan and Weinberg 2011). Copy number aberrations (CNAs) include gains and losses of individual chromosomes or chromosome arms, as well as focal amplifications and deletions. Additionally, many tumors also have abnormal overall DNA content, with ploidy that varies widely from less than 2N to more than 5N (Ewers et al. 1984; Hedley et al. 1985), possibly due to unequal cell divisions or endoreduplication. Copy number aberrations were originally detected with cytogenetic methods including fluorescence in situ hybridization (FISH) (Kallioniemi et al. 1992), which suffered from a limited number of sites which could be simultaneously quantified (Oltmann et al. 2018), and spectral karyotyping (Schröck et al. 1997), which suffered from low resolution. The adoption of microarrays (Pollack et al. 1999) and next-generation sequencing (Castle et al. 2010; Hayes et al. 2013) enabled megabase-level resolution of CNAs genomewide, but unlike cytogenetic methods represented an average of a large number of cells. CNAs also vary among individual cells of the same tumor, and such subclonal CNAs have been measured using single-cell whole genome sequencing (scWGS) in order to resolve subclones of tumor cells in studies of progression to invasive breast cancer (Casasent et al. 2018; Martelotto et al. 2017), as well as in the development of chemoresistance in breast cancer (Kim et al. 2018; Su et al. 2019).

Sequencing provides estimates of relative copy number, not absolute copy number, and converting to absolute copy numbers requires an estimate of the ploidy, defined here as the average copy number throughout the genome. Next-generation sequencing data produces millions of short reads, which are then aligned to the reference genome. Calling CNAs is done by dividing the reference genome into bins, and counting the number of reads that align to each bin. The bins are then joined into segments, which are sets of consecutive bins estimated to have the same underlying copy number. Normalizing the number of reads in a segment to the total number of reads sequenced yields a normalized read count, which is taken as an estimate of relative copy number: the copy number divided by the average copy number. Multiplying by the ploidy and rounding yields an estimate of absolute copy number. The statistical model underlying this estimation is

$$X_j = mC_j + \epsilon_j \qquad (12)$$

where X_j is a normalized read count of a segment of the reference genome, C_j is the number of copies of this segment in the cell, and ϵ_j is random noise. Equation (12) a "quantum model", meaning that the data are approximate multiples of unknown integers (Kendall 1986). m, called a "quantum", is an unknown parameter that represents the average contribution of a single copy to the normalized read count. m can be estimated using any experiment which measures the ploidy of the cell, because the ploidy is equal to $1/m$. There also exist techniques to estimate a quantum directly from the observed data, and outside of genomics, such quantum estimation techniques are often used in archaeology, where the quantum is an ancient unit of length (Cox 2009) or weight (Hewson 1980). In whole genome sequencing, the problem is more complicated, since it is also necessary to take into

account tumor purity and subclonal heterogeneity (Ha et al. 2014; Van Loo et al. 2010). However, in single-cell sequencing, such confounding factors are absent.

Fluorescence activated cell sorting (FACS) was used to isolate single cells in early single cell DNA sequencing methods, including DOP-PCR (Navin et al. 2011) and MDA (Wang et al. 2014a), and ploidy estimates could be obtained from the FACS data. More recent single-cell DNA sequencing methods have the advantage of being high-throughput and capable of profiling thousands of cells in parallel, but some are unable to measure ploidy of the sequenced cells, including droplet-based systems (Andor et al. 2020) . It is also not feasible to infer the quantum using a diploid cell as a reference, since the diploid cell would go through library preparation separately, and the contribution of a single copy to the read count would be different. Since in these techniques direct measurements of the ploidy or the quantum are not available, estimation of absolute copy number requires inferring ploidy directly from the data using quantum estimation. The primary techniques for quantum estimation are the least squares method (Broadbent 1955, 1956), and the cosine quantogram method (Kendall 1986), which involves taking a Fourier transform of the data distribution. The least squares method is used for estimating ploidy in the scWGS processing pipeline Ginkgo (Garvin et al. 2015). In another processing pipeline, SCOPE (Wang et al. 2020), least squares is used to provide an initial estimate, which is then refined by local optimization using expectation maximization. A similar criterion to least squares, but using log distance to rounded values than squared distance, is used by Laks et al. (2019) to choose settings for HMMCopy, a copy number inference method based on hidden Markov models. These examples show that quantum estimation methods an essential part

of scWGS pipelines, and that even in sophisticated probabilistic methods, simple methods such as least squares play an important role as subroutines.

Considering the importance of such quantum estimation methods in interpreting scWGS data, it is important understand their performance, and the circumstances under which they can be expected to work. The first progress on this question has been made in an extensive simulation study (Fan et al. 2019), which found effects of ploidy, coverage depth, and coverage nonuniformity on copy number inference. However, there is no information available about the effectiveness of quantum estimation in real scWGS data, the conditions required for its success, and whether these conditions are met in practice. In this work, performance of quantum estimation in practice is measured, by comparing ploidy estimated from scWGS data to experimentally measured ploidy for the same cell. Furthermore, a novel method, weighted cosine quantograms (WCQ), was developed by weighting the terms in the cosine quantogram to account for the heteroskedasticity of segment means.

Results

The proposed procedure for estimating ploidy and integer copy numbers from scWGS data is illustrated in figure 6a. The input are normalized read counts from a scWGS experiment. The first step is segmentation, and estimation of the mean of each segment and its standard error. The segment means have a multimodal distribution which resembles a periodic signal, and a weighted Fourier transform of this distribution is calculated. The real part of the weighted Fourier transform is called the weighted cosine quantogram (WCQ). The location of the peak of the WCQ is the estimate of ploidy. Then, relative copy numbers

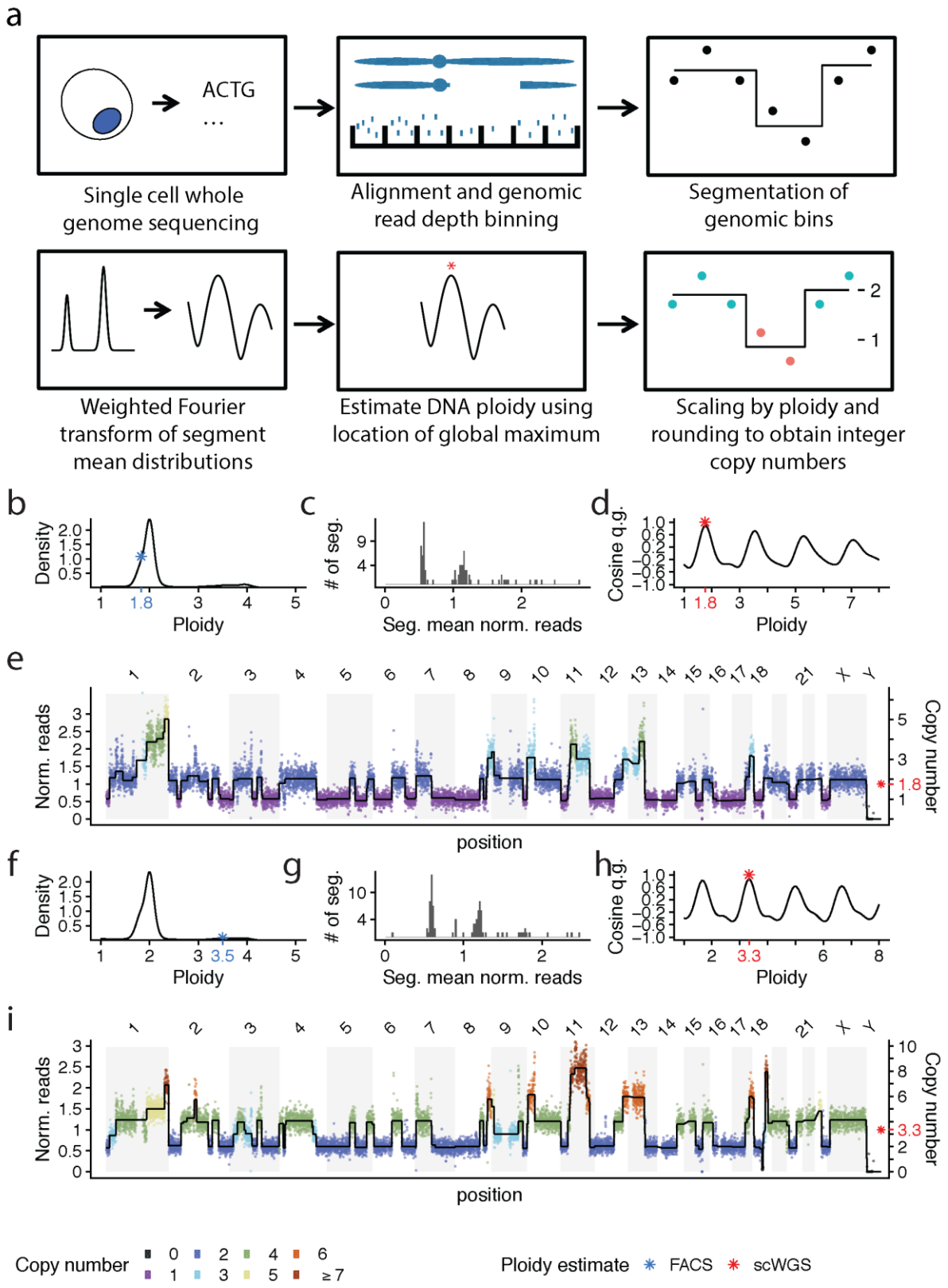

a
ACTG
...
Single cell whole genome sequencing
Alignment and genomic read depth binning
Segmentation of genomic bins
Weighted Fourier transform of segment mean distributions
Estimate DNA ploidy using location of global maximum
Scaling by ploidy and rounding to obtain integer copy numbers
- 2
- 1
b
Density
Ploidy
1.8
c
# of seg.
Seg. mean norm. reads
d
Cosine q.g.
Ploidy
1.8
e
Norm. reads
Copy number
position
1.8
f
Density
Ploidy
3.5
g
# of seg.
Seg. mean norm. reads
h
Cosine q.g.
Ploidy
3.3
i
Norm. reads
Copy number
position
3.3
Copy number 0 1 2 3 4 5 6 ≥7
Ploidy estimate FACS scWGS

