Computational methods for single cell RNA and genome assembly resolution using genetic variation.

Haynes Heaton

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

August 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
August 2021

Acknowledgements

It is often by luck or some of the 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 them, I learned more in hours than in months on my own.

The Brown Computer Science teacher assistant program was where I learned to teach and lead. It is also where I learned that a topic you cannot coherently 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 University Computer Science department since its inception. 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 even after graduating. 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 balance practicality with rigor is an inspiration to me to this day.

Another influence in my life was joining 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 much of statistics as I now understand it. 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 designing algorithms. Brendan is something of a stealth super contributor to the genomics and transcriptomics world. He is not particularly well known, but the field would be dramatically worse off without him. We went on to work together at 10x Genomics and hopefully some day we will have the opportunity to work together again. Our distinct yet semi-overlapping expertises allowed us to understand both the possibilities and limitations of each other's fields. This cross disciplinary understanding and communication was responsible for some of the most productive collaborations of my life thus far.

Through another one of life's serendipitous moments, I took what I thought was a throw away interview at a stealth genomics company. I decided to walk to the interview two miles away which turned out to be four miles away and so was late, but so was the hiring manager, Michael Schnall-Levin, who didn't balk at this and picked me up on his way in and brought me to the interview. After giving my job talk, I signed the non disclosure agreements and they told me about the technology they were building. It wasn't until later that evening that the implications and possibilities started to sink in, and I became 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 and loud that you are assured of its authenticity. 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 Saxonov and Ben Hindson for leading such a great company. It was a pleasure and honor to work 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, and the papers and patents I was able to contribute to while working there, 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. 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.

There is a wise saying that you should 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 like to think 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, the arts, and have supported me in all of my endeavors. Thanks especially to my mother who has always been my biggest fan, supporter, editor, interior designer, cooking (and eating) collaborator, and overall life advisor. I think she is both correct and completely unbiased in her opinions of me. I remember I was hiking in the Big Basin redwood forest in the summer of 2015 when I received a phone call from her. She said if I wanted to get a PhD, I should probably start planning. I said "I was just thinking of that myself." And fast forward to now, here I am. Also thanks to my father, who, at 74 years old, has continued treating patients in the hospital as a cardiologist through the pandemic. He is the single most dedicated and hardest working person I know and has always been an inspiration to me. Growing up, I constantly felt the impact of his work because, without exception, he invariably treated and potentially saved the life of a family member of the person I was interacting with.

Finally I'd like to thank my cat, Kasparov, for being a very cute kitty. But also thanks to my mother for secretly getting me a cat during the pandemic.

Abstract

Genetic variation and natural selection have driven the evolutionary history on this planet and are responsible for creating us and all other life as we know it. Over the past several decades, the genomic revolution has allowed us to assess population variation across humans and other species and use that to link genotypes with phenotypes and infer evolutionary histories. In this thesis, I explore computational methods for using genetic variation to demultiplex and disambiguate complex data.

In single cell RNAseq, problems of batch effects, doublets, and ambient RNA are each sources of noise that impede our ability to infer the functional states of cells and compare them between experiments. One new popular new experimental design promising to solve each of these while also reducing experimental costs is mixturing multiple individuals' cells into a single experiment. In chapter 2, I present a method for clustering cells by genotype, calling doublets, and using the cross-genotype signal in singletons to estimate and remove ambient RNA. I compare this methods to other existing methods including one that requires *a priori* information about the genotypes, and two which do not. I find that my method outperforms each of these methods across a wide range of data parameters and sample types.

In genome assembly, the recent higher throughput and lower cost of long read sequencing has revolutionized our ability to create reference quality genomes and has revitalized the assembly community. Now, massive efforts are taking place in the Darwin Tree of Life project and the Earth Biogenome project to create reference genomes for all multicelular eukaryotic life. This will create a scientific resource for the next generation of biological science, will serve as a conservation of data that could otherwise be lost in this time of mass extinction, and will allow for a much more broad understanding of evolution and the evolutionary history of life on Earth. While much progress has been made in data quality and assembly algorithms, some problems still exist. Until recently, the DNA input requirements for long read sequencing technologies made it impossible to sequence single individuals of these species with long reads. Also, high heterozygosity makes assembly more difficult due to the inherent ambiguity between heterozygous

