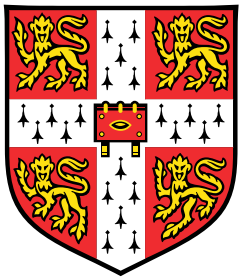
Exploiting properties of genetic variation.



Haynes Heaton

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2021

for E_2 & E_3

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Haynes Heaton
September 2021

Acknowledgements

It is often by luck or some other random vicissitudes of life through which the most opportunity and learning arise. In this, I would like to thank someone whose name I do not know for taking a semester off from Brown and opening up a single dorm room in technology house my sophomore year. And I would like to thank Jimmy Kaplowitz and Mike Katzourin for making that connection without which I would be a very different person today.

There I found my first unofficial mentors including Sean Smith, Lincoln Quirk, and Lucia Ballard from whom I learned intensely through both work and play. Through pair programming sessions with these three, I learned more in hours than months on my own.

The Brown Computer Science TA program was where I learned to teach and lead and where I came to understand that a topic you cannot teach is a topic you do not, yourself, understand. So I would first like to thank the founder of this program, Andy van Dam, who has been the driving force of not only this program, but the entire culture of the Brown Computer Science department for decades. Andy is an intense guy, but he also has a flair for the absurd. The undergraduate TA program brings a sense of ownership, membership, and community to the students who contribute to it. There I met my first official advisor, Sorin Istrail who saw much more potential in me than I saw in myself. I also met another mentor, Franco Preparata who I went on to work with for years to come. His creativity and infectious excitement for the work we did was perhaps what made me decide that science and computational biology was for me. I also met more friends and my first mentee who later really became another mentor for me. Dan Heller, or just "Heller" first came into my life as a student then as an applicant for a TA position under me in which he stated that "CS4 changed my life". At the time this seemed absurd, but in retrospect, it was absolutely true, and accepting his application changed my life as well. Heller's work ethic combined with his rare ability to combine practicality with rigor is an inspiration to me to this day.

I joined a company called Nabsys, which, despite not succeeding in its goal, succeeded in bringing together a number of talented people from whom I continued my intellectual journey. Peter Goldstein taught me statistics as I now understand it. Rather, he taught

me some of what I know and the rest we learned together by painstakingly unravelling some of the more esoteric papers (and theses) regarding our topic including several from Michael Waterman. At Nabsys I also met the most talented Biochemist I know, Brendan Galvin, who remains one of my closest friends and mentors. Brendan thinks of biochemical assays in a similar way to how I think of constructing an algorithm. Brendan is something of a stealth super contributor to the genomics and transcriptomics world. He is not particularly well know, but the field would be dramatically worse off without him. We went on to work together at another company and hopefully some day we will have the opportunity to work together again someday. Because of our distinct expertise, each of our understanding the possibilities and limitations of each other's fields, and our friendship, our collaborations have been some of the most productive of my life.

Through another on of life's serendipitous moments, I took what I thought was a throw away interview at a stealth genomics company. I was late for the interview due to poor planning, but Michael Schnall-Levin didn't balk at this and picked me up and brought me to the interview. After giving my job talk, I signed the non disclosure agreements and they told me what they were building. It wasn't until later that evening that the implications and possibilities started to sink in, and I started to become very excited. I took the job at 10x Genomics and still to this day I have never been in such a concentrated group of intellectual firepower. Patrick Marks is both one of the most talented computational biologists I have met and also the best manager I have had. The computational biology team as a whole is excellent because people follow truly talented and compassionate people like Pat. David Jaffe is one such person who became a friend and mentor to me. I miss his laugh, which is so absurd that you know he really means it. The rest of the company is also excellent and I would like to thank Alex Wong for running a great software team, Chris Hindson for never failing to deliver the secret sauce – the gel beads and oil for the microfluidic system, and Serge Saxanov and Ben Hindson for leading such a great company. It was a pleasure and honor to work there with such amazing people and create products that are still changing the face of biology today.

10x Genomics also contributed significantly to my opportunities going forward. Without the reputation of 10x Genomics as an innovative biotech company, my experience at 10x, and the papers and patents I was able to contribute to while working at 10x, I almost certainly would not have been considered for a PhD at some of the schools I was. I chose the Sanger Institute and Cambridge primarily because of Richard Durbin. His is one of the few textbooks that I have read cover to cover and I have followed his work throughout the years. The way he thinks comes through in his work and many of his

papers have not only been great contributions to the field, but mind expanding to me personally.

They say to never meet your heroes, and I have often felt the truth of this adage. However, Richard Durbin, or "The Durbinator" as I sometimes refer to him, consistently exceeds his already tremendous reputation. His algorithmic intuition is bar none. And while I would not be presumptuous enough to claim that we think alike, I would at least say that we have a similar algorithmic style. That is only part of what makes working with him incredible. He has the ability to see a global, decades-long plan for genomics and biology as well as the ability to work directly with the minute details of any of the diverse projects his group is working on.

Mara Lawniczak took me under her wing when I was struggling and gave me a home lab in which I could thrive. She has also been a true mentor to me in academia and life. I'd also like to thank her talented lab members including Ginny Howick, Arthur Talman, Juli Cudini, and others for their help and friendship. I'd also like to thank Sangjin Lee for his encouragement, support, and collaboration during the Covid19 pandemic. Without our pair programming sessions I would have been far less productive and less happy than I have been. During this past year, these sessions are often my only human contact in a given day.

Obviously I owe my family everything. They instilled in me a love of science, culture, literature, and art and have supported me every step of the way even when they didn't agree with some of my decisions. Thanks especially to my mother for trying to make the pandemic as good as possible for me.

And finally I'd like to thank my cat, Kasparov, for being a very cute kitty.

Abstract

Clustering single cell RNAseq by genotype using sparse mixture modeling.

While there are a number of methods for demultiplexing scRNAseq data on sample genotypes, they suffer from several error modes because they do not model the ambient RNA in the system. By not modeling this, the inferred genotypes of clusters are inaccurate and the cell doublet barcodes are dramatically overestimated. We present a method for clustering cells by genotype using a sparse probabilistic mixture model. We then run a co-inference of both the cluster genotypes and the ambient RNA allowing for more accurate genotype calls and the added benefit of being able to subtract off the expected expression of the ambient RNA from the transcription profile to give a more accurate view of the cell states. We have tested our model on real mixtures of human and *Plasmodium falciparum* with improved clustering compared to other methods showing both a performance improvement and applicability to a wide range of species and sample types.

Methods for genome assembly of challenging organisms.

While the technology improvements and cost reductions of third generation sequencing techniques have revolutionized genome assembly, problems remain especially for certain challenging organisms such as very small organisms and highly heterozygous organisms. We present a collection of methods aimed at addressing these issues and at validating the correctness of those assemblies. We show the first high quality reference genome made from a single mosquito and compare it to the current best reference. And we present several pieces of software and other algorithmic plans to address the lingering problems with assembling these organisms.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Assembly

1.1.1 DNA sequencing

1.1.1.1 Historically

1.1.1.2 Short reads

1.1.1.3 Long reads

1.1.2 Reference Genomes

1.1.2.1 Resequencing

1.1.2.2 Read Mapping

1.1.2.3 Variant Calling

1.1.2.4 Population Genomics

1.1.2.5 The old way

1.1.2.6 The new way and Darwin Tree of Life and Earth Biogenome Project

1.1.3 Haplotype phasing

1.1.3.1 Statistical

1.1.3.2 Direct / Read based

1.1.4 Assembly

1.1.4.1 Overlap, Layout, Consensus

1.1.4.2 De brujin graphs

1.1.4.3 String graphs

1.1.4.4 Repeats, Heterozygosity, and Errors

Chapter 2

Clustering single cell RNAseq by genotypes in mixed samples.

2.1 Background

Understanding the link between genotype and phenotype is a key goal of biology. The major efforts toward this can be thought of as top down and bottom up. Much effort has been made on the top down approach of linking genotype to macroscopic complex phenotypic traits such as height and diseases in the form of genome wide association studies (GWAS) [18]. More recently with the advent of RNAseq, we have begun to understand the link between genetic variants and mRNA expression levels via eQTL analysis [39].

Cells are a natural discrete building block of biology. And tissues are almost always complex arrangements of multiple different cell types. Bulk RNAseq is a blunt instrument measuring the average RNA content of many cells in a tissue. Advances in methods for the preparation of samples containing minuscule amounts of nucleic acids have made it possible to study the transcriptional state of single cells [59]. Single cell RNAseq (scRNAseq) is the process of measuring the transcriptional profile of each cell individually usually by physically separating cells and delivering distinct barcode sequences to templates generated from the mRNA of each cell [49]. Further advances in nanodroplet and nanowell technologies have made it possible to apply scRNAseq to thousands of cells simultaneously [38][69][12].

Many samples contain cells of mixed genotypes including those of single celled organisms such as mixed strain malaria infections, the gut microbiome, and environmental samples as well as intrinsically mixed samples such as maternal/fetal, transplant patient, or tumor samples and intentionally multiplexed samples. In order to properly analyze this

data we must first identify each cell's genotype of origin. Some tools and methods exist for this purpose [21] [58] [66] but each of them contain some downside such as the need for prior knowledge of the genotypes or a sample barcode and all of them contain several error modes arising from the lack of modeling ambient RNA in the system. Ambient RNA in single-cell RNAseq (also known as soup) is a recently described phenomenon in which RNA molecules from cells which have lysed before cell partitioning are included in partitions with cells from which they did not originate [67]. In these demultiplexing systems, not modeling the ambient RNA makes many cells appear to be doublets and makes many homozygous variant sites appear to be heterozygous.

2.2 Aims

The aims of this project are twofold. One is to build a general tool for dealing with mixed or multiplexed scRNAseq data. The other is to apply this tool to mixed strain malaria infections to learn about strain competition and selective sexual behaviour.

As a general tool for mixed scRNAseq data we wish to be able to determine the genotypes in a sample, assign cells to those genotypes, and identify intergenotypic doublet cell barcodes. In order to do this we must first identify a strategy for calling reliable variants in scRNAseq data and assigning allele counts to cell barcodes. And finally we must develop a clustering method which is robust to the dramatically sparse datatype as well as the inherent noise sources of ambiant RNA and doublet cell barcodes. Once these methods are in place we must characterize their performance on artificially mixed samples as well as actual mixed samples. To test the accuracy in true mixed samples we will use current tools which require the genotypes to be known a priori as a benchmark. We will do this with malaria samples and human samples created for this project.

In collaboration with other lab members we also wish to apply this tool to mixtures of malaria both from the lab and the field to study the effect of competition on sexual determination and selective mating. To do this we will compare mixed samples and uniform samples and look for differential expression as well as differential expression among the strains and attempt to identify quantitative trait loci to explain those differences.

2.3 Methods

2.3.1 Variant calling on scRNAseq data

Little work has been done on identifying genetic variants in bulk RNAseq [50] let alone scRNAseq [9]. Currently the most popular software for the initial analysis of scRNAseq data going from the reads to the expression matrix is cellranger [69].

2.3.1.1 Remapping

In the cellranger pipeline, the mapping component is done with the STAR aligner [5] which, while sufficient for the purpose of counting gene expression, produces artifacts in the alignments that produce many false positive variants. One such source of false positives is the soft clipping penalty which is not a parameter exposed to the user in the STAR software. It is often the case in WGS and even more so in RNAseq that the starts and ends of reads can be less reliable than the rest of the read. Because of this, mappers built for variant calling such as BWA [36] and minimap2 [35] have a relatively small one-time penalty for soft clipping any number of bases from the end of an alignment [34]. The STAR alignment soft clipping penalty is such that, in comparison with other aligners, it can create many false positive variant calls produced entirely by bases at the ends of reads. Another source of small variant errors caused by the STAR alignments is that the default indel penalty relative to the mismatch penalty is much higher than that of variant calling ready aligners. These penalties will, for example, prefer inducing 10 single base mismatches rather than a single 12 base indel. Further, because the sequences are correct, it will make the same error for all of the reads spanning the indel and these bases but not those spanning those bases but not spanning the indel. This penalty is exposed as a parameter to the user, but with the default parameters (and thus with the output of cellranger) these errors exist. And finally the last source of errors these alignments induce are due to the leniency of spliced alignments which STAR has. With its default parameters including a max intron length of 200kb, STAR will often include erroneous and statistically spurious spliced alignments of reads that otherwise don't align well. This creates alignments which match for some statistically significant portion in one location and then are spliced to other loci often for an 8-12 base segment which should occur by random chance alone. And due to the nature of mapping qualities being to the whole alignment and not each segment of the alignment, these sections are often denoted as having high mapping quality when in fact they should occur by chance. If, however, there is actually an alternative allele in one of these regions to which some reads have spurious matches, those reads, and thus those cells, are said to support the reference

allele with high probability. These alignments provide one further technical issue, which is that they dramatically slow down the pileup and fetch commands which are necessary for variant calling.

For these reasons we suggest first remapping with either BWA, minimap2, or hisat2. We have personally found the best results when remapping with minimap2 with a combination of long read splice parameters and short read parameters. Specifically, the parameters for which all analysis is done in this report are the following: minimap2 -ax splice -t 8 -G50k -k 21 -w 11 –sr -A2 -B8 -O12,32 -E2,1 -r200 -p.5 -N20 -f1000,5000 -n2 -m20 -s40 -g2000 -2K50m –secondary=no. We then perform deduplication by removing reads with the same unique molecular identifier (UMI) barcode, cell barcode, and have the same start and stop position.