Figure 6: ***Proposed workflow for copy number inference, illustrated with two cells from a breast tumor. a*** *Procedure for inference of absolute copy number from single-cell sequencing data.* ***b*** *Histogram of ploidies of cells observed during flow sorting, obtained by normalizing DAPI fluorescence. The measured ploidy of the sequenced cell is marked with a blue star.* ***c*** *The distribution of relative copy number estimates obtained from scWGS.* ***d*** *The weighted cosine quantogram of the distribution of relative copy number estimates.* ***e*** *Copy number profile obtained from scWGS. Relative copy number estimates are on the left axis, and absolute copy number estimates on the right axis.* ***f-i*** *Same as b-e, for a higher ploidy cell from the same tumor.*

are multiplied by the ploidy, and rounded to obtain estimates of absolute copy numbers. To illustrate inference of copy numbers using the weighted cosine quantogram (WCQ), two example cells from the same breast tumor are shown, which were isolated with FACS and then sequenced (figure 6b-d). The first cell was observed to be hypodiploid during FACS (figure 6b). After scWGS, the normalized read counts cluster around two values (figure 6c), and the WCQ has a peak at a ploidy estimate of 1.8 (figure 6d), consistent with the estimate from FACS. After converting to absolute copy numbers, the copy number profile shows that this cell, in most regions of the genome, has copy numbers of 1 or 2 (figure 6e). The second cell, according to FACS, is hypertriploid (figure 6f). In this cell, the distribution of normalized read counts has an additional mode between the two most prominent modes, which wasn't present in the hypodiploid cell (compare figure 6g with c). Consequently, the WCQ's highest peak is now at 3.4 (figure 6h). The hypertriploid cell has a similar copy number profile to the hypodiploid cell, but the most common copy number states are now 2 and 4, with just a few regions at copy number 3, on chromosomes 1, 3, 4 and 9 (figure 6i),

which account for the additional mode. This additional mode provides the information which the WCQ uses to infer a higher ploidy. In these two cells, the estimates of ploidy from the WCQ enable estimation of integer copy number using only the sequencing data, without requiring the FACS data.

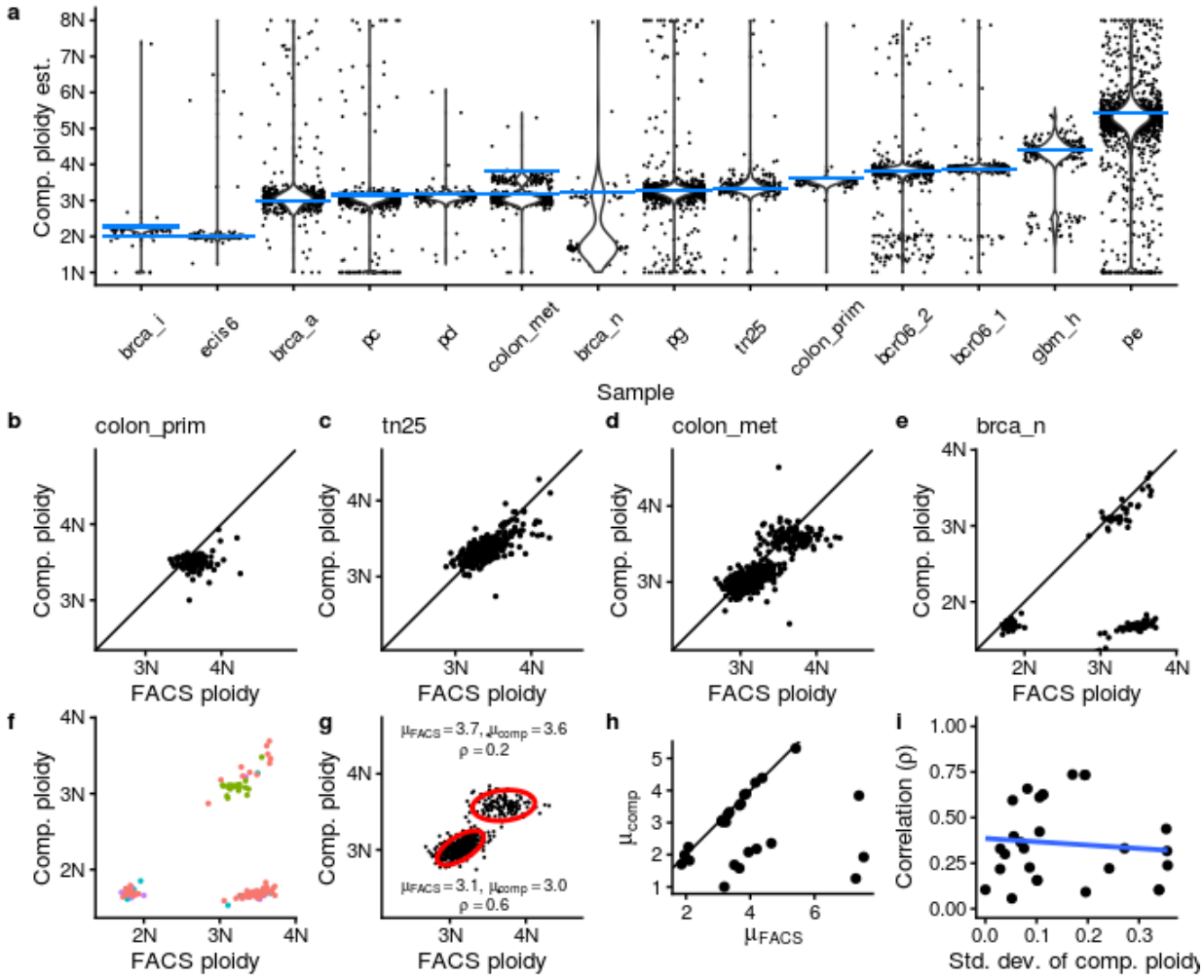


Figure 7: ***Comparisons of ploidy inferred computationally from scWGS data to ploidy inferred experimentally from FACS. a*** *Computational inferences of ploidy (dots) compared to the most common ploidy among cells in the tumor, according to FACS (blue line).* ***b-d*** *Comparison of FACS with computational ploidy for individual cells from four samples.* ***f***

Clustering of cells from brca_n. ***g*** *Illustration of fitting bivariate normal distributions, showing the 2 standard deviation contour of each distribution.* ***h*** *For each ploidy subpopulation of each sample, the mean ploidy estimates from a bivariate normal distribution fit.* ***i*** *For each ploidy subpopulation, the standard deviation of the computational ploidy estimates, and the Pearson correlation between the FACS and computational estimates, estimated from a bivariate normal distribution.*

To determine whether the WCQ ploidy estimates are reliable, the WCQ method was applied to scWGS data from individual cancer cells from 9 breast cancer patients, as well as one patient with colon cancer and one patient with brain cancer. The DAPI fluorescence values recorded during FACS were used to estimate the modes of the distribution of ploidy, and the WCQ results were compared to these modal values. The WCQ ploidy estimates of individual cells were concentrated at the modes of the ploidy distribution, except in brca_n where there was an additional concentration of values near the diploid peak (figure 7a). The average distance from WCQ ploidy estimates to the FACS mode varied between samples. To determine whether this variation is the result of error or intratumor heterogeneity, the FACS measurements of individual cells were examined next, using index sorting. Four samples are shown which are representative of the patterns in these data (figure 7b-d). In a colon tumor (colon_prim) with a very narrow distribution of WCQ ploidy estimates, although the FACS and WCQ estimates were the same on average, they were also uncorrelated, suggesting that the variation in each estimate is due to small random errors (figure 7b). However, in a breast tumor (tn25) with a wider distribution of WCQ ploidy estimates, the FACS and WCQ estimates were correlated, suggesting that the variation in both estimates reflects underlying intratumor heterogeneity of ploidy (figure 7c). In

colon_met, a liver metastasis from colon_prim, two aneuploid peaks were visible on FACS, and the FACS and WCQ estimates are generally in agreement about which ploidy population a cell came from (figure 7c). A breast tumor (brca_n) also had two populations with different ploidy, one of which was difficult to see in FACS since it overlapped with the diploid peak. In this tumor, some cells had FACS estimates of ploidy which were precisely double the WCQ estimates of ploidy (figure 7d). Though at first these appear to be errors, a clustering analysis shows that these cells have the same genotypes found in the low-ploidy population (figure 7e), suggesting that they are actually cells from the low-ploidy population in G2 phase. To summarize the results beyond these four examples, each distinct mode on a scatterplot was summarized with three numbers: its mean FACS ploidy, mean WCQ ploidy, and Pearson correlation coefficient (figure 7e). Across samples, the difference in means for a mode on a scatterplot is near zero in most cases (figure 7h), the exceptions including brca_n as previously discussed, as well as scatterplot modes from two other samples. Correlations between FACS and computational values for each scatterplot mode are shown in figure 7i.