sequence versus paralogous sequence when confronted with inexact homology. One solution to the DNA input requirements would be to pool individuals, but this only increases the heterozygosity of the sample and reduces assembly quality. In chapter 3, we present the first high quality assembly of a single mosquito using new library preparation methods with reduced DNA requirements. This reduces the number of haplotypes to two, improving the assembly quality. In chapter 4, we further address the problems brought on by heterozygosity in assembly. I present a suite of tools that use the phasing consistency of multiple heterozygous sequences as a signal for physical linkage, thus using genetic variation to our advantage rather than as a challenge to overcome. This tool creates phased, linked assemblies and phasing aware scaffolding. Further, I provide a tool for phasing aware scaffolding on existing assemblies. This includes a novel haplotype phasing algorithm with some unique beneficial properties. It is robust to non-heterozygous variants as input and can detect and correct those genotypes. And it naturally extends to polyploid genomes.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

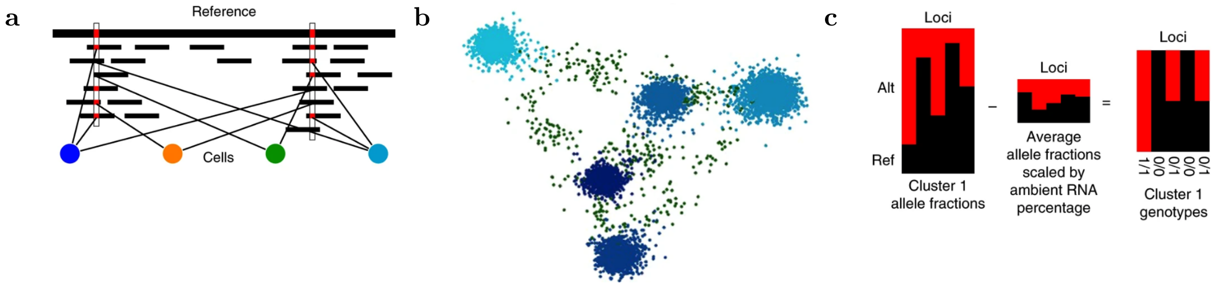
1.1 Genetic variation

Even before Gregor Mendel discovered the rules of genetic inheritance[229], the discovery that DNA was the molecule responsible for this[18], or its structure was known[331], humans have wondered at the variation among each other and all organisms. These discoveries have since made way for a rapid expansion in our ability to measure genetic variation from capillary sequencing[205] and single nucleotide polymophism (SNP) chips[203] to modern high throughput short read[24] and long read DNA sequencing[337]. We have sequenced thousands of individual humans and other organisms and explored the genetic variation of the human species [3][17][222][38][143] and used the genetic variation in population samples to impute population structure and evolutionary history[152][176][300][250][65][33]. In this thesis, I explore computational methods for using genetic variation to resolve mixtures of haplotypes in single cell RNA sequencing(scRNAseq), genome assembly, and scaffolding.

1.2 Introductiion

In single cell RNAseq (scRNAseq), the goal is not only to measure the transcriptome of many cells at a time, but also to compare the transcriptome of cells of different individuals or under different conditions such as disease state, pharmaceutical intervention, or a wide range of environmental differences. One problem with these comparisons is that there can be technical artifacts, or batch effects, between different experiments that bias the comparative results possibly even dwarfing the actual biological differences one is trying to measure. Additionally, scRNAseq has several sources of noise or errors (discussed in more detail in section 1.3.6). One is when two or more cells are erroneously partitioned into the same compartment and the data for what is supposed to be one cell is actually two cells (doublets). And another source of noise is when RNA from previously lysed cells that are in solution with the cell suspension is sequenced along with the RNA from an intact cell and those reads are given the same cellular barcode (ambient RNA). One solution to all of these problems is to pool the cells from different individuals into a single experiment. In Chapter 2 of this thesis, I present a method for demultiplexing cells from mixtures of individuals into their individual of origin using the genetic variation measured in the scRNAseq reads without requiring prior knowledge of the genotypes. In addition, I show how these mixtures can improve doublet detection and ambient RNA estimation and removal.