Once we have an alignment file, we must then proceed to variant calling. There are two strategies we can take for calling variants on scRNAseq. We can treat the sample as a population of cells and call variants with a population variant caller [11] [4] [37]. With this approach we would need to define each cell barcode as its own read group in the input bam and the variant caller would output a population VCF with genotype calls for each cell for each locus. Or we can treat the sample as an unknown mixture of individuals, call variants on that unknown mixture, and then for each read (and thus its cell barcode) decide whether it supports the reference allele or alternative allele which can be done with the tool vartrix[9]. Our analysis suggests these two strategies perform very similarly with minimal filtering. And because the latter strategy is much more computationally efficient, all further analysis is done with freebayes with parameters –pooled-continuous -iXu -C 2 -q 20 -n 3 -E 1 -m 30 –min-coverage 6 and vartrix with parameters –umi –mapq 30 –scoring-method coverage.

2.3.1.2 Variant Candidate Calling

2.3.1.3 Cell allele assignment

2.3.1.4 Validation: Genome in a Bottle

2.3.1.5 Post transcriptional modifications

2.3.2 Sparse mixture model clustering

In order to introduce this method, we must first motivate it with a description of the data type and its particular difficulties with respect to clustering by genotypes. Each cell barcode has reads from its transcription profile sampled very sparsely. Firstly, the test data I have is one mixed sample of six strains of *Plasmodium falciparum* (malaria) which

is a unicellular haploid parasitic organism and the sample contains cells coming from all cell types found in the life cycle in the human blood stage. And the other data set is three replicates of mixed samples of five human individuals from the human induced pluripotent stem cell project. Some sample statistics for malaria and human single cell data are shown in table 2.1 and some histograms of some of those metrics are shown in figure 2.1. We use a filter requiring at least four cells supporting each allele otherwise the variant is unlikely to be of almost any use in discriminating between different genotypes in this mixture. As you can see, the number of cells expressing any given locus is far fewer than the total number of cells and the number of variants with a given number of cells expressing that variant drops off roughly exponentially as cells expressing a given locus increases. It is also evident that while the human data contains more discriminating variants per cell, they are spread over many more total variants thus making the overlap between any two cells very low.

Table 2.1 Single cell data statistics

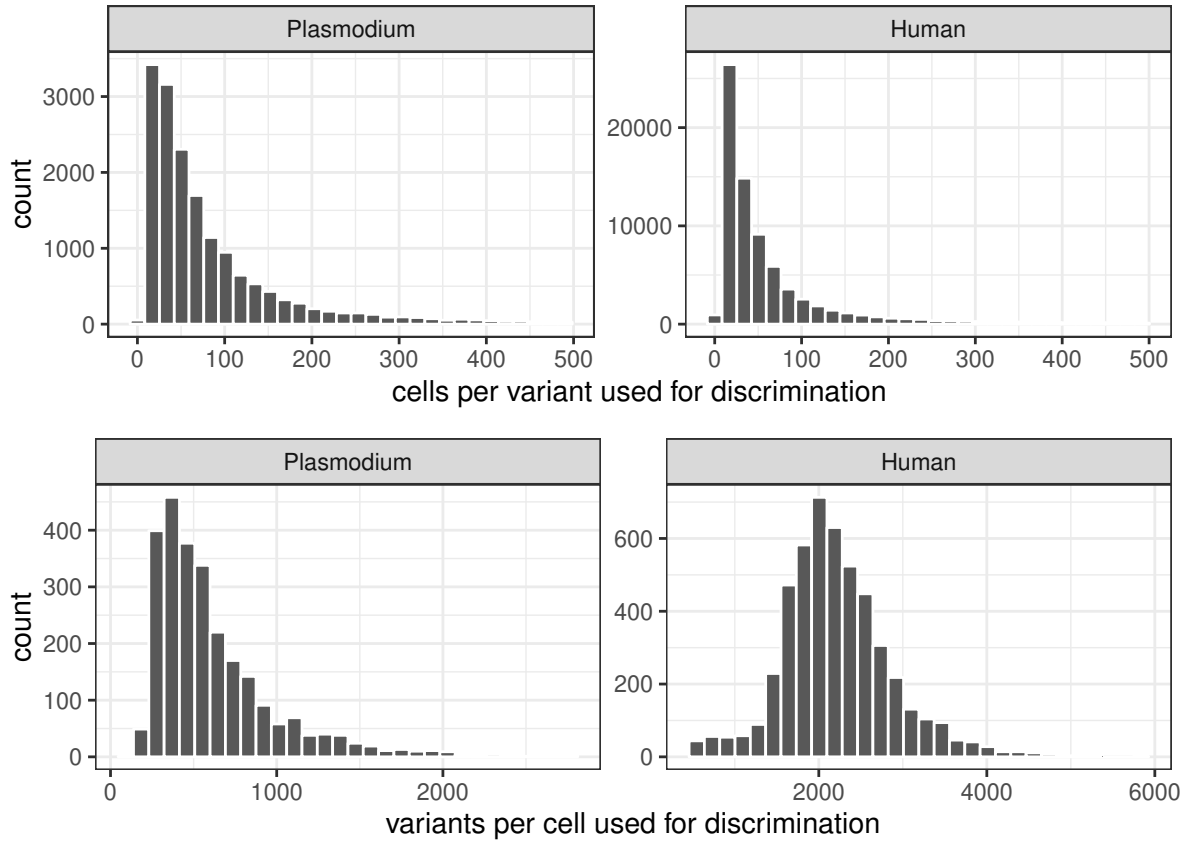
	malaria	human replicate 1
number of cells	2608	4925
median UMI per cell	995	25155
total variants	39487	194079
median cells per variant	24	18
median variants per cell	667	2642
total discriminating variants	16783	77878
median discriminating variant per cell	512	2147
median cells per discriminating variants	55	38
median genes per cell	571	4812

We wish to both determine which cells contain the same genotypes, but also what those genotypes are, and we achieve this with sparse mixture model clustering.

Definitions

- K: number of genotype clusters to be fixed at outset. Lower case k will be used for indexing and referring to a specific cluster.
- C: number of cells. Lower case c will be used for indexing and referring to a specific cell. This actually refers to a barcode which generally comes from a single droplet but could correspond to multiple droplets. This barcode could have 0, 1, or more

Fig. 2.1 Single cell sparsity



cells. It is important for some assumptions in this model that the majority of barcodes contain a single cell.

- L: number of variant loci. Lower case l will be used to index and refer to a specific loci. We will assume only biallelic variants. L_c will be a list of loci with observed data in cell c.

- A: Allele counts. $A_{l,c}$ is a vector of size 2 with the first number representing the number of reference alleles and the second value representing the number of alt alleles seen at locus l in cell c.

- $\phi_{k,l}$: mixture parameter for allele fractions of cluster k at locus l. This is a real number representing the fraction of ref alleles in this cluster at this locus. We expect this to be near 1.0 (homozygous reference), 0.5 (heterozygous), or 0.0 (homozygous alt) but will be skewed from these values by noise, doublets, and ambient RNA.

2.3.2.1 Model

Weuse a maximum likelihood strategy by maximizing $p(data)$ under a given model.

$$\underset{\phi}{\operatorname{argmax}}\, p(data, \phi) \tag{2.1}$$

For each cell we marginalize across the clusters it could belong to and at each locus the reference allele count is modeled by a binomial with n as the reference + alternative allele counts for that cell at that locus and p as the mixture parameter for that cluster at that locus.

$$p(A, \phi) = \prod_{c \in C} \sum_{k \in K} \frac{1}{K} \prod_{l \in L_c} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,0}} \phi_{k,l}^{A_{l,c,0}} (1 - \phi_{k,l})^{A_{l,c,1}} \tag{2.2}$$

Generally we minimize the negative log probability, which in this case is differentiable and thus susceptible to numerical optimization techniques such as gradient descent. It is worth noting here that the binomial probability of 0 counts with $n = 0$ is always 1 for any p and thus for a given cell c we can ignore loci which have no observed counts of either allele. We solve this optimization problem with the machine learning package tensorflow with a custom loss function and the Adam optimizer which is a type of gradient descent using decaying momentum [23]. In practice the loss function that we actually use is the following

$$loss = -log\left[\prod_{c \in C} \sum_{k \in K} \frac{1}{K} \prod_{l \in L_c} exp^{-\left(\frac{A_{l,c,0}}{A_{l,c,0}+A_{l,c,1}} - \phi_{k,l} \right)^2} \right] \tag{2.3}$$

due primarily to not understanding how to use tensorflow probability distributions in which the parameters of the distribution are variable values. We have since found a solution to this but have not yet implemented it. And treating this as a log probability we can calculate a pseudo-posterior cell assignments to clusters

$$p(c \in K_j) \approx \frac{\prod_{l \in L_c} exp^{-\left(\frac{A_{l,c,0}}{A_{l,c,0}+A_{l,c,1}} - \phi_{j,l} \right)^2}}{\sum_{k \in K} \prod_{l \in L_c} exp^{-\left(\frac{A_{l,c,0}}{A_{l,c,0}+A_{l,c,1}} - \phi_{k,l} \right)^2}} \tag{2.4}$$

And the posteriors for cell c being in cluster j are as follows

$$p(c \in K_j) = \frac{\prod_{l \in L_c} \binom{A_{l,c,0}+A_{l,c,1}}{A_{l,c,0}} \phi_{j,l}^{A_{l,c,0}} (1-\phi_{j,l})^{A_{l,c,1}}}{\sum_{k \in K} \prod_{l \in L_c} \binom{A_{l,c,0}+A_{l,c,1}}{A_{l,c,0}} \phi_{k,l}^{A_{l,c,0}} (1-\phi_{k,l})^{A_{l,c,1}}} \tag{2.5}$$

This method, as is the case with many clustering methods, may suffer from local minima in instances of poor initialization of the cluster centers ϕ. To get around this, we run this method a number of times (generally 15) and take the minimum loss across these randomized iterations.

2.3.3 Deterministic Annealing

2.3.4 Doublet cell barcode detection

One of the major aims of this work is to detect the barcodes which contain multiple cells with different genotypes. We are not, however, attempting to detect barcodes which contain multiple cells with the same genotype. We assume that the generation of doublet cell barcodes is a random poisson process and that the λ of this poisson process is low enough that the chance of multiplets > 2 are exceedingly unlikely. We view this problem as a multi-urn problem in which we each locus is an urn containing either the allele counts of the best fitting cluster for this cell or the allele counts of the combination of the top two clusters for this cell.

Doublet model: Urn problem Definitions

- $A_{k,l}$: Allele counts at locus l for all cells in cluster k according to the maximum probability cluster assignment from our clustering. This is a vector of size two with the ref and alt allele counts.
- K_c: Will be used to denote the list of clusters for cell c ordered from the maximum posterior probability to the minimum. Thus $K_{c,0}$ is the cluster with the highest posterior probability for cell c. And for ease of notation we will denote $A_{K_{c,0\cup 1},l}$ as the sum of the allele counts of those two clusters.

This can be computed directly.

$$p(c \in K_{c,0}) = \prod_{l \in L} \binom{A_{l,c,0}+A_{l,c,1}}{A_{l,c,0}} \frac{A_{K_{c,0},l,0}}{A_{K_{c,0},l,0}+A_{K_{c,0},l,1}}^{A_{l,c,0}} \left(1 - \frac{A_{K_{c,0},l,0}}{A_{K_{c,0},l,0}+A_{K_{c,0},l,1}}\right)^{A_{l,c,1}} \tag{2.6}$$

Beta binomial

$$p(c \in K_i) = \prod_{l \in L_c} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,0}} \frac{\beta(A_{l,c,0} + 1 + A_{i,l,0}, A_{l,c,1} + 1 + A_{i,l,1})}{\beta(1 + A_{i,l,0}, 1 + A_{i,l,1})} \tag{2.7}$$

$$alpha_{l,i,j} = 1 + \frac{\frac{A_{i,l,0}}{A_{i,l,0}+A_{i,l,1}} + \frac{A_{j,l,0}}{A_{j,l,0}+A_{j,l,1}}}{2} min(A_{i,l,0} + A_{i,l,1}, A_{j,l,0} + A_{j,l,1}) \tag{2.8}$$

$$beta_{l,i,j} = 1 + \frac{\frac{A_{i,l,1}}{A_{i,l,0}+A_{i,l,1}} + \frac{A_{j,l,1}}{A_{j,l,0}+A_{j,l,1}}}{2} min(A_{i,l,0} + A_{i,l,1}, A_{j,l,0} + A_{j,l,1}) \tag{2.9}$$

$$p(c \in K_i \cup K_j) = \prod_{l \in L_c} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,0}} \frac{\beta(A_{l,c,0} + alpha_{l,i,j}, A_{l,c,1} + beta_{l,i,j})}{\beta(alpha_{l,i,j}, beta_{l,i,j})} \tag{2.10}$$

$$p(doublet_c|c) = \frac{p(c \in K_i \cup K_j)p(doublet)}{p(c \in K_i \cup K_j)p(doublet) + p(c \in K_i)(1 - p(doublet))} \tag{2.11}$$

And the probability of the doublet case is

$$p(c \in K_{c,0} \cup K_{c,1}) = \prod_{l \in L} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,0}} \frac{A_{K_{c,0\cup1},l,0}}{A_{K_{c,0\cup1},l,0} + A_{K_{c,0\cup1},l,1}}^{A_{l,c,0}} \left(1 - \frac{A_{K_{c,0\cup1},l,0}}{A_{K_{c,0\cup1},l,0} + A_{K_{c,0\cup1},l,1}}\right)^{A_{l,c,1}} \tag{2.12}$$

And we can simply normalize these probabilities to sum to one for rough posteriors. We could also give the doublet vs singlet case a prior probability, but here we assume uniform chance. We then remove these doublets before moving on to the ambient RNA detection and cluster genotype inference.