To confirm that the variation in WCQ ploidy estimates is biological in origin, ploidy estimates were compared to copy number profiles from scWGS. In the colon metastasis (colon_met), the single-cell copy number profiles divide into several subclones with significant differences between them (figure 8f). One subclone has amplifications throughout nearly the entire genome, relative to the other two. This subclone also has the highest ploidy, according to the computational and experimental estimates, and can be identified with the right-most peak observed in flow cytometry. Within the major subclones, there are minor subclones, which also differ in ploidy. Most strikingly, the

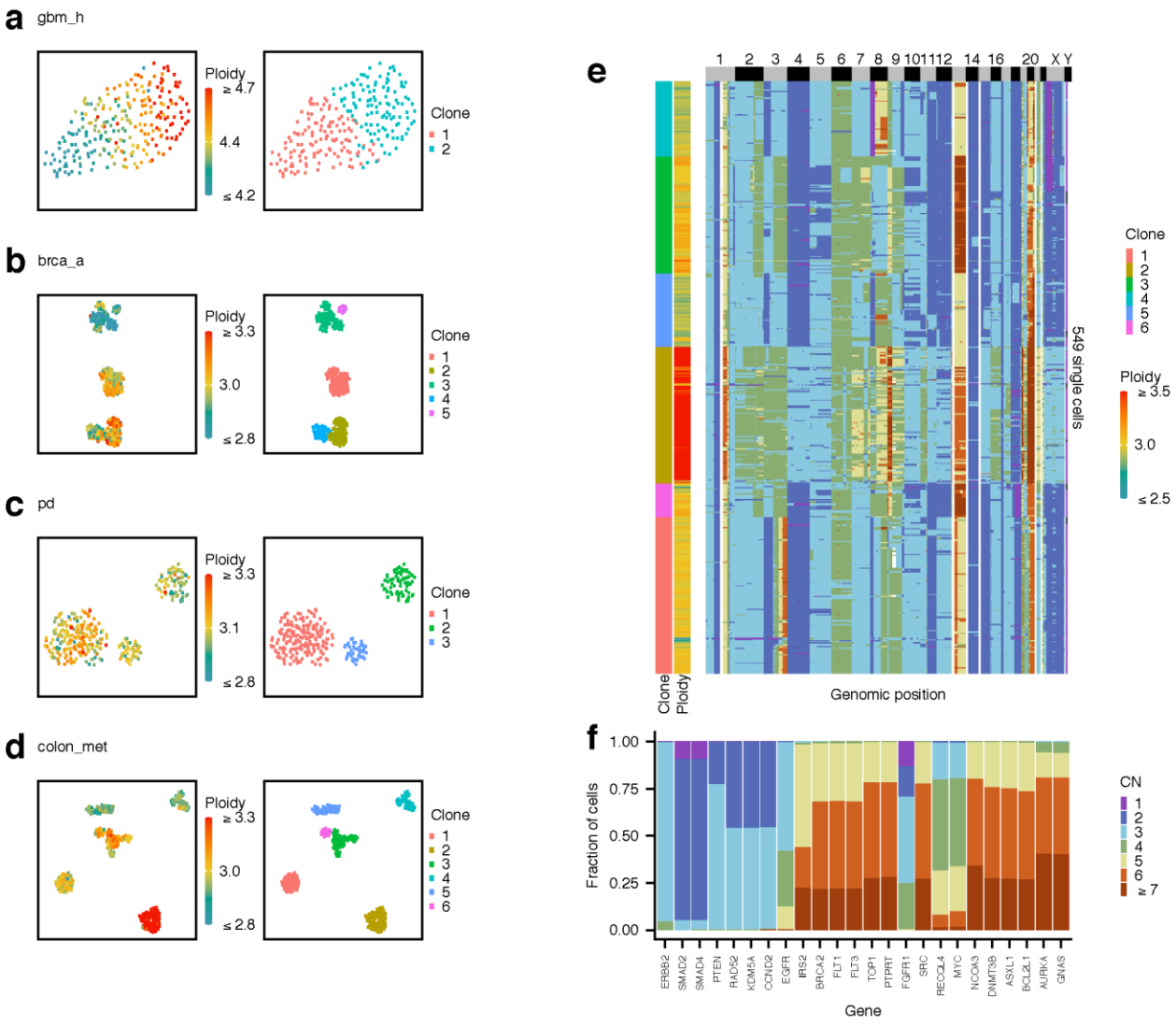


Figure 8: ***Example application. a-d*** *Dimensionality reduction labeled with ploidy.* ***e*** *Heatmap of cells from the colon metastasis colon_met labeled with ploidy and cluster.* ***f*** *Copy numbers of genes from single cells from the colon metastasis.*

highest-ploidy cells are related, and have extra copies of chromosomes 7p, 8q, and 15, even relative to other cells in the same major subclone. A similar pattern in which cells with similar genotypes have similar ploidy was found in other samples through dimensionality reduction (figure 8a-c). To illustrate the advantages of absolute copy numbers over

relative copy numbers, absolute copy numbers were estimated for each cell from colon_met. The frequency of estimated copy numbers for genes which are frequently affected by CNAs in colorectal cancer (Yaeger et al. 2018) are shown in figure 8f. Using absolute copy numbers reveals diversity among levels of deletions, as well as among levels of amplifications.

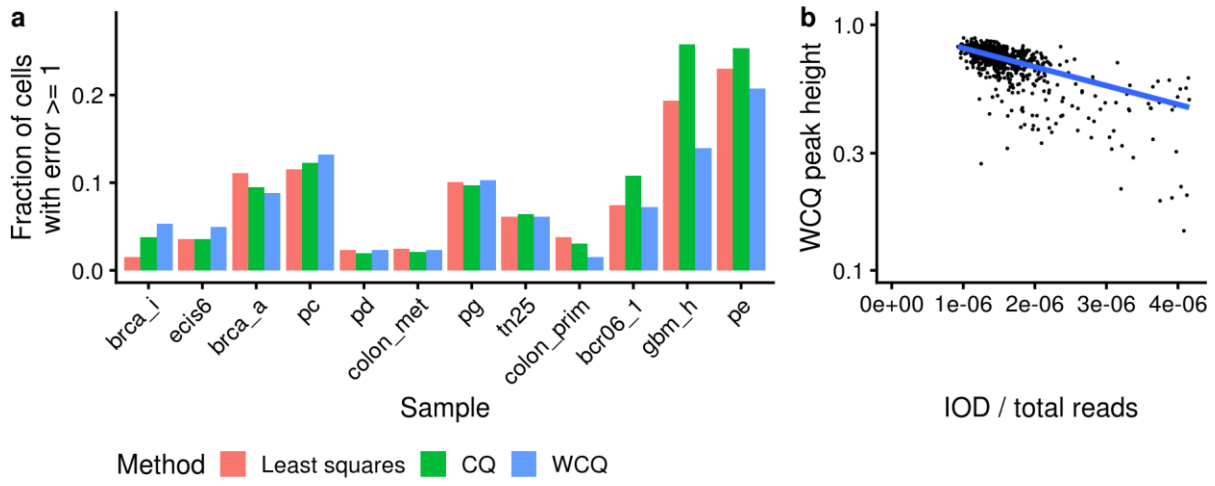


Figure 9: ***Benchmarking and trend in WCQ peak height. A*** *Fraction of cells for which the computationally estimated ploidy differed by more than 1N from the experimentally estimated ploidy, using three different methods for computational ploidy inference: least squares, unweighted cosine quantogram (CQ), and weighted cosine quantogram (WCQ).* ***B*** *In cells from the sample pc for which the difference between WCQ and FACS ploidy was less than 1N, the relationship between log WCQ peak height and the ratio of index of dispersion to reads sequenced is approximately linear.*

The WCQ differs from the least squares method used in Garvin et al. (2015) in that it weights segments according to confidence, giving the least weight to small, high-level amplifications. To determine the effects of these difference, cells with incorrect ploidy

inferences were counted, using the least squares method, as well as the unweighted cosine quantogram method proposed by Kendall (1986). An incorrect ploidy inference was defined as a difference of greater than 1 between the ploidy estimated from scWGS and the FACS estimate. Two samples (brca_n and bcr06_2) were excluded from this analysis because they may contain cells in G2 or M phase, in which case the computationally inferred and experimentally measured ploidies are not expected to match. The WCQ method made fewer errors than the least squares method in gbm_h and pe, high-ploidy samples with a high fraction of incorrect inferences, but in other samples there was no clear pattern about which method performs better (figure 9a).

To determine the causes of failures of the WCQ ploidy estimates, two measures of data quality were examined for the scWGS bincounts from each cell: the index of dispersion and the mean. The ratio of index of dispersion to reads sequenced was tested as a metric of data quality, using cells from the tumor pc for which the computational and experimental ploidy estimates differed by less than 1. It was found that the logarithm of the height of the WCQ peak was linearly related to this ratio (figure 9b). (The figure, and the linear fit, exclude 5.8% of cells because they were outside the range of the plot.) Since a shorter quantogram peak may not be distinguishable from background, and since index of dispersion and number of reads sequenced vary among cells and samples, this trend may explain some of the failures of the WCQ ploidy estimates.

Discussion

These results show that, using the weighted cosine quantogram (WCQ), integer copy numbers can be estimated from scWGS data by exploiting the periodicity of the distribution

of read counts. Through the lens of absolute copy number, a breast tumor with an apparent subclonal whole-genome doubling was studied (brca_n), as well as a colon metastasis with two populations that differ in ploidy (colon_met). Integer copy number estimation at the gene level revealed heterogeneity in levels of amplifications and deletions that may not have been apparent in relative copy numbers. Although in the cases examined in this work the absolute copy number could have been estimated from FACS data, closely matching answers were obtained using the WCQ, opening up applications to high-throughput scWGS protocols that do not involve FACS.