Fig. 1.1: Outline of single cell clustering by genotype



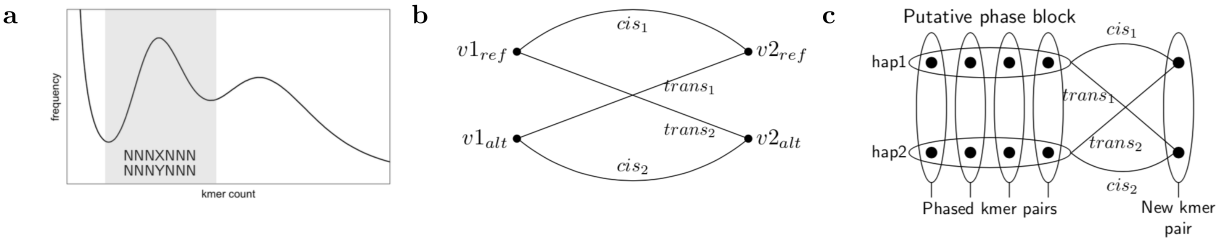
a) I find the variants in the reads from each cell barcode, **b)** cluster cells by their allele content and identify doublets, and **c)** use the bias in allele fraction from expected values to estimate and remove ambient RNA.

In genome assembly, the goal is to use the overlapping read sequence similarity to infer that those reads came from the same locus in the genome and build contiguous sequences (contigs) that represent (in part or whole) the organism's chromosomes. The inference that these reads originated from the same genomic locus is complicated by repeats, heterozygosity, and sequencing errors. With inexact homology, one must disambiguate whether the differences arose from errors, paralogous repeat sequences, or from the alternate haplotype. If one cannot make this distinction and no reads span this region into unique sequence, the contig must be broken, resulting in a fragmented assembly. If the contig is not broken, one risks a chimeric misassembly of sequences that are distant to one another being assembled together. In chapter 3 I discuss the first high quality assembly of a single mosquito. Prior to this, DNA requirements for long read sequencing (discussed in section 1.4.2) were too high to extract enough high molecular weight (HMW) DNA from many small organisms including mosquitos. This required pooling multiple individuals together in order to meet the DNA requirements for these sequencing technologies. Pooling individuals increases the number of haplotypes in the extracted DNA and makes distinguishing repeat from heterozygosity harder. Through recent advances in library preparation, the DNA requirements for long reads has been greatly reduced. By sequencing a single mosquito with long reads, we reduce the number of haplotypes from many down to two thus decreasing the potential ambiguities that arise from heterozygosity. I then compare this genome assembly to the current gold standard assembly of *Anopheles gambiae* that was created using bacterial artificial chromosome (BAC) Sanger sequencing (discussed in section 1.4.1.1), a dramatically higher cost method of creating high quality genome assemblies. We show many improvements in our assembly over the previous gold standard as well as highlight several issues that remain with the (then) current assembly state of the art.

I continue to address the problem of heterozygosity in chapter 4 by showing several ways in which haplotype phasing consistency can be used as a signal for physical linkage. Given two or more proximate heterozygous loci, sequencing reads containing them should segregate into distinct groups according to which alleles they contain. But how do we know that a site is heterozygous? Initially we do not. We can find sequences which have inexact homology each being read roughly half (assuming diploid) of the times homozygous sequences occur. These could be due to heterozygosity or paralogous sequences (if both loci are under sampled by random chance). In both cases, reads containing multiple of these alleles should segregate into two (assuming copy number of the repeat is two) groups. But when comparing a true heterozygous site with an inexact homologous sequence caused by paralogous sequence, the reads with both alleles of the heterozygous site will contain one of the presumed alleles caused by the repeat sequence. We can use this property to avoid misassemblies, create phased assemblies, and scaffold contigs in a phasing aware fashion. In chapter 4, I outline how we identify *de novo* candidate heterozygous sequences, define the phasing consistency criteria, build a phased assembly graph, and perform phasing aware scaffolding of contigs. If we wish to phasing aware scaffold an existing assembly, we must first phase its haplotypes. For this reason, I also provide an algorithm for haplotype phasing. This tool has the added benefit of being robust to being given non heterozygous sequences as input and can use the phasing inconsistency to correct those genotypes. I demonstrate these techniques on data from the butterfly *Vanessa atalanta*.

The remainder of this chapter contains the background and context for the work briefly described above. I first cover the history of single cell sequencing and analysis. I

Fig. 1.2: Phasing consistency as a *de novo* signal for physical linkage



a) We use the kmer count spectra to determine candidate heterozygous kmer pairs which we can then **b)** assess for phasing consistency based on the alleles on reads that contain one of each, and **c)** build a phased assembly graph.

then outline the biases and errors that occur in single cell sequencing and the various solutions and their downsides which motivates chapter 2. I then cover a short history of DNA sequencing methods as well as assembly and scaffolding algorithms. I discuss the inherent ambiguities that can occur and the errors that can arise and their causes, which motivates chapters 3 and 4.