2.3.5 Ambient RNA detection and Cluster genotype inference

One major goal of clustering scRNAseq by genotypes is calling the genotypes of each loci for each cluster. But as previously discussed, there can be lysed cells in solution prior to cell partitioning which contribute a background noise to both genotypes and transcriptional profiles. This ambient RNA gives a fuzzy picture of the transcriptional profile and makes cluster genotypes which are in truth homozygous appear heterozygous. Luckily with genotype mixtures, we can use our prior knowledge of the ploidy of the mixed organisms along with our genotype cluster assignments to make a co-inference of both the genotypes and level of ambient RNA in the experiment.

2.3.5.1 Mixture model of ambient RNA and cell RNA

Definitions

- ρ: mixture parameter representing the probability any given allele is arising from ambient RNA as opposed to from the cell associated with that barcode.
- P: ploidy. We assume ploidy is limited to 1 or 2.
- A_l: total allele expression at locus l. This is again a vector of length 2 denoting the reference and alternative allele counts.
- g: used to denote the number of copies of the reference allele. So the expected reference allele rate without ambient RNA is $\frac{g}{P}$ and g is an integer value $\in [0..P]$. And it is worth noting that this one number is sufficient to denote the genotypes for ploidy 1 and 2.
- $p(true)$: probability variant is a true positive.

Once again we solve this with a maximum likelihood approach.

$$\underset{\rho}{\operatorname{argmax}}\, p(data, \phi) \tag{2.13}$$

And the model treats each locus in each cluster as coming from one of three genotypes for diploid (0/0, 0/1, 1/1, here denoted by g=0,1, or 2) and two genotypes from haploid (0, 1). We treat each cluster as independent and each locus as independent. Then we marginalize across the possible genotypes. And so we model the allele counts in this cluster as having come from a mixture of ambient RNA and from this cells in this cluster. And then we model the observed allele fractions as being drawn from a binomial distribution with a probability which was skewed away from $p = \frac{g}{P}$ by the level of ambient RNA. And we believe that the ambient RNA is drawn from an average of all of the reads in the experiment. As such, the expected allele fraction coming from the soup is $\frac{A_{l,0}}{A_{l,0}+A_{l,1}}$. Thus, the probability of the binomial of the mixture of cell data is the following.

$$p_{tp} = (1-\rho)\frac{g}{P} + \rho\frac{A_{l,0}}{A_{l,0}+A_{l,1}} \tag{2.14}$$

$$p_{fp} = \frac{A_{l,0}}{A_{l,0}+A_{l,1}} \tag{2.15}$$

Now we use that binomial probability in the full likelihood.

$$p(data|\rho) = \prod_{l \in L} \Bigg[p(true) \prod_{k \in K} \sum_{g=0}^{P} \frac{1}{P} \binom{A_{k,l,0} + A_{k,l,1}}{A_{k,l,0}} p_{tp}^{A_{k,l,0}} (1 - p_{tp})^{A_{k,l,1}} + (1 - p(true)) \prod_{k \in K} \binom{A_{k,l,0} + A_{k,l,1}}{A_{k,l,0}} p_{fp}^{A_{k,l,0}} (1 - p_{fp})^{A_{k,l,1}} \Bigg] \tag{2.16}$$

2.3.5.2 Inference

And we solve this in a maximum likelihood fashion using STAN which is a domain specific language for probabilistic models.

And the posteriors for genotypes for each locus for each cluster can easily be computed by normalizing the binomial probabilities over all possible genotypes.

2.4 Results

In order to analyze our results we must first create a ground truth dataset to compare it to. We have the whole genome sequencing data for each of these strains and individuals, so we can use demuxlet [21] to assign cells to these known genotypes as well as call doublets. In doing this we ran into several problems. First, we ran demuxlet on the malaria data which is a mixture of six strains of *Plasmodium falciparum* and it claimed that every cell was from the 3D7 strain. As we found this to be unlikely given the experimental setup which aimed to get an even mixture of strains we looked into things further. We found that the indel calls on the WGS data were poor due to the extremely A/T rich nature of the *Plasmodium* genome. So we then removed the indels and reran demuxlet. It then showed all cells as 3D7 except for some doublets of 3D7 and other strains. It should be noted here that the reference used for mapping the WGS data was created from the 3D7 strain. So we suspected that reference bias might be a serious issue. We then remapped all of the WGS data to a reference created from a strain that this experiment did not contain. We then got a good mixture of cell assignments, but the doublet detection rate was an order of magnitude higher than the expected doublet rate given the number of input cells. On further analysis we have decided that this overestimation of doublets is due to the ambient RNA. So for our analysis we will compare to the demuxlet best call ignoring whether the cell was also called a doublet as well as to the demuxlet calls ignoring demuxlet called doublets.

2.4.1 Benchmarking: Synthetic human cell mixture vs real human cell mixture

2.4.1.1 Validation and comparison to other methods

Samples have been ordered such that the primary numbers show up on the diagonal.

Table 2.2 Clustering concordance of human scRNAseq replicate 1 with Demuxlet

		Demuxlet best sample				
		euts	babz	nufh	oaqd	ieki
Cluster	0	1110	1	0	9	0
	1	7	800	1	5	6
	2	4	5	691	9	3
	3	7	0	2	1640	7
	4	3	2	1	0	612

Table 2.3 Clustering concordance of human scRNAseq replicate 2 with Demuxlet

		Demuxlet best sample				
		ieki	babz	eutz	oaqd	nufh
Cluster	0	628	0	5	1	0
	1	2	794	8	3	1
	2	2	3	1133	3	0
	3	1	4	3	1523	2
	4	7	2	5	9	693

Table 2.4 Clustering concordance of human scRNAseq replicate 3 with Demuxlet

		Demuxlet best sample				
		nufh	eutz	babz	oaqd	ieki
Cluster	0	732	0	3	14	4
	1	2	1241	7	5	1
	2	3	2	805	4	4
	3	0	4	2	1655	4
	4	1	0	0	1	650

We currently do not have sc split run on these samples as the variant calling step is very computationally expensive in the way they suggest. We wish to have this comparison by the viva.

Giving adjusted rand indexes of 0.96, 0.97, and 0.97 respectively. If we remove the demuxlet called doublets, the concordance is perfect on all three replicates.

The ambient RNA estimation for these samples is 19.4%, 18.8%, and 19.2% respectively which are currently unvalidated as are the inferred genotype calls.

The doublet analysis code was written on a previous clustering and variant calling strategy which output different formats and needs reworking to get results. We wish to complete this analysis in time for the viva.

2.4.2 Maternal-Fetal data

2.4.3 Plasmodium

Samples have been ordered such that the primary numbers show up on the diagonal of the confusion matrix.

Table 2.5 Clustering concordance of malaria scRNAseq dataset with Demuxlet

		Demuxlet best strain					
		3D7	GB4	7G8	11.02	15.04	28.04
Cluster	0	1132	0	0	1	1	0
	1	1	414	0	0	0	0
	2	0	0	274	0	0	0
	3	0	0	1	242	0	0
	4	0	0	0	0	219	0
	5	1	0	0	0	0	321

Which gives an adjusted rand index of 0.995. Excluding cells called by demuxlet as doublets (23.7%) there is perfect concordance.

Sc split [65] was a tool recently released with this same goal in mind and we have run it on this dataset and here show the comparison of that tool to demuxlet. This is already removing cells called doublets by sc split (13%) but not removing cells called doublets by demuxlet.

This clustering seen in table 2.6 gives an adjusted rand index of 0.62 splitting 3D7 into two clusters and combining some others and also identifying nearly all of the 28.04 strain as doublets.

Table 2.6 sc split concordance of malaria scRNAseq dataset with Demuxlet

		Demuxlet best strain					
		7G8	3D7	GB4	11.02	15.04	28.04
Cluster	0	222	1	1	3	8	2
	1	0	559	0	1	0	1
	2	0	1	386	1	0	1
	3	1	2	12	224	79	5
	4	0	566	1	1	0	2
	5	52	1	7	11	111	7

2.4.4 Twenty one individual mixture demonstration

2.4.4.1 Contamination revealed

2.5 Discussion

We have presented a collection of novel methods for use in mixed sample single cell RNAseq datasets and have shown that they significantly out perform current competing methods for clustering cells by genotype. For doublet detection we have good reason to believe that our approach will yield more accurate doublet calls than either demuxlet or sc split in the presence of significant ambient RNA. And we also have good reason to believe that our genotype calls will be of higher quality. These last two things we hope to have more analysis on by the time of the first year viva.

Chapter 3

High quality assembly of a single Mosquito

3.1 Background

Exciting efforts to sequence the diversity of life are building momentum [32] but one of many challenges that these efforts face is the small size of most organisms. For example, arthropods, which comprise the most diverse animal phylum, are typically small. Beyond this, while levels of heterozygosity within species vary widely across taxa, intraspecific genetic variation is often highest in small organisms [31]. Over the past two decades, reference genomes for many small organisms have been built through considerable efforts of inbreeding organisms to reduce their heterozygosity levels such that many individuals can be pooled together for DNA extractions. This approach has varied in its success, for example working well for organisms that are easy to inbreed (e.g., many Drosophila species [6]), but less well for species that are difficult or impossible to inbreed (e.g., Anopheles [47]). Therefore, many efforts to sequence genomes of small organisms have relied primarily on short-read approaches due to the large amounts of DNA required for long-read approaches. For example, the recent release of 28 arthropod genomes as part of the i5K initiative used four different insert size Illumina libraries, resulting in an average contig N50 of 15 kb and scaffold N50 of 1 Mb [61].

Another way to overcome DNA input requirements, while also reducing the number of haplotypes present in a DNA pool, is to limit the number of haplotypes in the pool of individuals by using offspring from a single cross. This is easier than multiple generations of inbreeding, and can be successful. For example, a recent PacBio Aedes aegypti assembly used DNA extracted from the offspring of a single cross, thus reducing the maximum

number of haplotypes for any given locus to four, thereby improving the assembly process and achieving a contig N50 of 1.3 Mb [44].

However, for an initiative like the Earth BioGenome Project [32] that aims to build high-quality reference genomes for more than a million described species over the next decade, generating broods to reach sufficient levels of high molecular weight DNA for long-read sequencing will be infeasible for the vast majority of organisms. Therefore, new methods that overcome the need to pool organisms are needed to support the creation of reference-quality genomes from wild-caught individuals to increase the diversity of life for which reference genomes can be assembled. Here, we present the first high-quality genome assembled with unamplified DNA from a single individual insect using a new workflow that greatly reduces input DNA requirements.

Until recent advances in long read library prep [22], it was not possible to obtain enough DNA from a single individual of small organisms such as mosquitos to create a long-read sequencing library from one individual. But for many other smaller species, this remains the case. And it also remains the case for nanopore sequencing. Whether it is possible to decrease the input requirements for nanopore sequencing and to what extent are currently unknown.

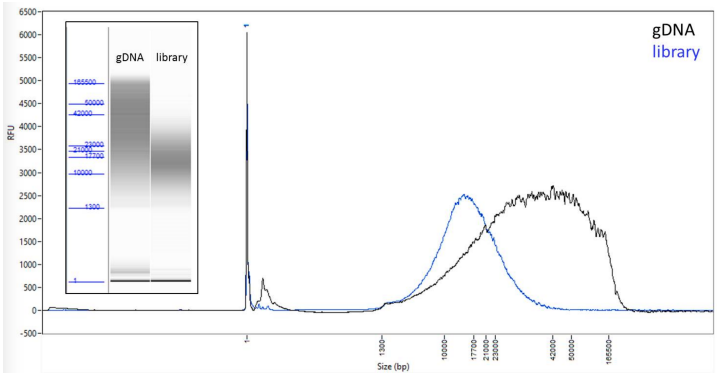
3.2 DNA Isolation

The DNA isolation was carried out by Juliana Cudini, a fellow PhD student.

High molecular weight (HMW) DNA was isolated from a single *Anopheles coluzzii* female from the Ngousso colony. This colony was created in 2006 from the broods of approximately 100 wild-caught pure *Anopheles coluzzii* females in Cameroon (pers. comm. Anna Cohuet). Although the colony has been typically held at >100 breeding individuals, given the long time since colonization, there is undoubtedly inbreeding. A single female was ground in 200 μl PBS using a pestle with several up and down strokes (i.e., no twisting), and DNA extraction was carried out using a Qiagen MagAttract HMW kit (PN-67653) following the manufacturer?s instructions, with the following modifications: 200 μl 1X PBS was used in lieu of Buffer ATL; PBS was mixed simultaneously with RNAse A, Proteinase K, and Buffer AL prior to tissue homogenisation and incubation; incubation time was shortened to 2 h; solutions were mixed by gently flicking the tube rather than pipetting; and subsequent wash steps were performed for one minute. Any time DNA was transferred, wide-bore tips were used. These modifications were in accordance with recommendations from 10X Genomics HMW protocols that aim to achieve >50 kb molecules. The resulting sample contained 250 ng of DNA, and we used

the FEMTO Pulse (Advanced Analytical, Ankeny, IA, USA) to examine the molecular weight of the resulting DNA. This revealed a relatively sharp band at 150 kb (Figure S1). The DNA was shipped from the U.K. to California on cold packs, and examined again by running 500 pg on the FEMTO Pulse. While a shift in the molecular weight profile was observed as a result of transport, showing a broader DNA smear with mode of 40 kb (Figure 1), it was still suitable for library preparation (note that this shifted profile is coincidentally similar to what is observed with the unmodified MagAttract protocol). DNA concentration was determined with a Qubit fluorometer and Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Waltham, MA, USA), and 100 ng from the 250 ng total was used for library preparation.