Integer copy number has been estimated from scWGS data in previous studies, using the least squares method, or essentially similar methods that use different measures of deviation besides squared distance. In this study, for the first time, such estimates have been systematically validated by experimental measurements of ploidy. Errors of ploidy inference were common in high-ploidy tumors, and could be reduced by weighting segments according to reliability during quantum estimation using the WCQ.

The proposed method for integer copy number inference with the WCQ operates on one cell at a time, using read counts as input. The limitations of this approach were revealed by the analysis of error rates and peak heights, which showed that difficulties emerged at low sequencing depth and poor coverage uniformity, and high ploidy. These problems may be addressed by combining data from multiple cells (increasing the effective read count) or using allele-specific copy numbers (halving the effective ploidy). Both approaches are incorporated in the recent method CHISEL (Zaccaria and Raphael 2019), which uses phylogenetic inference to structure the sharing of information between cells

and to estimate allele-specific copy numbers. However, the results above show that in many cases of practical relevance, all the information required to infer absolute copy numbers is present in the read counts of an individual cell, and that integer copy numbers can reliably be estimated as an early preprocessing step, rather than jointly with analysis of the evolutionary history of the tumor.

Methods

Data analysis tools

Data were analyzed using R (R Core Team 2019), with extensive use of dplyr (Wickham et al. 2019) and tidyr (Wickham and Henry 2019). Plots were made using ggplot2 (Wickham 2016) with the cowplot theme (Wilke 2019). Plots were combined with patchwork (Lin-Pedersen 2020).

Analysis of flow cytometry data

Fluorescence measurements from flow sorting and flow cytometry were exported in the CSV format, or in FCS format, in which case they were read using the R package flowCore (Ellis et al. 2009). Flow cytometry distributions were estimated using kernel density estimates.

Analysis of single-cell DNA sequencing data

Single-cell sequencing libraries were prepared with acoustic cell tagmentation (Conterno Minussi et al. n.d.), and sequenced using an Illumina platform to obtain short reads. Reads were aligned to the human reference genome version hg19, and duplicates

were removed using Picard MarkDuplicates. The reference genome was divided into 200kb bins, excluding regions that were not unique in 50bp windows (Derrien et al. 2012), and also excluding the ENCODE blacklisted regions. In each cell, the number of reads in each 200kb bin was counted. Then, bincounts were corrected for GC content using loess, and square root transformed to stabilize variance. The transformed bincounts were then segmented using the fused lasso algorithm of Johnson (2013) with a penalty of 25. Unless a cell had at least one segment covering at least 40 bins, with a bincount at least four thirds of the average bincount, it was assumed to be a stromal cell and detection of CNAs was not attempted. Unless a cell's segmented bincounts had Pearson correlation of at least 0.8 with at least one other cell from the same sample, it was assumed to be the result of a failed library preparation and detection of CNA's was not attempted. Segments less than 40 bins long were not used for ploidy estimation. Segment medians, rather than means, were used as input for ploidy estimation. Bivariate normal mixture models were fit to scatterplots using L2E (Scott 2004). Linear models were fit using robustbase (Maechler et al. 2019). Gene copy numbers were estimated using the estimated copy number of the segment which contains the gene.

Clustering and dimensionality reduction

For clustering and dimensionality reduction, sequencing data was processed using a different pipeline, described previously by Casasent et al. (2018). The resulting normalized bincounts were transformed with a square root, segmented using the fused lasso, and then smoothed using averages within a ten-bin sliding window. Distances between pairs of cells were calculated using Manhattan distance between the first differences of the resulting

transformed profiles. Using these distance matrices, dimensionality reduction was performed using UMAP (McInnes et al. 2018), and the UMAP graph was used for graph-based clustering with the Leiden algorithm (Traag et al. 2019).

6. Discussion

The methods described in this dissertation contribute to the statistical analysis of single cell DNA sequencing data of tumors. The first problem addressed was obtaining sufficiently many cells from each of the subclones of a tumor. I framed this problem as calculating probabilities from a multinomial distribution. An R package and a GUI for calculating probabilities from multinomial distributions were written to estimate the required number of cells. The R package was novel since no previous software provided exact multinomial probabilities and scaled to the relevant sample sizes, even though the needed algorithm was already in the literature (Levin 1981). However, the exact multinomial calculation turned out not to be necessary, since the relevant multinomial probabilities could be approximated sufficiently well by products of binomial probabilities. But even though it was not required for its original purpose of planning single cell sequencing experiments, the R package, pmultinom, has apparently filled a conspicuous void in the R ecosystem. pmultinom was downloaded 311 times per month on average in 2019, and has been used to answer a probability question on the Mathematics Stack Exchange (Lonza Leggiera 2019). The package is apparently being used despite the fact that the version currently on the R package repository, CRAN, is somewhat out of date and only implements a $O(n\log n)$ algorithm. Since the software seems useful I plan to continue to improve the package and update the version on CRAN.

After sequencing cells from a tumor, it must be decided whether sequencing additional cells from the same tumor is necessary. I framed this problem as predicting the number of subclonal mutations that would be discovered if the size of the sample of cells

was doubled. Both a non-parametric method and a model-based method were tested, but only the model-based method could make reasonable predictions when the sample size was doubled, whereas the nonparametric prediction suffered from a fast increase of variance with sample size. The model-based method was similar to a previous method applied to statistical genetics of human populations (Ionita-Laza et al. 2009), but was novel in that it used a population genetic model appropriate for tumor cell populations (Durrett 2013). However, the population genetic model had an unwelcome implication: if single cell sequencing of a tumor is conducted in sequential batches (for example, flow sorting cells into 384 well plate, sequencing them, and repeating), then each batch will reveal just as many subclonal mutations as the last batch. Therefore, predictions of the number of mutation discoveries in a hypothetical second experiment will never be able to serve as an argument that the first experiment was sufficient. Although this means that the research described here cannot meet its original goal, the result is important and worth publicizing, since it contradicts claims that have been made in the past. For example, the authors of the first single-cell exome sequencing study claimed that "statistic analysis showed that sequencing more cells would almost not increase the number of somatic mutations called from the cell population" (Hou et al. 2012). Their conclusion was based on a rarefaction curve that did not account for consensus filtering. But the neutral exponential growth model predicts that sequencing more cells would always increase the number of somatic mutations called. The claim that sequencing more cells is unnecessary will have to be made based on the specific aim of the study, and how much diversity must be observed in order to meet that aim.

In order to detect subclones in recent high-throughput single-cell whole genome sequencing (scWGS) datasets, I addressed the detection of copy number aberrations from these data, using the counts of sequencing reads in bins of the reference genome. The first step was to determine the statistical properties of the bincounts. It was found that the variance of the bincounts within a segment of the reference genome was proportional to the copy number within that segment. Furthermore, the index of dispersion varied between cells and especially between different experiments, providing a measure of the noise in the data. This simple observation will be of great value to people who are using maximum likelihood or Bayesian methods, since it can guide the choice of a probabilistic model. This model was calculating weights during quantum estimation.

To address the problem of estimating copy number from the bincounts from scWGS, I framed the problem as quantum estimation: estimation of an unknown quantum of which the data are small integer multiples. Instead of the previously used least-squares method (Baslan et al. 2012), I chose to use the "cosine quantogram" method based on the empirical characteristic function, because mathematically it was easy to analyze using previously established theory, and conceptually it had a simple interpretation as extracting the principal frequency using a Fourier transform of the distribution of the data. Another option would have been a Bayesian analysis, which has been applied before in quantum estimation from archaeology data (Freeman 1976). A Bayesian analysis would have had the advantage of easily taking into account the differences in reliability between segment means. However, staying within the framework of the cosine quantogram method, I solved this problem by weighting contributions of segment means. The mathematical tractability of the cosine quantogram method was an advantage in enabling a simple derivation of the

theoretically optimal weights. Besides being theoretically justified, the weighting procedure also turned out to be practically useful in samples like the breast tumor "pe", where the data did not determine the ploidy with perfect reliability in all cells. In these difficult samples, including the weights reduced the error rate of ploidy inference relative to an unweighted version of the method.

In order to calculate error rates of ploidy estimation, it was necessary to have some way of validating ploidy estimates. Besides testing the improvements from including weights, it was also important to test previous methods for ploidy estimation. I used experimental measurements of ploidy for individual cells made prior to sequencing, using indexed flow sorting. I found that the computational and experimental estimates of ploidy agreed in most cells in all samples, and that the weighting procedure reduced error rates. Besides testing novel methodology, I also provided the first systematic measurement of the performance of the least-squares method used by the popular software Ginkgo (Garvin et al. 2015). Ginkgo's integer copy number estimates, based on the least-squares ploidy estimates, have been used in several studies (Alexander et al. 2018; Perez-Rodriguez et al. 2019), but before this work the least-squares ploidy estimates had never been experimentally tested.

The proposed method for estimating ploidy and copy numbers is fast and easy to implement. Furthermore, the method operates on one cell at a time, using the bincount data which are already computed in data processing pipelines. This makes it easy to integrate the method into existing pipelines. However, the most challenging cases seem to be the ones with the highest ploidy. Haplotype specific read counts would solve the

problem by cutting the "effective" ploidy in two, since for example a cell with ploidy 5 should have an average copy number of around 2.5 for the maternal DNA, and likewise for the paternal DNA. It is possible that such haplotype specific read counts can be obtained, since even in a relatively small segment there are thousands of SNP sites, which can be partitioned into haplotype blocks using 1000 Genomes Project data. Estimating haplotype specific copy number would also be useful to distinguish tumor subclones which differ due to copy number neutral loss of heterozygosity. Obtaining haplotype specific copy numbers is therefore an important next step in bioinformatics processing of single cell whole genome sequencing data.