1.3 Single Cell RNAseq

The etymology of the word cell comes from the latin *cella* meaning storeroom or chamber. These entities separate the physical space into compartments which interact selectively with their environments. This partitioning the cell provides is necessary for life due to the second law of thermodynamics and the nature of life. In Erwin Schrodinger's classic lecture series and book titled "What is Life"[290], he noted that while closed physical systems will always tend toward increased entropy (stated by the 2nd law of thermodynamics[41][153][221]), life must maintain (on average) a neutral or negative entropy[1] in the portion of the system in which it resides[347][126][235]. In order to do so, this requires the expenditure of energy. Biological evolution found an economical way of solving this problem with the bilipid membrane with various embedded molecules giving it the property of semipermeability—allowing some molecules in and not out or vice versa in a dynamic fashion. These cells proved, over time, to be so successful as to become the primary unit and building block of biological life on this planet.

[1] Incidentally, Erwin Schrodinger, as the father of quantum mechanics, along with Josiah Gibbs, as the father of statistical mechanics, will appear again later in this thesis as some of the algorithms described take ideas inspired by these fields for search strategies in optimization problems.

When studying the state of a cell and its current function, we could try to measure many different properties such as the proteome, transcriptome, genome, chromatin accessability, environmental conditions (such as hormone content, pH, etc), cell surface proteins, etc. But we are somewhat limited by the tools available, and when addressing function, we first turn to the central dogma of molecular biology[58] that states that information in general passes from DNA to RNA and then to proteins. If we could easily inspect the protein content of many cells in a high throughput fashion, that would be desirable, but protein detection and sequencing methods are often only limited to one or a few proteins at a time and/or are not high throughput[55][302][315][36][107][14][192][7]. But we do have a very high throughput detector for DNA and thus RNA by converting it to complementary DNA (cDNA) via reverse transcription.

1.3.1 Bulk RNAseq

Scientists have been sequencing cDNA libraries of RNA isolated from many cells mixed together since the advent of next generation DNA sequencing technologies (discussed in section 1.4.1.2) became available[20][236]. Because these use extractions from pools of cells, they are denoted as ‘bulk’ RNAseq. These experiments have driven many biological discoveries, but for some applications their usefulness is limited because they represent the average transcriptome across a population of potentially diverse cells. This blurs the data and makes inference on minority cell types difficult if not impossible. The amount of RNA in a human single cell is roughly 10-30pg[46] that until recently was not enough to create a complex cDNA library even with amplification. Some researchers have isolated specific cell types using Fluorescence Activated Cell Sorting (FACS)[56][93] prior to RNAseq with some success[277], but due to FACS cell stress and it only accessing one cell type at a time it was of limited utility.

1.3.2 Single cell RNA sequencing

In the past decade, technical advances in methods for the preparation of samples containing minuscule amounts of nucleic acids have made it possible to study the transcriptome of single cells[310]. This has changed the way biologists could access the functional state of individual cells within a complex and diverse population of cells in tissues across different states of organisms, shedding light on the cellular response to diseases, drugs, development, and more.

There are many types of RNA in the cell including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), micro RNA, and small nucleolar RNA, and non coding RNA. mRNA makes up only 3-7% of the cell's total RNA by mass[252], but it is what is translated by ribosomes into proteins, that conduct a large amount of the function of the cell. scRNAseq targeting other types of RNA have also been developed for alternative types of RNA for specific purposes[90], but in this thesis, scRNAseq is referring to a system that enriches for mRNA by using the 3' polyadenylation most mRNAs have (with some exceptions[345]).

1.3.3 Technologies

mRNA from a single cell was first isolated and amplified to measurable levels with polymerase chain reaction (PCR)[239][238] in the early 1990s[29] before the sequencing revolution. Without a high throughput detector, and due to the exponential nature of PCR causing abundant mRNA to dominate the sample, very few genes were detected. Detection improved with linear amplification achieved by multiple cycles of transcription of antisense RNA from the initial cDNA using T7 RNA polymerase[322] and the advent of oligonucleotide microarrays[203] allowing for RNA microarray studies detecting over a thousand genes[180][87]. However, the number of genes measured were still a fraction of the total genes expressed. The first sequencing of single cell cDNA sequencing increased the number of genes detected by 75%[310] and provided a hypothesis free measurement of the transcriptome a single cell. Since then, scRNAseq has grown greatly in the number of cells processed per experiment, the number of genes detected per cell, and its uptake by the scientific community[309].