Fig. 3.1 *Anopheles coluzzii* input and resulting library DNA lengths



(a) FEMTO Pulse traces and ?gel? images (inset) of the genomic DNA input (black) and the final library (blue) before sequencing.

3.3 Library prep and Sequencing

Library prep and sequencing were performed by Sarah Kingan, Senior Scientist at PacBio.

A SMRTbell library was constructed using an early access version of SMRTbell Express Prep kit v2.0 (Pacific Biosciences, Menlo Park, CA, USA). Because the genomic DNA was already fragmented with the majority of DNA fragments above 20 kb, shearing was not necessary. 100 ng of the genomic DNA was carried into the first enzymatic reaction to remove single-stranded overhangs followed by treatment with repair enzymes to repair any damage that may be present on the DNA backbone. After DNA damage repair, ends of the double stranded fragments were polished and subsequently tailed with an A-overhang. Ligation with T-overhang SMRTbell adapters was performed at 20 C for 60 min. Following ligation, the SMRTbell library was purified with two AMPure

PB bead clean up steps (PacBio, Menlo Park, CA), first with 0.45X followed by 0.80X AMPure. The size and concentration of the final library (Figure 3.1) were assessed using the FEMTO Pulse and the Qubit Fluorometer and Qubit dsDNA HS reagents Assay kit (Thermo Fisher Scientific, Waltham, MA, USA), respectively. Sequencing primer v4 and Sequel DNA Polymerase 3.0 were annealed and bound, respectively, to the SMRTbell library. The library was loaded at an on-plate concentration of 5?6 pM using diffusion loading. SMRT sequencing was performed on the Sequel System with Sequel Sequencing Kit 3.0, 1200 min movies with 120 min pre-extension and Software v6.0 (PacBio). A total of 3 SMRT Cells were run.

3.4 Assembly

As previously discussed, recent advances in Pacbio library prep have reduced the DNA input requirement to a level (100ng) at which it is possible to create a long read library from a single mosquito. Of course sequencing a single individual is beneficial to assembly over sequencing a pool of individuals as the problem that heterozygosity imposes on assembly is greatly exacerbated due to the additional haplotypes in a pool of individuals over the two haplotypes contained in a single individual.

High molecular weight DNA was isolated from a single An. coluzzii female mosquito and shipped to Pacbio for early access low input library preparation and sequencing. We obtained three SMRT cells of Pacbio long read data. This data was assembled using the Falcon-Unzip software with the median length subread per ZMW for a for a total of 12.8 Gb of sequence which equates to roughly 48x coverage of the expected genome size. The resulting primary assembly consisted of 372 contigs totaling 266 Mb in length, with a contig N50 of 3.5 Mb and a secondary haplotype assembly totalling 78.5 Mb. 3.1 shows various assembly statistics of the long read assembly and the current reference assembly of the closely related *Anopheles gambiae*[19][55][56][30].

3.5 Curation

3.6 Assembly statistics

The contigs were screened by the Sanger assembly curation team to identify contaminants and mitochondrial sequence identifying two mitochondrial contigs and one complete assembly (4.24 Mb single contig) of *Elizabethkingia anophelis*, which is a common gut microbe in *Anopheles* mosquitoes [27].

Table 3.1 Assembly statistics

		Pacbio Raw	Pacbio Curated	Sanger Assembly
Primary Assembly	Size (Mb)	266	251	224
	No. Contigs	372	206	27,063
	Contig N50 (Mb)	3.52	3.47	0.025
Alternate Haplotigs	Size (Mb)	78.5	89.2	unresolved
	No. Contigs	665	830	N/A
	Contig N50 (Mb)	0.22	0.199	N/A

3.6.1 Quality assessment

According to Busco [63] analysis of the primary assembly, 110 genes were duplicated indicating some resolved heterozygosity (haplotigs) remaining in the primary assembly. The presence of duplicated haplotypes in a reference genome can result in erroneously low mapping qualities in resequencing studies and cause problems when scaffolding. We used the Purge Haplotigs software [51] and identified 165 primary contigs totalling 10.6 Mb as likely haplotigs which were then moved to the alternate haplotigs fasta.

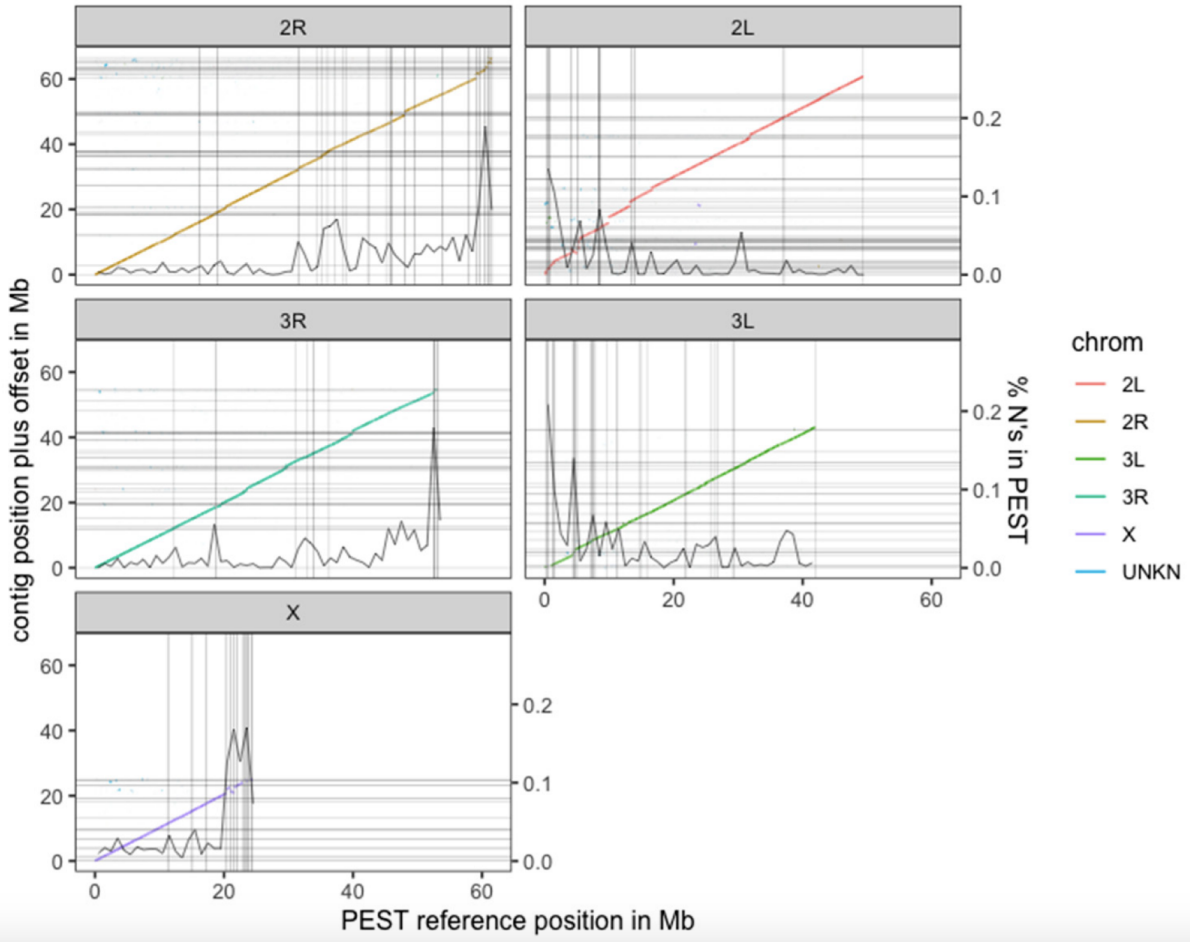
There are many problems with the current *Anopheles gambiae* reference including 6302 gaps of Ns in the primary chromosome scaffolds ranging from 20 bases to 36 kb and 55 gaps of 10 kb that the AGP (A Golden Path) file on Vectorbase annotates as "contig" endings. This reference also contains a large bin of unplaced contigs (27.3 Mb excluding Ns) designated as the "UNKN" (unknown) chromosome. However, it is the previously best characterized *Anopheles* assembly which should be very closely related to the *coluzzii* species so we endeavored to make comparisons between them. We aligned the assembly contigs to the reference with minimap2 and then attempted to assign contigs to chromosomes as well as order and orient them. The new assembly is highly concordant with the reference over the entire genome, allowing the placement of the long PacBio contigs into chromosomal contexts (figure 3.2). We also showed that the assembly correctly expanded long repeats that had been collapsed in the reference (figure 3.4) and that the assembly resolved an incorrect order and orientation of the scaffolding of chromosome X (figure 3.6).

We also found that despite running Purge Haplotigs, there remained some haplotig sequence at the ends of contigs (figure 3.10) due to the fact that Purge Haplotigs only looks for full contigs that have evidence of being a haplotig.

The PEST annotation also retains a large bin of unplaced contigs (27.3 Mb excluding Ns) designated as the ?UNKN? (unknown) chromosome. We compared the alignments

3.7 Comparison to *Anopheles gambiae* PEST reference

Fig. 3.2 Comparison of the assembly with the PEST reference

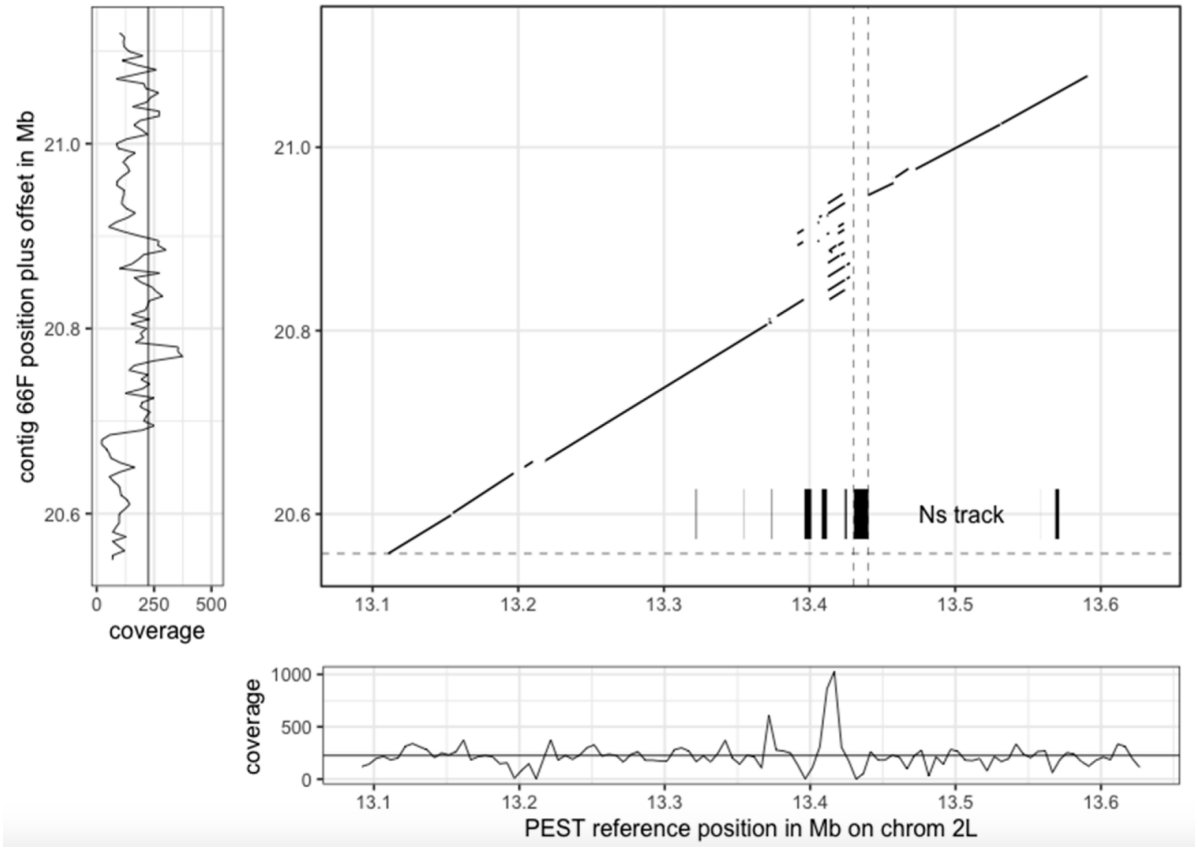


(a) Alignment of the curated PacBio contigs to the AgamP4 PEST reference [21]. Alignments are colored by the primary PEST reference chromosome to which they align but are placed in the panel and Y offset to which the contig as a whole aligns best. Contig ends are denoted by horizontal lines in the assembly and vertical lines in PEST. However, there are many Ns in PEST not annotated as contig breaks so the percent Ns per megabase of PEST is overlaid (scale on the right Y axis). There are no Ns in the PacBio assembly.

of contigs from the PEST chromosomes (X, 2, 3) versus the contigs from the UNKN to the new assembly. Any regions with a mapping quality score (mapq) 60 alignments of both UNKN and chromosomal contigs are likely to be haplotigs in the UNKN. In total, we find that 7.27 Mb are haplotigs (i.e., also have PEST chromosomal alignments to the same location in the assembly) and another 10.9 Mb are newly placed sequence