Ploidy estimates can be improved by calculating them for all cells jointly, instead of individually, borrowing strength in the estimate of each individual cell's ploidy. This can be expected to help since a stronger signal was obtained from cells from which more reads were sequenced, and considering multiple cells simultaneously is raising the effective read count. Although cells are not guaranteed to have the same ploidy, closely related cells do, so borrowing strength can be done through a clustering or a phylogenetic structure. A recent tool called CHISEL jointly estimates integer copy numbers with the phylogeny, and also attempts to calculate allele-specific copy numbers, using the phylogeny rather than the above-described strategy based on haplotypes (Zaccaria and Raphael 2019). This kind of joint inference is very different from the strategy in the weighted cosine quantogram method, which treats integer copy number estimation as an early processing step performed on data from individual cells one by one, prior to joint analyses such as phylogeny inference. Which approach will be most useful in the future depends upon how experimental methods develop. Low depth, low quality data from a large number of cells

demands a joint analysis. But, if the trend is towards getting better data from individual cells, then it will be more straightforward to calculate integer copy number as a preprocessing step like the WCQ proposed in this work.

Estimating ploidy, and using it to convert bincounts to integer copy number estimates, is important for learning about clonal evolution. For testing population genetic models, it is important to be able to obtain a frequency spectrum: for each k, the number of mutations present in k cells. In my work on judging completeness of sampling, I relied on the frequency spectrum, and therefore had to use single-cell whole exome sequencing datasets. Reliable integer copy number estimation is one step towards being able to obtain reliable frequency spectra from scWGS datasets as well. Additionally, for phylogenetic inference, obtaining integer copy numbers enables inference of a phylogenetic tree using Steiner minimum trees, which have been used previously for fluorescence *in situ* hybridization measurements of copy number (Chowdhury et al. 2013). However, a remaining problem is that copy number aberrations overlap with one another. More work is required to deconvolute a copy number profile into a list of individual, potentially overlapping events.

In this dissertation, two main contributions have been made to the statistical analysis of intratumor heterogeneity (ITH) with DNA sequencing. Evidence has been obtained that, contrary to the assumptions made in previous research, it is not possible to saturate the diversity of a tumor. Therefore, the decision that sufficiently many cells have been sequenced will have to be made in consideration of the specific hypothesis being tested. It has been found that the read count data from a single cell are sufficient to determine its

ploidy and estimate integer copy numbers. These contributions represent steps towards the goal of understanding the message about clonal evolution communicated by ITH.

Applications of this work are anticipated in the study of clonal evolution. Researchers testing population genetic models of tumor growth will benefit from the digitization of copy number aberrations accomplished by estimating integer copy numbers. Besides basic research, the connection between ITH and tumor evolution means that ITH may play a role in future cancer treatment. ITH has been shown to be a prognostic biomarker in the lung TracerX trial (Jamal-Hanjani et al. 2017), although not in TracerX renal (Turajlic et al. 2018). The TracerX studies measured ITH by using multiregion sequencing to detect region-specific CNAs. Future clinical trials may prefer to use scWGS, which will require both estimation of the number of cells required, and estimating integer copy numbers from scWGS data.

Vita

Alexander Davis was born in Livingston, New Jersey on August 20, 1990. After completing his work at Morristown High School, Morriston, New Jersey in 2008, he entered Rutgers University in New Brunswick, New Jersey. He received the degree of Bachelor of Arts with a major in biomathematics from Rutgers in May, 2013. In August of 2013 he entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

1885 El Paseo St, Apt 35311 Houston, TX, 77054

Bibliography

Alexander, J., Kendall, J., McIndoo, J., Rodgers, L., Aboukhalil, R., Levy, D., Stepansky, A., Sun, G., Chobardjiev, L., Riggs, M., Cox, H., Hakker, I., Nowak, D. G., Laze, J., Llukani, E., Srivastava, A., Gruschow, S., Yadav, S. S., Robinson, B., Atwal, G., Trotman, L. C., Lepor, H., Hicks, J., Wigler, M., and Krasnitz, A. (2018), "Utility of single-cell genomics in diagnostic evaluation of prostate cancer," *Cancer Research*, American Association for Cancer Research Inc., 78, 348–358. https://doi.org/10.1158/0008-5472.CAN-17-1138.

Alves, J. M., Prieto, T., and Posada, D. (2017), "Multiregional Tumor Trees Are Not Phylogenies," *Trends in Cancer*, Elsevier Inc., 10, e1003703. https://doi.org/10.1016/j.trecan.2017.06.004.

Andor, N., Lau, B. T., Catalanotti, C., Sathe, A., Kubit, M., Chen, J., Blaj, C., Cherry, A., Bangs, C. D., Grimes, S. M., Suarez, C. J., and Ji, H. P. (2020), "Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution," *NAR Genomics and Bioinformatics*, Oxford University Press, 2, 1–13. https://doi.org/10.1093/nargab/lqaa016.

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P.,

Resch, A. M., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Fulton, L., Fulton, R., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C. L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L. J. M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Banks, E., Bhatia, G., Angel, G. del, Genovese, G., Li, H., Kashin, S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Rausch, T., Fritz, M. H., Stütz, A. M., Beal, K., Datta, A., Herrero, J., Ritchie, G. R. S., Zerbino, D., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Herwig, R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Simpson, J. T., Shriver, M. D., Rosenfeld, J. A., Bustamante, C. D., Montgomery, S. B., De La Vega, F. M., Byrnes, J. K., Carroll,

A. W., DeGorter, M. K., Lacroute, P., Maples, B. K., Martin, A. R., Moreno-Estrada, A., Shringarpure, S. S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F. C. L., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A. A., Sinari, S. A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E. G., Hernandez, R. D., Gignoux, C. R., Haussler, D., Katzman, S. J., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E. T., Devine, S. E., Min Kang, H., Kidd, J. M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi, X., Quitadamo, A., Lunter, G., Marchini, J. L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D. K., Oleksyk, T. K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B. L., Browning, S. R., Hormozdiari, F., Sudmant, P. H., Khurana, E., Tyler-Smith, C., Albers, C. A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M. B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A. O., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Y., Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Paolo Casale, F., Stegle, O., Lameijer, E.-W., Hall, I., Bafna, V., Michaelson, J., Gardner, E. J., Mills, R. E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M. J., Huddleston, J., Malig, M., Nelson, B. J., Parrish, N. F., Blackburne, B., Lindsay, S. J., Ning, Z., Zhang, Y., Lam, H., Sisu, C., Challis, D., Evani, U. S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., Berg Jespersen, J., Horn, H., Kim, D., Desalle, R., Narechania, A., Wilson Sayres, M. A., Mendez, F. L., David Poznik, G., Underhill, P. A., Coin, L., Mittelman, D., Banerjee, R., Cerezo, M., Fitzgerald, T. W., Louzada, S., Massaia, A., Ritchie, G.

R., Yang, F., Kalra, D., Hale, W., Dan, X., Barnes, K. C., Beiswanger, C., Cai, H., Cao, H., Henn, B., Jones, D., Kaye, J. S., Kent, A., Kerasidou, A., Mathias, R., Ossorio, P. N., Parker, M., Rotimi, C. N., Royal, C. D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J. C., Mathias, R. A., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Deng, X., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Sisay Joof, F., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Tịnh Hiê‘n, T., Dunstan, S. J., Thuy Hang, N., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Schieffelin, J., Grant, D. S., Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Vaydylevich, Y., Duncanson, A., Dunn, M., and Schloss, J. A. (2015), “A global reference for human genetic variation,” *Nature*, 526, 68–74. https://doi.org/10.1038/nature15393.

Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., Wigler, M., Navin, N., and Hicks, J. (2012), “Genome-wide copy number analysis of single cells.” *Nature protocols*, Nature Publishing Group, 7, 1024–41. https://doi.org/10.1038/nprot.2012.039.

Broadbent, S. R. (1955), “Quantum hypotheses,” *Biometrika*, 42, 45–57. https://doi.org/10.1214/11-AOS933.

Broadbent, S. R. (1956), “Examination of a quantum hypothesis based on a single set of data,” *Biometrika*, 43, 32–44.

Bryant, D., Tapper, W., Weston-Bell, N. J., Bolomsky, A., Song, L., Xu, S., Collins, A. R., Zojer, N., and Sahota, S. S. (2018), “Single-cell exomes in an index case of amp1q21 multiple

myeloma reveal more diverse mutanomes than the whole population," *Blood*, 132, 232–235. https://doi.org/10.1182/blood-2018-01-829291.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., and Stratton, M. R. (2008), "Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing," *Proceedings of the National Academy of Sciences*, 105, 13081–13086. https://doi.org/10.1073/pnas.0801523105.

Carothers, A. D. (1973), "Capture-Recapture Methods Applied to a Population with Known Parameters," *The Journal of Animal Ecology*, 42, 125. https://doi.org/10.2307/3408.

Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E., and Navin, N. E. (2018), "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing," *Cell*, Elsevier Inc., 172, 205–210.e12. https://doi.org/10.1016/j.cell.2017.12.007.

Castle, J. C., Biery, M., Bouzek, H., Xie, T., Chen, R., Misura, K., Jackson, S., Armour, C. D., Johnson, J. M., Rohl, C. A., and Raymond, C. K. (2010), "DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing," *BMC Genomics*, 11. https://doi.org/10.1186/1471-2164-11-244.

Chao, A. (1984), "Non-parametric estimation of the classes in a population," *Scandinavian Journal of Statistics*, 11, 265–270. https://doi.org/10.2307/4615964.

Cheng, C.-W., Hung, Y.-C., and Balakrishnan, N. (2012), *rBeta2009: The Beta Random Number and Dirichlet Random Vector Generating Functions*.

Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. a, and Schwartz, R. (2013), "Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations." *Bioinformatics (Oxford, England)*, 29, i189–98. https://doi.org/10.1093/bioinformatics/btt205.

Conterno Minussi, D., Nicholson, M., Ye, H., Davis, A., Wang, K., Sei, E., Du, H., Rabbani, M., Peng, C., Hu, M., Bai, S., McDonald, T., Schalck, A., Casasent, A., Barrera, A., Chen, H., Lim, B., Arun, B., Meric-Bernstam, F., Michor, F., and Navin, N. (n.d.). "Breast Tumors Maintain a Reservoir of Subclonal Diversity During Primary Expansion."

Cox, S. M. (2009), "Determining Greek Architectural Design Units in the Sanctuary of the Great Gods, Samothrace: Application of and Extensions to the Cosine Quantogram Method," PhD thesis, Emory University.

Csörgö, S. (1981), "Limit Behaviour of the Empirical Characteristic Function," *The Annals of Probability*, 9, 130–144.

Daley, T., and Smith, A. D. (2013), "Predicting the molecular complexity of sequencing libraries," *Nature Methods*, 10, 325–327. https://doi.org/10.1038/nmeth.2375.

Deng, C., Daley, T., Calabrese, P., Ren, J., and Smith, A. D. (2020), "Predicting the Number of Bases to Attain Sufficient Coverage in High-Throughput Sequencing Experiments," *Journal of Computational Biology*, 27, 1130–1143. https://doi.org/10.1089/cmb.2019.0264.

Derrien, T., Estellé, J., Sola, S. M., Knowles, D. G., Raineri, E., Guigó, R., and Ribeca, P. (2012), "Fast computation and applications of genome mappability," *PLoS ONE*, 7. https://doi.org/10.1371/journal.pone.0030377.

Durrett, R. (2013), "Population genetics of neutral mutations in exponentially growing cancer cell populations," *Annals of Applied Probability*, 23, 230–250. https://doi.org/10.1214/11-AAP824.

Efron, B., and Thisted, R. (1976), "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, 63, 435–447. https://doi.org/10.1093/biomet/63.3.435.

Ellis, B., Haaland, P., Hahne, F., Le Meur, N., Gopalakrishnan, N., Spidlen, J., and Jiang, M. (2009), "flowCore: basic structures for flow cytometry data," *R package version*, 1.

Ewens, W. J., and Wilf, H. S. (2007), "Computing the distribution of the maximum in balls-and-boxes problems with application to clusters of disease cases," *Proceedings of the National Academy of Sciences of the United States of America*, 104, 11189–11191. https://doi.org/10.1073/pnas.0704691104.

Ewers, S., Baldetorp, E. L., and Killander, D. (1984), "Flow-Cytometric DNA analysis in primary breast carcinomas and clinicopthological correlations," *Cytometry*, 5, 408–419. https://doi.org/10.1002/cyto.990050419.

Fan, X., Edrisi, M., Navin, N., and Nakhleh, L. (2019), "Benchmarking Tools for Copy Number Aberration Detection from Single-cell DNA Sequencing Data," *bioRxiv*, 696179. https://doi.org/10.1101/696179.

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009), "Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior," *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), 71, 993–1008. https://doi.org/10.1111/j.1467-9868.2009.00717.x.

Feuerverger, A., and Mureika, R. A. (1977), "The Empirical Characteristic Function and Its Applications," *The Annals of Statistics*, 5, 88–97. https://doi.org/10.1214/aos/1176343742.

Fisher, R. A. (1922), "On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, 85, 87. https://doi.org/10.2307/2340521.

Fisher, R. A., Corbet, S. A., and Williams, C. B. (1943), "The relation between the number of species and the number of individuals in a random sample of an animal population," *Journal of Animal Ecology*, 12, 42–58. https://doi.org/citeulike-article-id:3755508.

Freeman, P. R. (1976), "A Bayesian analysis of the megalithic yard," *Journal of the Royal Statistical Society. Series A*, 139, 20–55.

Frigo, M., and Johnson, S. G. (2005), "The design and implementation of FFTW3," *Proceedings of the IEEE*, 93, 216–231. https://doi.org/10.1109/JPROC.2004.840301.

Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., Meric-Bernstam, F., Michor, F., and Navin, N. E. (2016), "Punctuated copy number evolution and clonal stasis in triple-negative breast cancer," *Nature Genetics*, 48, 1–15. https://doi.org/10.1038/ng.3641.

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., Wigler, M., and Schatz, M. C. (2015), "Interactive analysis and assessment of single-cell copy-number variations," *Nature Methods*, 12, 1058–1060. https://doi.org/10.1038/nmeth.3578.

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. a, Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, P. A., Stewart, A., and Swanton, C. (2014), "Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing." *Nature genetics*, Nature Publishing Group, 46, 225–33. https://doi.org/10.1038/ng.2891.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012), "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing," *New England Journal of Medicine*, 366, 883–892. https://doi.org/10.1056/NEJMoa1113205.

Good, I. J., and Toulmin, G. H. (1956), "The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased," *Biometrika*, 43, 45. https://doi.org/10.2307/2333577.

Gotelli, N., and Colwell, R. (2011), "Estimating species richness," in *Biological diversity. Frontiers in measurement and assessment*, pp. 39–54. https://doi.org/10.2307/3547060.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., and Bustamante, C. D. (2011), "Demographic history and rare allele sharing among

human populations," *Proceedings of the National Academy of Sciences of the United States of America*, 108, 11983–11988. https://doi.org/10.1073/pnas.1019276108.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M. A., Gilks, C. B., Huntsman, D. G., McAlpine, J. N., Aparicio, S., and Shah, S. P. (2014), "TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data," *Genome Research*, 24, 1881–1893. https://doi.org/10.1101/gr.180281.114.

Hanahan, D., and Weinberg, R. A. (2011), "Hallmarks of Cancer: The Next Generation," *Cell*, Elsevier Inc., 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

Hansen, H. H., Christensen, L. J., Spang-Thomsen, M., Hirsch, F. R., Hansen, M., and Nissen, N. I. (1980), "Clonal Heterogeneity of Small-Cell Anaplastic Carcinoma of the Lung Demonstrated by Flow-Cytometric DNA Analysis," *Cancer Research*, 40, 4295–4300.

Hayes, J. L., Tzika, A., Thygesen, H., Berri, S., Wood, H. M., Hewitt, S., Pendlebury, M., Coates, A., Willoughby, L., Watson, C. M., Rabbitts, P., Roberts, P., and Taylor, G. R. (2013), "Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation," *Genomics*, Elsevier B.V., 102, 174–181. https://doi.org/10.1016/j.ygeno.2013.04.006.

Hedley, D. W., Friedlander, M. L., and Taylor, I. W. (1985), "Application of DNA flow cytometry to paraffin-embedded archival material for the study of aneuploidy and its clinical significance," *Cytometry*, 6, 327–333. https://doi.org/10.1002/cyto.990060409.

Heppner, G. H. (1984), "Tumor Heterogeneity," *Cancer Research*, 44, 2259–2265. https://doi.org/10.1007/978-94-009-8219-2_4.

Hewson, A. (1980), "The Ashanti weights—A statistical evaluation," *Journal of Archaeological Science*, 7, 363–370. https://doi.org/10.1016/S0305-4403(80)80041-0.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., and Wang, J. (2012), "Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm." *Cell*, Elsevier Inc., 148, 873–85. https://doi.org/10.1016/j.cell.2012.02.028.

Hung, Y. C., Balakrishnan, N., and Cheng, C. W. (2011), "Evaluation of algorithms for generating Dirichlet random vectors," *Journal of Statistical Computation and Simulation*, 81, 445–459. https://doi.org/10.1080/00949650903409999.

Ionita-Laza, I., Lange, C., and M Laird, N. (2009), "Estimating the number of unseen variants in the human genome." *Proceedings of the National Academy of Sciences of the United States of America*, 106, 5008–13. https://doi.org/10.1073/pnas.0807815106.

Jamal-Hanjani, M., Hackshaw, A., Ngai, Y., Shaw, J., Dive, C., Quezada, S., Middleton, G., Bruin, E. de, Le Quesne, J., Shafi, S., Falzon, M., Horswell, S., Blackhall, F., Khan, I., Janes, S., Nicolson, M., Lawrence, D., Forster, M., Fennell, D., Lee, S.-M., Lester, J., Kerr, K., Muller, S., Iles, N., Smith, S., Murugaesu, N., Mitter, R., Salm, M., Stuart, A., Matthews, N., Adams, H., Ahmad, T., Attanoos, R., Bennett, J., Birkbak, N. J., Booton, R., Brady, G., Buchan, K., Capitano, A., Chetty,

M., Cobbold, M., Crosbie, P., Davies, H., Denison, A., Djearman, M., Goldman, J., Haswell, T., Joseph, L., Kornaszewska, M., Krebs, M., Langman, G., MacKenzie, M., Millar, J., Morgan, B., Naidu, B., Nonaka, D., Peggs, K., Pritchard, C., Remmen, H., Rowan, A., Shah, R., Smith, E., Summers, Y., Taylor, M., Veeriah, S., Waller, D., Wilcox, B., Wilcox, M., Woolhouse, I., McGranahan, N., and Swanton, C. (2014), "Tracking genomic cancer evolution for precision medicine: the lung TRACERx study." *PLoS biology*, 12, e1001906. https://doi.org/10.1371/journal.pbio.1001906.