Current scRNAseq protocols convert RNA to cDNA using a reverse transcriptase primed off of the poly-A tail of the mRNA. In this process, a cellular barcode oligo as well as template switching oligo (for subsequent PCR) are added to the construct. Often, a unique molecular identifier (UMI) is also added. This is used to determine which cDNA molecules were amplified from the same RNA source molecule to avoid counting PCR duplicates multiple times[169][145].

To deliver a barcode oligo to all of the reads originating from one cell and different barcodes to different cells, physical separation of one form or another is generally used. The separation could be in different tubes, plate wells, nanowells[106][27][89], or more recently with microfluidic systems creating reverse emulsion droplets[211][170][355]. These methods vary in several parameters including the scalability of number of cells per experiment, the mRNA capture rate, and the technical variation between experiments among others[308][141][356]. Which method is best to use ultimately depends on the biological question. For experiments where sampling the whole population of cells is important, droplet and nanowell methods are better whereas if capture rate and amount of data per cell is paramount, plate based systems might be more appropriate.

1.3.3.1 10x Genomics

In this thesis, all of the single cell data presented was generated using the 10x Genomics Chromium platform, a reverse emulsion droplet based system. Figure 1.3 outlines how this system works. It uses a microfluidic system to deliver gel beads, reagents, and cells into reverse emulsion droplets where the reverse transcription occurs. The gel bead contains a construct with the cell barcode oligo, a UMI, and adapters for Illumina sequencing[355][2].

1.3.4 Analysis of scRNAseq data

This construct is then sequenced on a next generation Illumina sequencer [24][110] (discussed in detail in section 1.4.1.2) giving paired-end reads with one read containing the cell barcode and UMI and the other read containing the cDNA of the mRNA transcript. Other sequencing platforms have been used, notably long read sequencing platforms, to get the whole transcript read to investigate alternative isoforms, but will not be discussed further here[115][155][299]. The most common software package used to produce cell by genes expression counts matrix and initial cell type clustering is cellranger[255] and while alternatives exist[247], they do largely the same steps.

1.3.4.1 Genome alignment

First, the template switch oligo and polyadenylation are trimmed from the 5' and 3' ends of read two respectively. Then the read two is mapped to the given reference genome using the STAR splicing aware RNA aligner[75]. Other aligners exist for this purpose such as HISAT2 [163] and TopHat2 [162]. Also there are psuedo-aligners (Kallisto[30] and Salmon[253]) which are much faster and robust to sequencing errors but do not provide base level alignments[342]. The reads are then marked as exonic or intronic using the given transcript annotation gene transfer format (GTF) file and confident or not

[2] I was fortunate to have worked at 10x Genomics between 2014 and 2017. While I did not work much directly on the single cell technology (I primarily worked on the company's first product, linked reads), I gained much insight into the data by being present for its creation.