3.7.1 Expansion of previously collapsed repeat

Fig. 3.4 Example of expansion of previously collapsed repeat

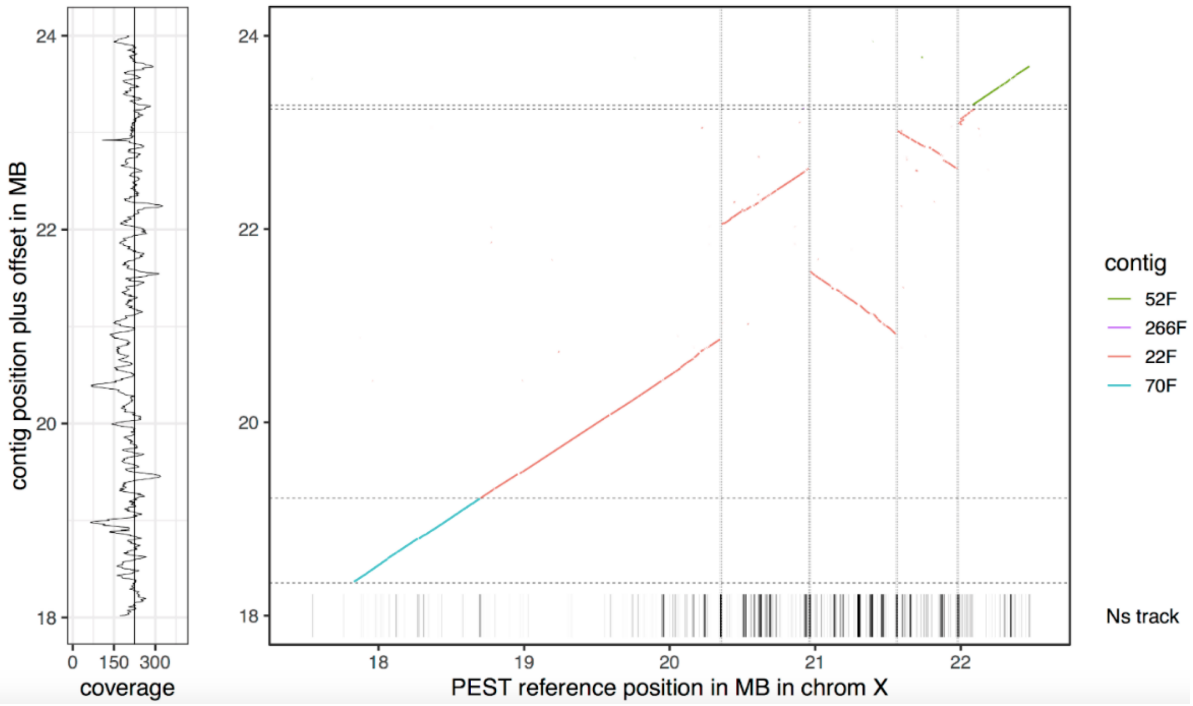


(a) Example of a compressed repeat in PEST that has been expanded by the PacBio assembly. Dotted vertical lines represent a gap in the PEST assembly (10,000 Ns) between scaffolds, which is now spanned by the single PacBio contig. Coverage plot of the PacBio subreads aligned to PEST (bottom) highlights the region where excess coverage indicates a collapsed repeat in PEST, in contrast the coverage of PacBio subreads aligned to the PacBio contig (left) is more uniform.

that do not overlap with PEST chromosomal alignments. The UNKN bin also contains 737 annotated genes. Remarkably, our single-insect assembly now places 667 (>90%) of these formerly unplaced genes into their appropriate chromosomal contexts (2L:148 genes; 2R:162 genes; 3L: 126 genes; 3R:91 genes; X:140 genes; unplaced:70 genes; details on specific genes can be found in Table S4), which together with their flanking sequence comprise 8.9 Mb of sequence. Altogether, this means that 40% of the UNKN chromosome is now placed in the genome, along with 90% of the genes that were contained within it.

3.7.2 Corrected order and orientation vs PEST scaffolding

Fig. 3.6 Resolved order and orientation error in PEST scaffolding



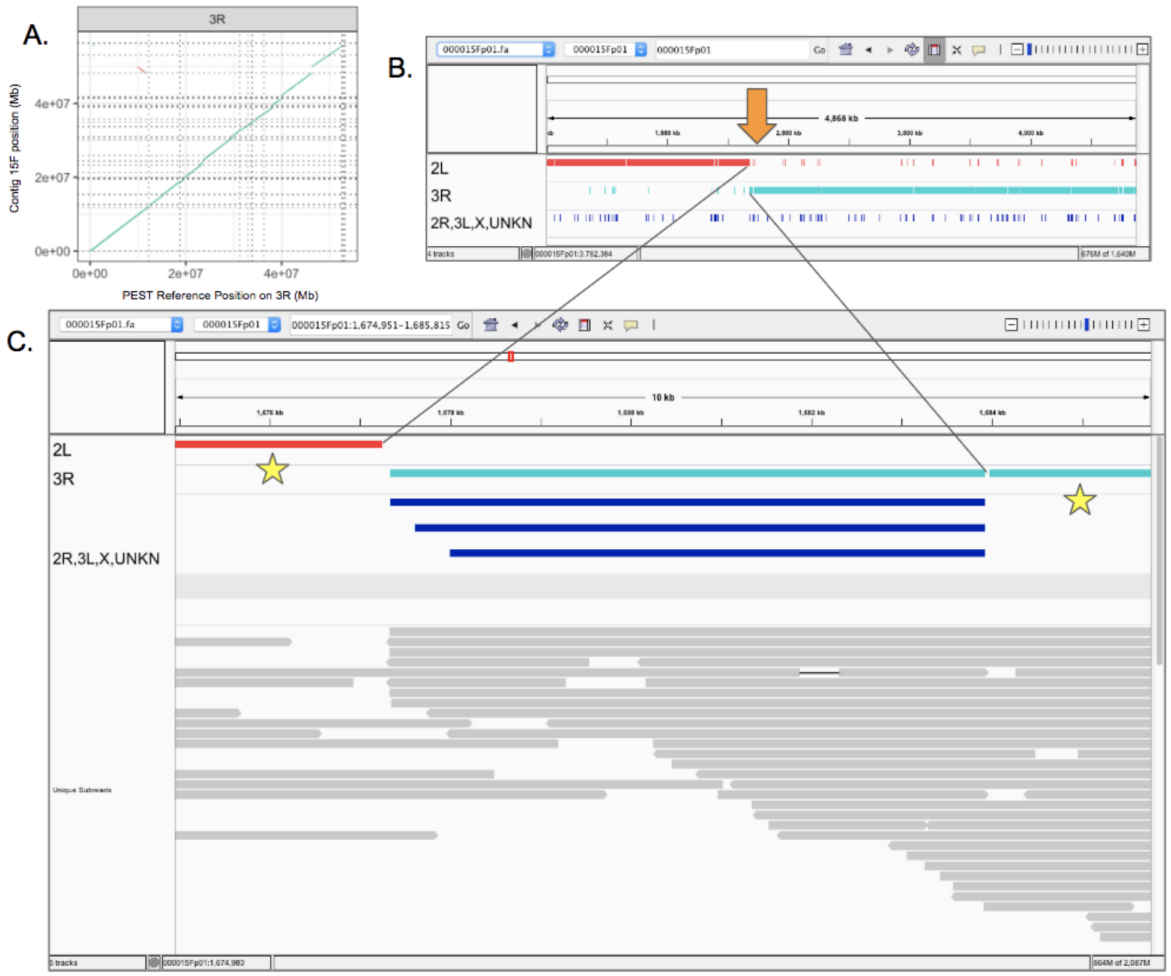
(a) Alignment of X pericentromeric contigs to PEST, highlighting likely order and orientation issues in the PEST assembly that are resolved by a single PacBio contig.

3.7.5 Placement of previously unplaced genes

Chromosome arm	Number of placed genes
2L	148
2R	162
3L	126
3R	91
X	140
unplaced	70

3.7.3 Identification and correction of misassembly

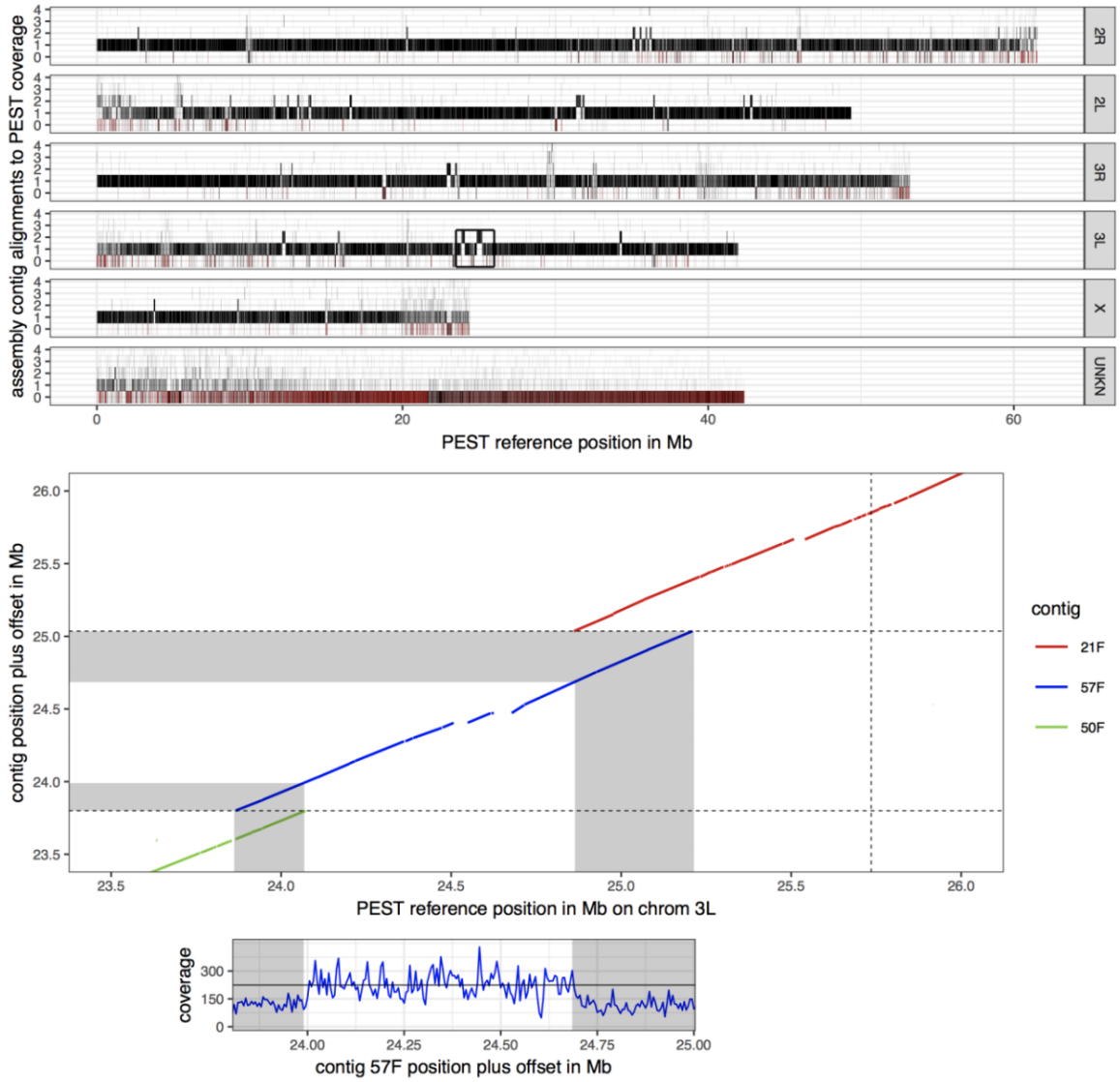
Fig. 3.8 Chimeric assembly



(a) A chimeric contig between 2L and 3R. A. Alignment of PacBio contigs to PEST identifies a candidate chromosomal rearrangement. B. IGV screenshot of breakpoint (orange arrow) localized by alignment of contig to PEST. Red: alignment to 2L, turquoise: alignments to 3R, navy blue: alignments to other chromosomes and unplaced contigs. C. IGV visualization of mapped unique subreads at breakpoint shows 0 subreads mapping across the central repetitive region into the unique flanking sequence on the left (2L) and right (3R) (stars). A count of spanning reads was also determined with bedtools bamtobed utility. The 6.5kb central region aligns to four loci in the PEST genome and has 370 bp of sequence similarity to the Tc1-like transposase gene in *Anopheles gambiae*.

3.7.4 Remaining haplotig sequence on ends of contigs

Fig. 3.10 Evidence of remaining haplotig contig ends.



(a) Alignment and coverage plot (top) of the PacBio assembly contigs relative to PEST, and magnification of one area of excess coverage (bottom). In the top panel, the number of alignments of PacBio contigs to PEST are represented by black bars, with most of the genome showing a 1:1 correspondence to PEST. Red denotes N?s in the reference. Isolated areas of higher number of contig alignments are visible, one of which (black box) is magnified in the bottom panel. Here, the ends of neighboring contigs overlap, which is currently not resolved with the Purge Haplotigs software since the overlap is only partial. The sequencing depth of PacBio reads for the central (blue) contig (57F) corroborate this interpretation, exhibiting half of the expected coverage in the greyed regions of contig overlap, and with the corresponding ends of the red and green contigs complementing with the other half of coverage, respectively (not shown for clarity).