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., Lee, S.-M., Forster, M. D., Ahmad, T., Hiley, C. T., Abbosh, C., Falzon, M., Borg, E., Marafioti, T., Lawrence, D., Hayward, M., Kolvekar, S., Panagiotopoulos, N., Janes, S. M., Thakrar, R., Ahmed, A., Blackhall, F., Summers, Y., Shah, R., Joseph, L., Quinn, A. M., Crosbie, P. A., Naidu, B., Middleton, G., Langman, G., Trotter, S., Nicolson, M., Remmen, H., Kerr, K., Chetty, M., Gomersall, L., Fennell, D. A., Nakas, A., Rathinam, S., Anand, G., Khan, S., Russell, P., Ezhil, V., Ismail, B., Irvin-Sellers, M., Prakash, V., Lester, J. F., Kornaszewska, M., Attanoos, R., Adams, H., Davies, H., Dentro, S., Taniere, P., O'Sullivan, B., Lowe, H. L., Hartley, J. A., Iles, N., Bell, H., Ngai, Y., Shaw, J. A., Herrero, J., Szallasi, Z., Schwarz, R. F., Stewart, A., Quezada, S. A., Le Quesne, J., Van Loo, P., Dive, C., Hackshaw, A., and Swanton, C. (2017), "Tracking the Evolution of Non–Small-Cell Lung Cancer," *New England Journal of Medicine*, NEJMoa1616288. https://doi.org/10.1056/NEJMoa1616288.

Jaynes, E. T., and Bretthorst, G. L. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press.

Jenkins, D., Faits, T., Khan, M., Briars, E., Carrasco Pro, S., and Johnson, W. (2018), "singleCellTK: Interactive Analysis of Single Cell RNA-Seq Data." https://doi.org/10.18129/B9.bioc.singleCellTK.

Johnson, N. A. (2013), "A dynamic programming algorithm for the Fused Lasso and L 0-segmentation," *Journal of Computational and Graphical Statistics*, 22, 246–260. https://doi.org/10.1080/10618600.2012.681238.

Kallioniemi, O. P., Kallioniemi, A., Kurisu, W., Thor, A., Chen, L. C., Smith, H. S., Waldman, F. M., Pinkel, D., and Gray, J. W. (1992), "ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization," *Proceedings of the National Academy of Sciences of the United States of America*, 89, 5321–5325. https://doi.org/10.1073/pnas.89.12.5321.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., and Ding, L. (2013), "Mutational landscape and significance across 12 major cancer types," *Nature*, Nature Publishing Group, 502, 333–339. https://doi.org/10.1038/nature12634.

Kendall, D. (1986), "Quantum hunting," *Encyclopedia of statistical sciences*.

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N. E. (2018), "Chemoresistance Evolution in Triple-Negative Breast Cancer

Delineated by Single-Cell Sequencing," *Cell*, Elsevier Inc., 173, 879–893.e13. https://doi.org/10.1016/j.cell.2018.03.041.

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko, O., Poon, S., Eirew, P., Kabeer, F., Ruiz de Algara, T., Lee, S. R., Taghiyar, M. J., Huebner, C., Ngo, J., Chan, T., Vatrt-Watts, S., Walters, P., Abrar, N., Chan, S., Wiens, M., Martin, L., Scott, R. W., Underhill, T. M., Chavez, E., Steidl, C., Da Costa, D., Ma, Y., Coope, R. J., Corbett, R., Pleasance, S., Moore, R., Mungall, A. J., Mar, C., Cafferty, F., Gelmon, K., Chia, S., Hannon, G. J., Battistoni, G., Bressan, D., Cannell, I., Casbolt, H., Jauset, C., Kovačević, T., Mulvey, C., Nugent, F., Ribes, M. P., Pearsall, I., Qosaj, F., Sawicka, K., Wild, S., Williams, E., Aparicio, S., Li, Y., O'Flanagan, C., Smith, A., Ruiz, T., Balasubramanian, S., Lee, M., Bodenmiller, B., Burger, M., Kuett, L., Tietscher, S., Windager, J., Boyden, E., Alon, S., Cui, Y., Emenari, A., Goodwin, D., Karagiannis, E., Sinha, A., Wassie, A. T., Caldas, C., Bruna, A., Callari, M., Greenwood, W., Lerda, G., Lubling, Y., Marti, A., Rueda, O., Shea, A., Harris, O., Becker, R., Grimaldi, F., Harris, S., Vogl, S., Joyce, J. A., Hausser, J., Watson, S., Shah, S., Vázquez-García, I., Tavaré, S., Dinh, K., Fisher, E., Kunes, R., Walton, N. A., Al Sa'd, M., Chornay, N., Dariush, A., Solares, E. G., Gonzalez-Fernandez, C., Yoldas, A. K., Millar, N., Zhuang, X., Fan, J., Lee, H., Duran, L. S., Xia, C., Zheng, P., Marra, M. A., Hansen, C., and Shah, S. P. (2019), "Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing," *Cell*, 179, 1207–1221.e22. https://doi.org/10.1016/j.cell.2019.10.026.

Levin, B. (1981), "A Representation for Multinomial Cumulative Distribution Functions," *Annals of Statistics*, 9, 1123–1126. https://doi.org/10.1214/aos/1176345593.

Li, C., Wu, S., Yang, Z., Zhang, X., Zheng, Q., Lin, L., Niu, Z., Li, R., Cai, Z., and Li, L. (2017), "Single-cell exome sequencing identifies mutations in KCP, LOC440040, and LOC440563 as drivers in renal cell carcinoma stem cells," *Cell Research*, 27, 590–593. https://doi.org/10.1038/cr.2016.150.

Lin-Pedersen, T. (2020), "patchwork: The Composer of Plots."

Lonza Leggiera (2019), "Probability of at least 3 people sharing the same birthday in a group of n people," Mathematics Stack Exchange.

Macrae, F. (2018), "pmultinom: R package for computing the multinomial cumulative distribution function (CDF)."

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and Anna di Palma, M. (2019), *robustbase: Basic Robust Statistics*.

Martelotto, L. G., Baslan, T., Kendall, J., Geyer, F. C., Burke, K. A., Spraggon, L., Piscuoglio, S., Chadalavada, K., Nanjangud, G., Ng, C. K., Moody, P., D'Italia, S., Rodgers, L., Cox, H., Da Cruz Paula, A., Stepansky, A., Schizas, M., Wen, H. Y., King, T. A., Norton, L., Weigelt, B., Hicks, J. B., and Reis-Filho, J. S. (2017), "Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples," *Nature Medicine*, Nature Publishing Group, 23, 376–385. https://doi.org/10.1038/nm.4279.

McInnes, L., Healy, J., and Melville, J. (2018), "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction."

Navin, N. E. (2015), "The first five years of single-cell cancer genomics and beyond." *Genome research*, 25, 1499–507. https://doi.org/10.1101/gr.191098.115.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011), "Tumour evolution inferred by single-cell sequencing." *Nature*, Nature Publishing Group, 472, 90–4. https://doi.org/10.1038/nature09807.

Negrao, M. V., Quek, K., Zhang, J., and Sepesi, B. (2018), "TRACERx: Tracking tumor evolution to impact the course of lung cancer," *Journal of Thoracic and Cardiovascular Surgery*, The American Association for Thoracic Surgery, 155, 1199–1202. https://doi.org/10.1016/j.jtcvs.2017.10.134.

Nowell, P. (1976), "The clonal evolution of tumor cell populations," *Science*, 194, 23–28. https://doi.org/10.1126/science.959840.

Oltmann, J., Heselmeyer-Haddad, K., Hernandez, L. S., Meyer, R., Torres, I., Hu, Y., Doberstein, N., Killian, J. K., Petersen, D., Zhu, Y. J., Edelman, D. C., Meltzer, P. S., Schwartz, R., Gertz, E. M., Schäffer, A. A., Auer, G., Habermann, J. K., and Ried, T. (2018), "Aneuploidy, TP53 mutation, and amplification of MYC correlate with increased intratumor heterogeneity and poor prognosis of breast cancer patients," *Genes Chromosomes and Cancer*, 57, 165–175. https://doi.org/10.1002/gcc.22515.

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016), "Optimal prediction of the number of unseen species." *Proceedings of the National Academy of Sciences of the United States of America*, 113, 13283–13288. https://doi.org/10.1073/pnas.1607774113.

Peng, L., Xing, R., Liu, D., Bao, L., Cheng, W., Wang, H., Yu, Y., Liu, X., Jiang, L., Wu, Y., An, Z., Liang, Q., Kim, R. N., Shin, Y. K., Yang, H., Wang, J., Yu, J., Zhang, X., Xu, X., Yang, J., Wu, K., Zhu, S., and Lu, Y. (2019), "Characterization and validation of somatic mutation spectrum to reveal heterogeneity in gastric cancer by single cell sequencing," *Science Bulletin*, 64, 236–244. https://doi.org/https://doi.org/10.1016/j.scib.2018.12.014.

Perez-Rodriguez, D., Kalyva, M., Leija-Salazar, M., Lashley, T., Tarabichi, M., Chelban, V., Gentleman, S., Schottlaender, L., Franklin, H., Vasmatzis, G., Houlden, H., Schapira, A. H., Warner, T. T., Holton, J. L., Jaunmuktane, Z., and Proukakis, C. (2019), "Investigation of somatic CNVs in brains of synucleinopathy cases using targeted SNCA analysis and single cell sequencing," *Acta Neuropathologica Communications*, Acta Neuropathologica Communications, 7, 1–22. https://doi.org/10.1186/s40478-019-0873-5.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999), "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, 23, 41–46. https://doi.org/10.1038/12640.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014), "PyClone: statistical inference of clonal population structure in cancer." *Nature methods*, 11, 396–8. https://doi.org/10.1038/nmeth.2883.