Chapter 4

Methods for assembly of challenging organisms

4.1 Background

Reference genomes have enabled a range of genomic analysis by providing prior knowledge of the sequence and giving genomic context as well as a common coordinate system by which to compare multiple genomes [1] [53]. Assembling reference genomes is complicated by repetitive sequences, heterozygosity, and sequencing errors. Historically reference genomes were created by large haploid clone libraries [29] which solve many of the repeat issues as well as the heterozygosity problem, but these methods are too costly to apply to many genomes. More recently the cost reductions of long read technologies [8] [13] as well as the emergence of other long range genetic information technologies [68] [54] [20] have converged to make high quality, cost effective reference genomes. This has then resparked interest in assembly as well as large reference generation projects such as the Earth BioGenome Project [33] and the Darwin Tree of Life Project. For these technologies there are now assembly algorithms that deal with each data type [3] [64] [57] as well as combinations of multiple technologies [60] [45] [43]. These methods try to co-assemble both haplotypes arriving at a haploid consensus [52] [25] or a diploid assembly [26] [64], but heterozygosity injects complexity and ambiguity on top of a haploid assembly process. It requires that the assembler disambiguate between paralogous sequence and differences between haplotypes. When the level of heterozygosity is high, the differences between haplotypes can be even greater than the differences between paralogous sequences. One method for dealing with the problem of heterozygosity is inbreeding organisms to a point of low heterozygosity [46], but this is not possible for all organisms. Trio-sga used pedigree sequencing information in the assembly algorithm [40] but does not work on long

read data. Recently Koren et al. described trio binning which uses a mother-father-child trio to separate long reads into their haplotype of origin prior to assembly [24]. While this method is very effective, creating such a cross would be infeasible for many species. And even with this method, many assembly artifacts remain.

4.2 Aims

The unifying aim of this project is to create high quality reference genomes for species that thus far have been challenging to assemble well. This project will have a focus of developing methods to address problems posed by small, highly heterozygous organisms, with a specific focus on the Anopheles genus. But where possible, we will develop methods that are generalizable to many more species.

We aim to develop methods to split reads into individual haplotypes prior to assembly to address the inherent problem of disambiguating near repeats versus haplotype differences. We will take two main strategies. One will involve sequencing multiple individuals from a pedigree with short accurate reads from which we will determine the haplotype distinguishing kmers. With these kmers we will bin the long reads into their respective haplotypes. The other strategy we will take is to use multiple technologies on a single individual in order to find heterozygous loci, phase them, and use those phased kmers to bin long reads into their respective haplotypes prior to assembly. This strategy has the bonus of not requiring multiple related individuals but the downside of requiring enough DNA from a single individual for every technology used.

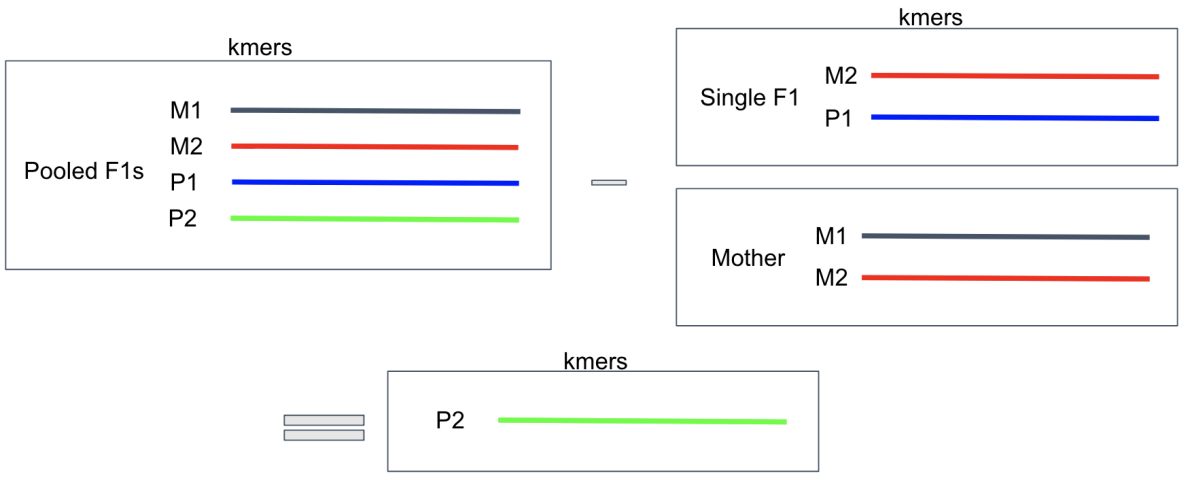
Both of these methods still suffer from one further downside. While separating haplotypes prior to assembly solves the disambiguation of paralogues from heterozygosity, we are then making two assemblies each with half of the total coverage. We aim to solve this either by using each assembly to scaffold the other, or to construct a graph containing both assemblies. From this graph we could choose a single path as the linear reference.

4.3 Pedigree sample strategies

As discussed already there are simple strategies for splitting long reads by haplotypes using short accurate reads from both parents. But for many species such as wild caught samples it may be difficult to obtain the paternal sample whereas it is possible to capture a fertilized female and grow her brood in isolation. For this case we propose to use three short read samples: one maternal sample, one sample from a single F1, and

another sample from a pool of multiple F1 offspring. From these samples we can obtain distinguishing kmers or probabilistic distinguishing kmers for each haplotype of both the mother and father.

Fig. 4.1 Pseudotrio



As shown in 4.1 we can obtain distinguishing kmers for the P1 haplotype by set subtraction of kmers in the maternal and single F1 sample from the those in the pooled samples. In a similar way we can subtract kmers in the maternal sample from the single F1 sample to get probabilistic distinguishing kmers from the P2 haplotype. We say probabilistic distinguishing kmers because these kmers could also occur on the P1 haplotype as well. And in the same way we can get probabilistic distinguishing kmers for the M1 haplotype by subtracting kmers in the single F1 sample from those in the maternal sample. And finally we can obtain probabilistic distinguishing kmers for the M2 haplotype will be obtained by selecting kmers which are shared between the maternal and single F1 sample and occur at haploid counts in both samples. Once we have these distinguishing kmers we can use them to separate long reads by haplotype for libraries created from any of these samples. We have created software for collecting these distinguishing kmers [14] which uses counting bloom filters to minimize memory usage and a mod based system allowing distribution of jobs to many worker nodes. And we also have software for binning long reads based on those distinguishing kmers [17].

4.4 Phasstools: phasing and assembly tools

4.4.1 Heterozygous kmer pairs and detection

First we tackle the subject of identifying heterozygous variants in a reference-free manner. We will do this using a kmer approach, as many reference-free methods do. Many people have focused on identifying kmers which occur at roughly half counts in short read data [41] and various software exists to count kmers [42] and to model the mixture of expected distributions (errors, haploid, diploid, duplication kmers) [62]. Identifying heterozygous kmers in this way suffers from multiple problems from the perspective of de novo phasing. 1. Many of these identified as half counts will be either randomly high count error kmers or randomly low count homozygous kmers. 2. You end up with $K-1$ overlapping kmers for a given variant which is needlessly redundant information which will both slow down any phasing algorithm and likely break key independence assumptions. And 3. while you have heterozygous kmers, you don't know which kmers are alternative alleles of each other. We propose finding pairs of kmers which vary only in the center position which are also both roughly at half counts. These heterozgyous SNP kmers will be much more robust and have the benefit of knowing that one is the alternative allele of the other. We have software for this purpose [16] which uses counting bloom filters to minimize memory usage and a mod based method to distribute the work to many worker nodes.

4.4.2 Phasing consistency

4.4.3 Data types and uses

4.4.4 Phasst phase: Reference or assembly based phasing

4.4.4.1 Sparse *Bernoulli* mixture model clustering

4.4.4.2 Polyploid phasing

4.4.4.3 Phasing consistency genotype correction

4.4.5 Phasst a: phased assembly

4.4.5.1 Phasing consistent heterozygous kmer recruitment

4.4.5.2 Haplotype and chromosome read binning

4.4.5.3 Haploid chromosome assemly

4.4.6 Phasst scaff: phasing aware assembly scaffolding

4.4.6.1 Chromosome binning

4.4.6.2 Ordering and Orienting

In order to separate long reads by haplotype for a single individual, we must develop an algorithm for de novo haplotype phasing. In reference based systems, haplotype phasing begins by mapping reads to the reference and calling variants from the reference. Then physical linkage information of two heterozygous variants occurring on data known to be generated from a single haplotype is used to determine which alleles come from which of the sister chromosomes across either some region (denoted as a phase block) or across whole chromosomes [68] [48] [7] [2]. At each heterozygous variant locus, one allele can arbitrarily be denoted as 0 and the second allele denoted 1. Then without loss of generality we choose to represent the series of alleles located on whichever chromosome has the 0 allele of the first variant in a phase block. So now the problem can be seen as determining the binary sequence of which alleles are on that same chromosome that are most consistent with the linkage data we have. To do this, each of these methods uses different search strategies (beam search, dynamic programming, graph based heuristic search) to find the configuration that maximizes the probability of the data under some error model. These algorithms are also aided by the knowledge of which variants are close to each other (and thus are most likely to contain linkage information) on the

genome. It is common to proceed in a directed manner, determining the optimal solution for variants $[0..n-1]$ before determining the phase of variant n. So for de novo haplotype phasing we will need to 1. identify heterozygous alleles 2. determine an ordering of those alleles and 3. create a search strategy to maximize a probabilistic model of the data.

4.4.6.3 Diploid assembly validation

While certain tools for assembly validation are available [41] [63], the validation of diploid assemblies is a difficult problem still in its infancy. These heterozygous kmers are also useful for validating diploid assemblies because you expect to see one of the heterozygous kmer pairs in the primary assembly and the other one in the secondary haplotigs. If, for instance, you find both versions of a heterozygous kmer pair in the primary assembly, this could either represent residual heterozygosity in the primary assembly which should be removed, or perhaps the kmers came from paralogous regions and both kmers had lower than homozygous counts due to random sampling error. Another error type that exists is lacking either of the kmer pair in the primary which could represent low base level accuracy, over collapse of near repeat regions, or some other type of error. Other cases exist which may have their own interpretations. We have developed software for this analysis which is now in use by the Sanger assembly curation team [15].

The next step for de novo phasing is determining a rough ordering of heterozygous SNPs. If we view the heterozygous kmers as nodes in a graph and edges between nodes as linkage information of long reads or other long genetic distance linkage information with weights as the number of those links, we could view our problem as a graph based traveling salesman problem in which we want to find the maximum scoring traversal of nodes in which each node is only visited once. While the traveling salesman problem is known to be an NP-hard problem, greedy approximation algorithms exist with fast implementations [10], and we don't need the true optimal ordering. We just need an ordering such that heterozygous SNPs near one another in the ordering are likely to share linkage information in our data.

Once we have a pseudo ordering of heterozygous SNP kmers, we could proceed in many different ways including methods which already exist [68]. Or we could approach the problem from a novel angle with some benefits. We propose to solve the haplotype phasing search problem with mixture model clustering. In this case, each cluster center would be a vector of length n where n is the number of heterozgyous kmers in some chunk which we believe should share linkage and the values in that vector would represent the allele fraction on that haplotype of one of the alleles chosen arbitrarily. When optimized, these values would tend toward 0 and 1 assuming the data supports a diploid structure.

Similar to our single cell genotype clustering algorithm, we marginalize each read across the possible haplotypes. Then the negative log likelihood should be differentiable and susceptible to numerical optimization techniques. This method has several potential advantages over current approximation search algorithms. One is that is is trivially extendable to polyploid genomes. In the polyploid case, the only thing that changes is that there are $P = ploidy$ cluster centers instead of just two. Another benefit is that it can handle cases in which one of the variants chosen is not actually heterozygous but is homozygous for one of the alleles. Including these cases can dramatically slow down current algorithms, but is beneficial in error correcting variant calls [28]. This algorithm could also be used in reference based systems.

4.5 Discussion

Chapter 5

Conclusions

References

[1] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012.

[2] Stefano Beretta, Murray D Patterson, Simone Zaccaria, Gianluca Della Vedova, and Paola Bonizzoni. HapCHAT: adaptive haplotype assembly for efficiently leveraging high coverage in long reads. *BMC Bioinformatics*, 19(1):252, July 2018.

[3] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R Ecker, Dario Cantu, David R Rank, and Michael C Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, 13(12): 1050–1054, December 2016.

[4] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[5] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[6] Drosophila 12 Genomes Consortium, Andrew G Clark, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, Therese A Markow, Thomas C Kaufman, Manolis Kellis, William Gelbart, Venky N Iyer, Daniel A Pollard, Timothy B Sackton, Amanda M Larracuente, Nadia D Singh, Jose P Abad, Dawn N Abt, Boris Adryan, Montserrat Aguade, Hiroshi Akashi, Wyatt W Anderson, Charles F Aquadro, David H Ardell, Roman Arguello, Carlo G Artieri, Daniel A Barbash, Daniel Barker, Paolo Barsanti, Phil Batterham, Serafim Batzoglou, Dave Begun, Arjun Bhutkar, Enrico Blanco, Stephanie A Bosak, Robert K Bradley, Adrianne D Brand, Michael R Brent, Angela N Brooks, Randall H Brown, Roger K Butlin, Corrado Caggese, Brian R Calvi, A Bernardo de Carvalho, Anat Caspi, Sergio Castrezana, Susan E Celniker, Jean L Chang, Charles Chapple, Sourav Chatterji, Asif Chinwalla, Alberto Civetta, Sandra W Clifton, Josep M Comeron, James C Costello, Jerry A Coyne, Jennifer Daub, Robert G David, Arthur L Delcher, Kim Delehaunty, Chuong B Do,

Heather Ebling, Kevin Edwards, Thomas Eickbush, Jay D Evans, Alan Filipski, Sven Findeiss, Eva Freyhult, Lucinda Fulton, Robert Fulton, Ana C L Garcia, Anastasia Gardiner, David A Garfield, Barry E Garvin, Greg Gibson, Don Gilbert, Sante Gnerre, Jennifer Godfrey, Robert Good, Valer Gotea, Brenton Gravely, Anthony J Greenberg, Sam Griffiths-Jones, Samuel Gross, Roderic Guigo, Erik A Gustafson, Wilfried Haerty, Matthew W Hahn, Daniel L Halligan, Aaron L Halpern, Gillian M Halter, Mira V Han, Andreas Heger, Ladeana Hillier, Angie S Hinrichs, Ian Holmes, Roger A Hoskins, Melissa J Hubisz, Dan Hultmark, Melanie A Huntley, David B Jaffe, Santosh Jagadeeshan, William R Jeck, Justin Johnson, Corbin D Jones, William C Jordan, Gary H Karpen, Eiko Kataoka, Peter D Keightley, Pouya Kheradpour, Ewen F Kirkness, Leonardo B Koerich, Karsten Kristiansen, Dave Kudrna, Rob J Kulathinal, Sudhir Kumar, Roberta Kwok, Eric Lander, Charles H Langley, Richard Lapoint, Brian P Lazzaro, So-Jeong Lee, Lisa Levesque, Ruiqiang Li, Chiao-Feng Lin, Michael F Lin, Kerstin Lindblad-Toh, Ana Llopart, Manyuan Long, Lloyd Low, Elena Lozovsky, Jian Lu, Meizhong Luo, Carlos A Machado, Wojciech Makalowski, Mar Marzo, Muneo Matsuda, Luciano Matzkin, Bryant McAllister, Carolyn S McBride, Brendan McKernan, Kevin McKernan, Maria Mendez-Lago, Patrick Minx, Michael U Mollenhauer, Kristi Montooth, Stephen M Mount, Xu Mu, Eugene Myers, Barbara Negre, Stuart Newfeld, Rasmus Nielsen, Mohamed A F Noor, Patrick O'Grady, Lior Pachter, Montserrat Papaceit, Matthew J Parisi, Michael Parisi, Leopold Parts, Jakob S Pedersen, Graziano Pesole, Adam M Phillippy, Chris P Ponting, Mihai Pop, Damiano Porcelli, Jeffrey R Powell, Sonja Prohaska, Kim Pruitt, Marta Puig, Hadi Quesneville, Kristipati Ravi Ram, David Rand, Matthew D Rasmussen, Laura K Reed, Robert Reenan, Amy Reily, Karin A Remington, Tania T Rieger, Michael G Ritchie, Charles Robin, Yu-Hui Rogers, Claudia Rohde, Julio Rozas, Marc J Rubenfield, Alfredo Ruiz, Susan Russo, Steven L Salzberg, Alejandro Sanchez-Gracia, David J Saranga, Hajime Sato, Stephen W Schaeffer, Michael C Schatz, Todd Schlenke, Russell Schwartz, Carmen Segarra, Rama S Singh, Laura Sirot, Marina Sirota, Nicholas B Sisneros, Chris D Smith, Temple F Smith, John Spieth, Deborah E Stage, Alexander Stark, Wolfgang Stephan, Robert L Strausberg, Sebastian Strempel, David Sturgill, Granger Sutton, Granger G Sutton, Wei Tao, Sarah Teichmann, Yoshiko N Tobari, Yoshihiko Tomimura, Jason M Tsolas, Vera L S Valente, Eli Venter, J Craig Venter, Saverio Vicario, Filipe G Vieira, Albert J Vilella, Alfredo Villasante, Brian Walenz, Jun Wang, Marvin Wasserman, Thomas Watts, Derek Wilson, Richard K Wilson, Rod A Wing, Mariana F Wolfner, Alex Wong, Gane Ka-Shu Wong, Chung-I Wu, Gabriel Wu, Daisuke Yamamoto, Hsiao-Pei Yang, Shiaw-Pyng Yang, James A Yorke, Kiyohito Yoshida, Evgeny Zdobnov, Peili Zhang, Yu Zhang, Aleksey V Zimin, Jennifer Baldwin, Amr Abdouelleil, Jamal Abdulkadir, Adal Abebe, Brikti Abera, Justin Abreu, St Christophe Acer, Lynne Aftuck, Allen Alexander, Peter An, Erica Anderson, Scott Anderson, Harindra Arachi, Marc Azer, Pasang Bachantsang, Andrew Barry, Tashi Bayul, Aaron Berlin, Daniel Bessette, Toby Bloom, Jason Blye, Leonid Boguslavskiy, Claude Bonnet, Boris Boukhgalter, Imane Bourzgui, Adam Brown, Patrick Cahill, Sheridon Channer, Yama Cheshatsang, Lisa Chuda, Mieke Citroen, Alville Collymore, Patrick Cooke, Maura Costello, Katie D'Aco, Riza Daza, Georgius De Haan, Stuart DeGray, Christina DeMaso, Norbu Dhargay, Kimberly Dooley, Erin Dooley, Missole Doricent, Passang Dorje, Kunsang Dorjee, Alan Dupes, Richard Elong, Jill Falk, Abderrahim Farina, Susan Faro, Diallo Ferguson, Sheila Fisher, Chelsea D Foley, Alicia Franke, Dennis Friedrich, Loryn

Gadbois, Gary Gearin, Christina R Gearin, Georgia Giannoukos, Tina Goode, Joseph Graham, Edward Grandbois, Sharleen Grewal, Kunsang Gyaltsen, Nabil Hafez, Birhane Hagos, Jennifer Hall, Charlotte Henson, Andrew Hollinger, Tracey Honan, Monika D Huard, Leanne Hughes, Brian Hurhula, M Erii Husby, Asha Kamat, Ben Kanga, Seva Kashin, Dmitry Khazanovich, Peter Kisner, Krista Lance, Marcia Lara, William Lee, Niall Lennon, Frances Letendre, Rosie LeVine, Alex Lipovsky, Xiaohong Liu, Jinlei Liu, Shangtao Liu, Tashi Lokyitsang, Yeshi Lokyitsang, Rakela Lubonja, Annie Lui, Pen MacDonald, Vasilia Magnisalis, Kebede Maru, Charles Matthews, William McCusker, Susan McDonough, Teena Mehta, James Meldrim, Louis Meneus, Oana Mihai, Atanas Mihalev, Tanya Mihova, Rachel Mittelman, Valentine Mlenga, Anna Montmayeur, Leonidas Mulrain, Adam Navidi, Jerome Naylor, Tamrat Negash, Thu Nguyen, Nga Nguyen, Robert Nicol, Choe Norbu, Nyima Norbu, Nathaniel Novod, Barry O'Neill, Sahal Osman, Eva Markiewicz, Otero L Oyono, Christopher Patti, Pema Phunkhang, Fritz Pierre, Margaret Priest, Sujaa Raghuraman, Filip Rege, Rebecca Reyes, Cecil Rise, Peter Rogov, Keenan Ross, Elizabeth Ryan, Sampath Settipalli, Terry Shea, Ngawang Sherpa, Lu Shi, Diana Shih, Todd Sparrow, Jessica Spaulding, John Stalker, Nicole Stange-Thomann, Sharon Stavropoulos, Catherine Stone, Christopher Strader, Senait Tesfaye, Talene Thomson, Yama Thoulutsang, Dawa Thoulutsang, Kerri Topham, Ira Topping, Tsamla Tsamla, Helen Vassiliev, Andy Vo, Tsering Wangchuk, Tsering Wangdi, Michael Weiand, Jane Wilkinson, Adam Wilson, Shailendra Yadav, Geneva Young, Qing Yu, Lisa Zembek, Danni Zhong, Andrew Zimmer, Zac Zwirko, David B Jaffe, Pablo Alvarez, Will Brockman, Jonathan Butler, Cheewhye Chin, Sante Gnerre, Manfred Grabherr, Michael Kleber, Evan Mauceli, and Iain MacCallum. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218, November 2007.

[7] Peter Edge, Vineet Bafna, and Vikas Bansal. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.*, 27(5):801–812, May 2017.

[8] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, January 2009.

[9] Sarah Garcia, Rajiv Bharadwaj, Stéphane Boutet, Claudia Catalanotti, Valeria Giangerra, Josephine Lee, Jessica Terry, Stephen Williams, Grace X Zheng, Tarjei Mikkelsen, Michael Schnall-Levin, Ben Hindson, and Deanna M Church. Abstract 281: Identifying genetic variation and cellular heterogeneity with a comprehensive cancer analysis toolkit. *Cancer Res.*, 78(13 Supplement):281–281, July 2018.

[10] Erik Garrison. Neighborly-tour. https://github.com/ekg/neighborly-tour, 2018.

[11] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. July 2012.

[12] Todd M Gierahn, Marc H Wadsworth, 2nd, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*, 14(4):395–398, April 2017.

[13] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.*, 25 (11):1750–1756, November 2015.

[14] Haynes Heaton. Distinguishing kmers. https://github.com/wheaton5/distinguishing_kmers, 2019.

[15] Haynes Heaton. Haplovalidate. https://github.com/wheaton5/haplovalidate, 2019.

[16] Haynes Heaton. Heterozygous snp kmer detection. https://github.com/wheaton5/het_snp_kmers2, 2019.

[17] Haynes Heaton. Long read binner. https://github.com/wheaton5/long_read_binner, 2019.

[18] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6(2):95–108, February 2005.

[19] Robert A Holt, G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, Andrew G Clark, José M C Ribeiro, Ron Wides, Steven L Salzberg, Brendan Loftus, Mark Yandell, William H Majoros, Douglas B Rusch, Zhongwu Lai, Cheryl L Kraft, Josep F Abril, Veronique Anthouard, Peter Arensburger, Peter W Atkinson, Holly Baden, Veronique de Berardinis, Danita Baldwin, Vladimir Benes, Jim Biedler, Claudia Blass, Randall Bolanos, Didier Boscus, Mary Barnstead, Shuang Cai, Angela Center, Kabir Chaturverdi, George K Christophides, Mathew A Chrystal, Michele Clamp, Anibal Cravchik, Val Curwen, Ali Dana, Art Delcher, Ian Dew, Cheryl A Evans, Michael Flanigan, Anne Grundschober-Freimoser, Lisa Friedli, Zhiping Gu, Ping Guan, Roderic Guigo, Maureen E Hillenmeyer, Susanne L Hladun, James R Hogan, Young S Hong, Jeffrey Hoover, Olivier Jaillon, Zhaoxi Ke, Chinnappa Kodira, Elena Kokoza, Anastasios Koutsos, Ivica Letunic, Alex Levitsky, Yong Liang, Jhy-Jhu Lin, Neil F Lobo, John R Lopez, Joel A Malek, Tina C McIntosh, Stephan Meister, Jason Miller, Clark Mobarry, Emmanuel Mongin, Sean D Murphy, David A O'Brochta, Cynthia Pfannkoch, Rong Qi, Megan A Regier, Karin Remington, Hongguang Shao, Maria V Sharakhova, Cynthia D Sitter, Jyoti Shetty, Thomas J Smith, Renee Strong, Jingtao Sun, Dana Thomasova, Lucas Q Ton, Pantelis Topalis, Zhijian Tu, Maria F Unger, Brian Walenz, Aihui Wang, Jian Wang, Mei Wang, Xuelan Wang, Kerry J Woodford, Jennifer R Wortman, Martin Wu, Alison Yao, Evgeny M Zdobnov, Hongyu Zhang, Qi Zhao, Shaying Zhao, Shiaoping C Zhu, Igor Zhimulev, Mario Coluzzi, Alessandra della Torre, Charles W Roth, Christos Louis, Francis Kalush, Richard J Mural,

Eugene W Myers, Mark D Adams, Hamilton O Smith, Samuel Broder, Malcolm J Gardner, Claire M Fraser, Ewan Birney, Peer Bork, Paul T Brey, J Craig Venter, Jean Weissenbach, Fotis C Kafatos, Frank H Collins, and Stephen L Hoffman. The genome sequence of the malaria mosquito anopheles gambiae. *Science*, 298(5591): 129–149, October 2002.

[20] J Jing, J Reed, J Huang, X Hu, V Clarke, J Edington, D Housman, T S Anantharaman, E J Huff, B Mishra, B Porter, A Shenker, E Wolfson, C Hiort, R Kantor, C Aston, and D C Schwartz. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.*, 95(14):8046–8051, July 1998.

[21] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, 36(1):89–94, January 2018.

[22] Sarah B Kingan, Haynes Heaton, Juliana Cudini, Christine C Lambert, Primo Baybayan, Brendan D Galvin, Richard Durbin, Jonas Korlach, and Mara K N Lawniczak. A High-Quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes*, 10(1), January 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[24] S Koren, A Rhie, B P Walenz, A T Dilthey, D M Bickhart, and others. Complete assembly of parental haplotypes with trio binning. *bioRxiv*, 2018.

[25] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27(5):722–736, May 2017.

[26] Z N Kronenberg, R J Hall, S Hiendleder, TPL Smith, and others. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*, 2018.