Sanders, H. L. (1968), "Marine Benthic Diversity: A Comparative Study," *The American Naturalist*, 102, 243–282. https://doi.org/10.1086/282541.

Schröck, E., Veldman, T., Padilla-Nash, H., Ning, Y., Liyanage, M., Macville, M., Manoir, S. du, and Ried, T. (1997), "Spectral karyotyping," *Cancer Genetics and Cytogenetics*, 98, 146. https://doi.org/10.1016/s0165-4608(97)90230-2.

Scott, D. W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics*, 43, 274–285. https://doi.org/10.1198/004017001316975880.

Scott, D. W. (2004), "Outlier detection and clustering by partial mixture modeling," *COMPSTAT Symposium*, 453–465.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., Bashashati, A., Prentice, L. M., Khattra, J., Burleigh, A., Yap, D., Bernard, V., McPherson, A., Shumansky, K., Crisan, A., Giuliany, R., Heravi-Moussavi, A., Rosner, J., Lai, D., Birol, I., Varhol, R., Tam, A., Dhalla, N., Zeng, T., Ma, K., Chan, S. K., Griffith, M., Moradian, A., Cheng, S.-W. G., Morin, G. B., Watson, P., Gelmon, K., Chia, S., Chin, S.-F., Curtis, C., Rueda, O. M., Pharoah, P. D., Damaraju, S., Mackey, J., Hoon, K., Harkins, T., Tadigotla, V., Sigaroudinia, M., Gascard, P., Tlsty, T., Costello, J. F., Meyer, I. M., Eaves, C. J., Wasserman, W. W., Jones, S., Huntsman, D., Hirst, M., Caldas, C., Marra, M. a., and Aparicio, S. (2012), "The clonal and mutational evolution spectrum of primary triple-negative breast cancers," *Nature*, Nature Publishing Group, 486, 395–399. https://doi.org/10.1038/nature10933.

Shen, T., Chao, A., and Lin, C. (2003), "Predicting the Number of New Species in Further Taxonomic Sampling," *Ecology*, 84, 798–804.

Shi, H., Hugo, W., Kong, X., Hong, A., Koya, R. C., Moriceau, G., Chodon, T., Guo, R., Johnson, D. B., Dahlman, K. B., Kelley, M. C., Kefford, R. F., Chmielowski, B., Glaspy, J. A., Sosman, J. A., Van Baren, N., Long, G. V., Ribas, A., and Lo, R. S. (2014), "Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy," *Cancer Discovery*, 4. https://doi.org/10.1158/2159-8290.CD-13-0642.

Sobel, M., and Frankowski, K. S. (2004), "Handbook of Beta Distribution and Its Applications," in *Handbook of beta distribution and its applications*, eds. A. K. Gupta and S. Nadarajah, Boca Raton: CRC Press, pp. 319–360. https://doi.org/10.1201/9781482276596.

Su, Z., Wang, Z., Ni, X., Duan, J., Gao, Y., Zhuo, M., Li, R., Zhao, J., Ma, Q., Bai, H., Chen, H., Wang, S., Chen, X., An, T., Wang, Y., Tian, Y., Yu, J., Wang, D., Xie, X. S., Bai, F., and Wang, J. (2019), "Inferring the evolution and progression of small-cell lung cancer by single-cell sequencing of circulating tumor cells," *Clinical Cancer Research*, 25, 5049–5060. https://doi.org/10.1158/1078-0432.CCR-18-3571.

Sun, R., Hu, Z., Sottoriva, A., Graham, T. A., Harpak, A., Ma, Z., Fischer, J. M., Shibata, D., and Curtis, C. (2017), "Between-region genetic divergence reflects the mode and tempo of tumor evolution," *Nature Genetics*, Nature Publishing Group, 1–13. https://doi.org/10.1038/ng.3891.

Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2019), "Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq," *bioRxiv*, Cold Spring Harbor Laboratory. https://doi.org/10.1101/762773.

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A., and Teichmann, S. A. (2017), "Power analysis of single-cell rnA-sequencing experiments," *Nature Methods*, Nature Publishing Group, 14, 381–387. https://doi.org/10.1038/nmeth.4220.

Traag, V. A., Waltman, L., and Eck, N. J. van (2019), "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, 9, 1–12. https://doi.org/10.1038/s41598-019-41695-z.

Turajlic, S., and Swanton, C. (2017), "TRACERx Renal: tracking renal cancer evolution through therapy," *Nature Reviews Urology*, 14, 575–576. https://doi.org/10.1038/nrurol.2017.112.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Horswell, S., Chambers, T., O'Brien, T., Lopez, J. I., Watkins, T. B., Nicol, D., Stares, M., Challacombe, B., Hazell, S., Chandra, A., Mitchell, T. J., Au, L., Eichler-Jonsson, C., Jabbar, F., Soultati, A., Chowdhury, S., Rudman, S., Lynch, J., Fernando, A., Stamp, G., Nye, E., Stewart, A., Xing, W., Smith, J. C., Escudero, M., Huffman, A., Matthews, N., Elgar, G., Phillimore, B., Costa, M., Begum, S., Ward, S., Salm, M., Boeing, S., Fisher, R., Spain, L., Navas, C., Grönroos, E., Hobor, S., Sharma, S., Aurangzeb, I., Lall, S., Polson, A., Varia, M., Horsfield, C., Fotiadis, N., Pickering, L., Schwarz, R. F., Silva, B., Herrero, J., Luscombe, N. M., Jamal-Hanjani, M., Rosenthal, R., Birkbak, N. J., Wilson, G. A., Pipek, O., Ribli, D., Krzystanek, M., Csabai, I., Szallasi, Z., Gore, M., McGranahan, N., Van Loo, P., Campbell, P., Larkin, J., and Swanton, C. (2018), "Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal," *Cell*, 173, 595–610.e11. https://doi.org/10.1016/j.cell.2018.03.043.

Van Loo, P., and Campbell, P. J. (2012), "ABSOLUTE cancer genomics," *Nature Biotechnology*, Nature Publishing Group, 30, 620–621. https://doi.org/10.1038/nbt.2293.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A. L., and Kristensen, V. N. (2010), "Allele-specific copy number analysis of tumors," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 16910–16915. https://doi.org/10.1073/pnas.1009843107.

Wang, H., Nettleton, D., and Ying, K. (2014a), "Copy number variation detection using next generation sequencing read counts." *BMC bioinformatics*, 15, 109. https://doi.org/10.1186/1471-2105-15-109.

Wang, R., Lin, D. Y., and Jiang, Y. (2020), "SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing," *Cell Systems*, Elsevier Inc., 10, 445–452.e6. https://doi.org/10.1016/j.cels.2020.03.005.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., and Navin, N. E. (2014b), "Clonal evolution in breast cancer revealed by single nucleus genome sequencing," *Nature*, Nature Publishing Group, 512, 155–160. https://doi.org/10.1038/nature13600.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. (2019), *dplyr: A Grammar of Data Manipulation*.

Wickham, H., and Henry, L. (2019), *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*.

Wilke, C. O. (2019), *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.

Wu, H., Zhang, X.-Y., Hu, Z., Hou, Q., Zhang, H., Li, Y., Li, S., Yue, J., Jiang, Z., Weissman, S. M., Pan, X., Ju, B.-G., and Wu, S. (2016), "Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing," *Oncogene*, 1–11. https://doi.org/10.1038/onc.2016.438.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., and Iacobuzio-Donahue, C. A. (2010), "Distant metastasis occurs late during the genetic evolution of pancreatic cancer." *Nature*, 467, 1114–7. https://doi.org/10.1038/nature09515.

Yaeger, R., Chatila, W. K., Lipsyc, M. D., Hechtman, J. F., Cercek, A., Sanchez-Vega, F., Jayakumaran, G., Middha, S., Zehir, A., Donoghue, M. T. A., You, D., Viale, A., Kemeny, N., Segal, N. H., Stadler, Z. K., Varghese, A. M., Kundra, R., Gao, J., Syed, A., Hyman, D. M., Vakiani, E., Rosen, N., Taylor, B. S., Ladanyi, M., Berger, M. F., Solit, D. B., Shia, J., Saltz, L., and Schultz, N. (2018), "Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer." *Cancer cell*, 33, 125–136.e3. https://doi.org/10.1016/j.ccell.2017.12.004.

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y.,

Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C., and Campbell, P. J. (2015), "Subclonal diversification of primary breast cancer revealed by multiregion sequencing." *Nature medicine*, 21, 751–759. https://doi.org/10.1038/nm.3886.

Yu, C., Yu, J., Yao, X., Wu, W. K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y., Wu, R., Jian, M., Chen, R., Zhang, F., Xu, L., Fan, F., He, J., Liang, Q., Wang, H., Hu, X., He, M., Zhang, X., Zheng, H., Li, Q., Wu, H., Chen, Y., Yang, X., Zhu, S., Xu, X., Yang, H., Wang, J., Zhang, X., Sung, J. J., Li, Y., and Wang, J. (2014), "Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing." *Cell research*, Nature Publishing Group, 24, 701–12. https://doi.org/10.1038/cr.2014.43.

Zaccaria, S., and Raphael, B. J. (2019), "Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL," *bioRxiv*, 837195. https://doi.org/10.1101/837195.

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S. P., Aparicio, S., and Hansen, C. L. (2017), "Scalable whole-genome single-cell library preparation without preamplification," *Nature Methods*, 14, 167–173. https://doi.org/10.1038/nmeth.4140.

Zhang, C.-z., Adalsteinsson, V. A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K. L., Meyerson, M., and Love, J. C. (2015), "Calibrating genomic and allelic coverage bias in single-cell sequencing," *Nature Communications*, Nature Publishing Group, 6, 6822. https://doi.org/10.1038/ncomms7822.