[27] Phanidhar Kukutla, Bo G Lindberg, Dong Pei, Melanie Rayl, Wanqin Yu, Matthew Steritz, Ingrid Faye, and Jiannong Xu. Insights from the genome annotation of elizabethkingia anophelis from the malaria vector anopheles gambiae. *PLoS One*, 9 (5):e97715, May 2014.

[28] Sofia Kyriazopoulou-Panagiotopoulou, Patrick Marks, Michael Schnall-Levin, Xinying Zheng, Mirna Jarosz, Serge Saxonov, Kristina Giorda, Patrice Mudivarti, Heather Ordonez, Jessica Terry, et al. Systems and methods for determining structural variation and phasing using variant call data, August 11 2016. US Patent App. 15/019,928.

[29] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[30] M K N Lawniczak, S J Emrich, A K Holloway, A P Regier, M Olson, B White, S Redmond, L Fulton, E Appelbaum, J Godfrey, C Farmer, A Chinwalla, S-P Yang, P Minx, J Nelson, K Kyung, B P Walenz, E Garcia-Hernandez, M Aguiar, L D Viswanathan, Y-H Rogers, R L Strausberg, C A Saski, D Lawson, F H Collins, F C Kafatos, G K Christophides, S W Clifton, E F Kirkness, and N J Besansky. Widespread divergence between incipient anopheles gambiae species revealed by whole genome sequences. *Science*, 330(6003):512–514, October 2010.

[31] Ellen M Leffler, Kevin Bullaughey, Daniel R Matute, Wynn K Meyer, Laure Ségurel, Aarti Venkat, Peter Andolfatto, and Molly Przeworski. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.*, 10(9):e1001388, September 2012.

[32] Harris A Lewin, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, Melissa M Goldstein, Igor V Grigoriev, Kevin J Hackett, David Haussler, Erich D Jarvis, Warren E Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S Soltis, Xun Xu, Huanming Yang, and Guojie Zhang. Earth BioGenome project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.*, 115(17):4325–4333, April 2018.

[33] Harris A Lewin, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, Melissa M Goldstein, Igor V Grigoriev, Kevin J Hackett, David Haussler, Erich D Jarvis, Warren E Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S Soltis, Xun Xu, Huanming Yang, and Guojie Zhang. Earth BioGenome project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.*, 115(17):4325–4333, April 2018.

[34] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, October 2014.

[35] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, May 2018.

[36] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[37] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[38] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

[39] Jacek Majewski and Tomi Pastinen. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, 27(2):72–79, February 2011.

[40] Milan Malinsky, Jared T Simpson, and Richard Durbin. trio-sga: facilitating de novo assembly of highly heterozygous genomes with parent-child trios. May 2016.

[41] Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J Clavijo. KAT: a k-mer analysis toolkit to quality control NGS datasets and genome assemblies, 2016.

[42] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, March 2011.

[43] B J Matthews, O Dudchenko, S Kingan, S Koren, and others. Improved aedes aegypti mosquito reference genome assembly enables biological discovery and vector control. *bioRxiv*, 2017.

[44] Benjamin J Matthews, Olga Dudchenko, Sarah B Kingan, Sergey Koren, Igor Antoshechkin, Jacob E Crawford, William J Glassford, Margaret Herre, Seth N Redmond, Noah H Rose, Gareth D Weedall, Yang Wu, Sanjit S Batra, Carlos A Brito-Sierra, Steven D Buckingham, Corey L Campbell, Saki Chan, Eric Cox, Benjamin R Evans, Thanyalak Fansiri, Igor Filipović, Albin Fontaine, Andrea Gloria-Soria, Richard Hall, Vinita S Joardar, Andrew K Jones, Raissa G G Kay, Vamsi K Kodali, Joyce Lee, Gareth J Lycett, Sara N Mitchell, Jill Muehling, Michael R Murphy, Arina D Omer, Frederick A Partridge, Paul Peluso, Aviva Presser Aiden, Vidya Ramasamy, Gordana Rašić, Sourav Roy, Karla Saavedra-Rodriguez, Shruti Sharan, Atashi Sharma, Melissa Laird Smith, Joe Turner, Allison M Weakley, Zhilei Zhao, Omar S Akbari, William C Black, 4th, Han Cao, Alistair C Darby, Catherine A Hill, J Spencer Johnston, Terence D Murphy, Alexander S Raikhel, David B Sattelle, Igor V Sharakhov, Bradley J White, Li Zhao, Erez Lieberman Aiden, Richard S Mann, Louis Lambrechts, Jeffrey R Powell, Maria V Sharakhova, Zhijian Tu, Hugh M Robertson, Carolyn S McBride, Alex R Hastie, Jonas Korlach, Daniel E Neafsey, Adam M Phillippy, and Leslie B Vosshall. Improved reference genome of aedes aegypti informs arbovirus vector control. *Nature*, 563(7732):501–507, November 2018.

[45] Yulia Mostovoy, Michal Levy-Sakin, Jessica Lam, Ernest T Lam, Alex R Hastie, Patrick Marks, Joyce Lee, Catherine Chu, Chin Lin, Željko Džakula, Han Cao, Stephen A Schlebusch, Kristina Giorda, Michael Schnall-Levin, Jeffrey D Wall, and Pui-Yan Kwok. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods*, 13(7):587–590, July 2016.

[46] E W Myers, G G Sutton, A L Delcher, I M Dew, D P Fasulo, M J Flanigan, S A Kravitz, C M Mobarry, K H Reinert, K A Remington, E L Anson, R A Bolanos, H H Chou, C M Jordan, A L Halpern, S Lonardi, E M Beasley, R C Brandon, L Chen, P J Dunn, Z Lai, Y Liang, D R Nusskern, M Zhan, Q Zhang, X Zheng, G M Rubin, M D Adams, and J C Venter. A whole-genome assembly of drosophila. *Science*, 287 (5461):2196–2204, March 2000.

[47] Daniel E Neafsey, Robert M Waterhouse, Mohammad R Abai, Sergey S Aganezov, Max A Alekseyev, James E Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, Lauren A Assour, Hamidreza Basseri, Aaron Berlin, Bruce W Birren, Stephanie A Blandin, Andrew I Brockman, Thomas R Burkot, Austin Burt, Clara S Chan, Cedric Chauve, Joanna C Chiu, Mikkel Christensen, Carlo Costantini, Victoria L M Davidson, Elena Deligianni, Tania Dottorini, Vicky Dritsou, Stacey B Gabriel, Wamdaogo M Guelbeogo, Andrew B Hall, Mira V Han, Thaung Hlaing, Daniel S T Hughes, Adam M Jenkins, Xiaofang Jiang, Irwin Jungreis, Evdoxia G Kakani, Maryam Kamali, Petri Kemppainen, Ryan C Kennedy, Ioannis K Kirmitzoglou,

Lizette L Koekemoer, Njoroge Laban, Nicholas Langridge, Mara K N Lawniczak, Manolis Lirakis, Neil F Lobo, Ernesto Lowy, Robert M MacCallum, Chunhong Mao, Gareth Maslen, Charles Mbogo, Jenny McCarthy, Kristin Michel, Sara N Mitchell, Wendy Moore, Katherine A Murphy, Anastasia N Naumenko, Tony Nolan, Eva M Novoa, Samantha O'Loughlin, Chioma Oringanje, Mohammad A Oshaghi, Nazzy Pakpour, Philippos A Papathanos, Ashley N Peery, Michael Povelones, Anil Prakash, David P Price, Ashok Rajaraman, Lisa J Reimer, David C Rinker, Antonis Rokas, Tanya L Russell, N'fale Sagnon, Maria V Sharakhova, Terrance Shea, Felipe A Simão, Frederic Simard, Michel A Slotman, Pradya Somboon, Vladimir Stegniy, Claudio J Struchiner, Gregg W C Thomas, Marta Tojo, Pantelis Topalis, José M C Tubio, Maria F Unger, John Vontas, Catherine Walton, Craig S Wilding, Judith H Willis, Yi-Chieh Wu, Guiyun Yan, Evgeny M Zdobnov, Xiaofan Zhou, Flaminia Catteruccia, George K Christophides, Frank H Collins, Robert S Cornman, Andrea Crisanti, Martin J Donnelly, Scott J Emrich, Michael C Fontaine, William Gelbart, Matthew W Hahn, Immo A Hansen, Paul I Howell, Fotis C Kafatos, Manolis Kellis, Daniel Lawson, Christos Louis, Shirley Luckhart, Marc A T Muskavitch, José M Ribeiro, Michael A Riehle, Igor V Sharakhov, Zhijian Tu, Laurence J Zwiebel, and Nora J Besansky. Highly evolvable malaria vectors: The genomes of 16 anopheles mosquitoes. *Science*, 347(6217), January 2015.

[48] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth. WhatsHap: Weighted haplotype assembly for Future-Generation sequencing reads. *J. Comput. Biol.*, 22(6):498–509, June 2015.

[49] Simone Picelli, Omid R Faridani, Asa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protoc.*, 9(1):171–181, January 2014.

[50] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, 93(4):641–651, October 2013.

[51] Michael J Roach, Simon A Schmidt, and Anthony R Borneman. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1):460, November 2018.

[52] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. January 2019.

[53] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T Simpson, Glen Threadgold, James Torrance, Jonathan M Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M Phillippy, Richard Durbin, Richard K Wilson, Paul Flicek, Evan E Eichler, and Deanna M Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, 2017.

[54] Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, 31(12):1111–1118, December 2013.

[55] Maria V Sharakhova, Martin P Hammond, Neil F Lobo, Jaroslaw Krzywinski, Maria F Unger, Maureen E Hillenmeyer, Robert V Bruggner, Ewan Birney, and Frank H Collins. Update of the anopheles gambiae PEST genome assembly. *Genome Biol.*, 8(1):R5, 2007.

[56] Maria V Sharakhova, Phillip George, Irina V Brusentsova, Scotland C Leman, Jeffrey A Bailey, Christopher D Smith, and Igor V Sharakhov. Genome mapping and characterization of the anopheles gambiae heterochromatin. *BMC Genomics*, 11:459, August 2010.

[57] Jennifer M Shelton, Michelle C Coleman, Nic Herndon, Nanyan Lu, Ernest T Lam, Thomas Anantharaman, Palak Sheth, and Susan J Brown. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics*, 16:734, September 2015.

[58] Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck, 3rd, Peter Smibert, and Rahul Satija. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.*, 19(1):224, December 2018.

[59] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382, May 2009.

[60] Emma C Teeling, Sonja C Vernes, Liliana M Dávalos, David A Ray, M Thomas P Gilbert, Eugene Myers, and Bat1K Consortium. Bat biology, genomes, and the Bat1K project: To generate Chromosome-Level genomes for all living bat species. *Annu Rev Anim Biosci*, 6:23–46, February 2018.

[61] Gregg W C Thomas, Elias Dohmen, Daniel S T Hughes, Shwetha C Murali, Monica Poelchau, Karl Glastad, Clare A Anstead, Nadia A Ayoub, Phillip Batterham, Michelle Bellair, Gretta J Binford, Hsu Chao, Yolanda H Chen, Christopher Childers, Huyen Dinh, Harshavardhan Doddapaneni, Jian J Duan, Shannon Dugan, Lauren A Esposito, Markus Friedrich, Jessica Garb, Robin B Gasser, Michael A D Goodisman, Dawn E Gundersen-Rindal, Yi Han, Alfred M Handler, Masatsugu Hatakeyama, Lars Hering, Wayne B Hunter, Panagiotis Ioannidis, Joy C Jayaseelan, Divya Kalra, Abderrahman Khila, Pasi K Korhonen, Carol Eunmi Lee, Sandra L Lee, Yiyuan Li, Amelia R I Lindsey, Georg Mayer, Alistair P McGregor, Duane D McKenna, Bernhard Misof, Mala Munidasa, Monica Munoz-Torres, Donna M Muzny, Oliver Niehuis, Nkechinyere Osuji-Lacy, Subba R Palli, Kristen A Panfilio, Matthias Pechmann, Trent Perry, Ralph S Peters, Helen C Poynton, Nikola-Michael Prpic, Jiaxin Qu, Dorith Rotenberg, Coby Schal, Sean D Schoville, Erin D Scully, Evette Skinner, Daniel B Sloan, Richard Stouthamer, Michael R Strand, Nikolaus U Szucsich, Asela Wijeratne, Neil D Young, Eduardo E Zattara, Joshua B Benoit, Evgeny M Zdobnov, Michael E Pfrender, Kevin J Hackett, John H Werren, Kim C

Worley, Richard A Gibbs, Ariel D Chipman, Robert M Waterhouse, Erich Bornberg-Bauer, Matthew W Hahn, and Stephen Richards. The genomic basis of arthropod diversity. August 2018.

[62] Gregory W Vurture, Fritz J Sedlazeck, Maria Nattestad, Charles J Underwood, Han Fang, James Gurtowski, and Michael C Schatz. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14):2202–2204, July 2017.

[63] Robert M Waterhouse, Mathieu Seppey, Felipe A Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V Kriventseva, and Evgeny M Zdobnov. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, December 2017.

[64] Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. *Genome Res.*, 27(5):757–767, May 2017.

[65] J Xu, C Falconer, and L Coin. Genotype-free demultiplexing of pooled single-cell RNA-seq. *bioRxiv*, 2019.

[66] J Xu, C Falconer, and L Coin. Genotype-free demultiplexing of pooled single-cell RNA-seq. *bioRxiv*, 2019.

[67] M D Young and S Behjati. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv*, 2018.

[68] Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson, and Hanlee P Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, 34(3):303–311, March 2016.

[69] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, January 2017.