Fig. 1.3: 10x Genomics single cell RNAseq

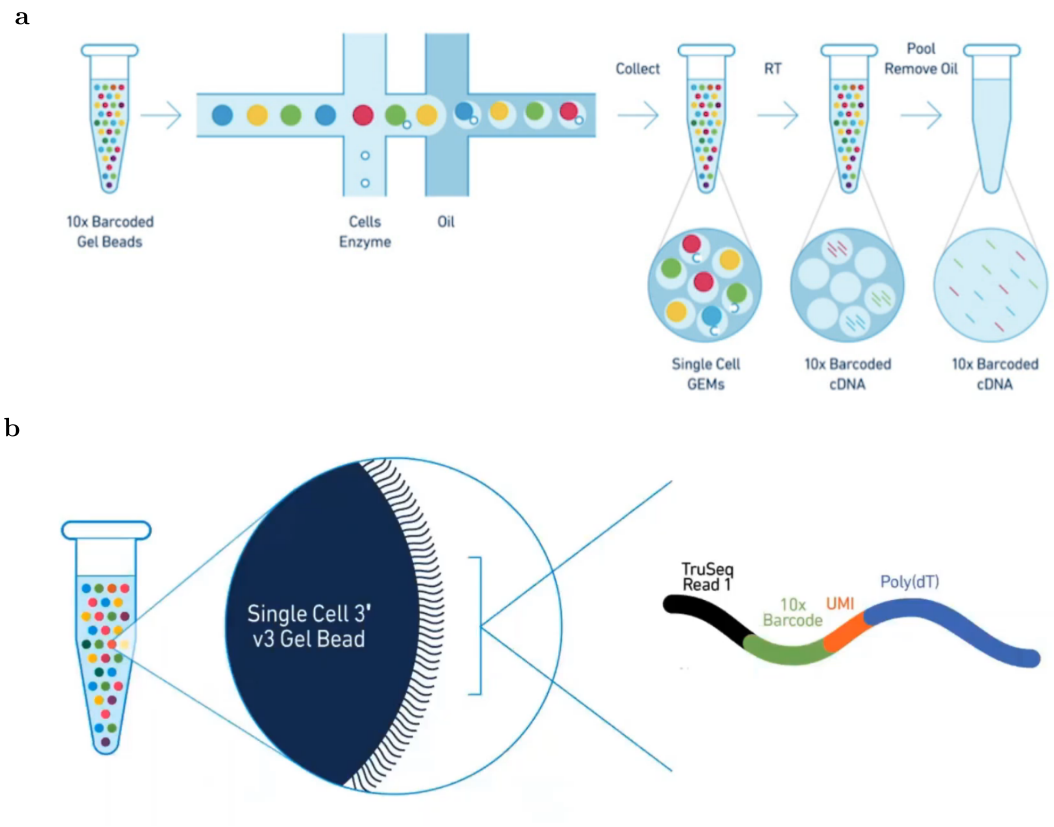


Diagram outlining 10x Genomics single cell sequencing technology (image credit: 10x Genomics website[4]).

based on if the read overlaps an exon for >50% of its length and if the mapping quality (mapq) is 255 which indicates that the read aligned uniquely. Confident exonic reads are carried forward to the UMI counting step[218].

1.3.4.2 Barcode correction

Before counting UMIs, cellranger attempts to do barcode correction on the cell barcodes. 10x Genomics uses a designed barcode set of either 737k or 3m barcodes each with a hamming distance[117] of at least two to any other barcode in the set. The barcodes that make up this designed set are the barcodes we expect to see and this set is termed the whitelist. In order to error correct barcodes, first the frequency of each barcode in the whitelist is counted. Then for every barcode that is not in the whitelist, each barcode that is one hamming distance from this sequence and is on the whitelist is found. A posterior probability is computed with the priors set by the frequency of that whitelist barcode and the base quality of the changed base used to determine likelihood of that error. For a barcode correction to then take place, the posterior probability of that whitelist barcode must be over 97.5%[217].

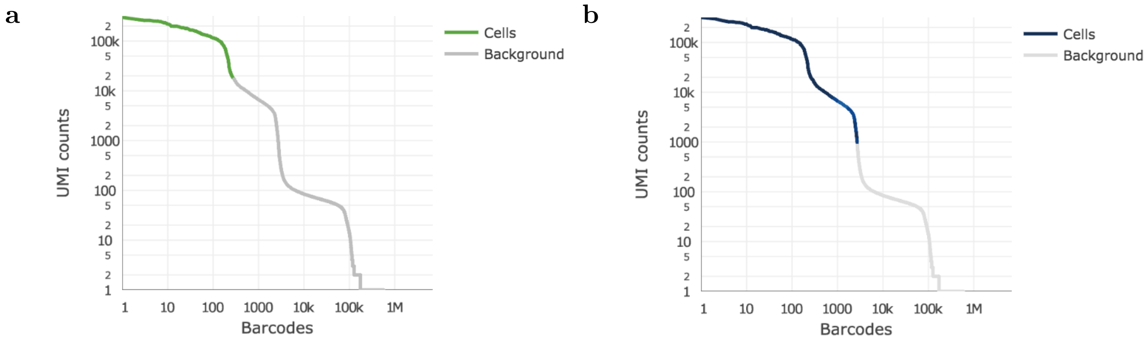
1.3.4.3 UMI counting

PCR duplicates are then removed. If any reads have cell barcode, gene, and UMI, all are removed except one. The remaining reads will be counted to create the cell barcode by gene count matrix.

1.3.4.4 Cell-barcode detection

Note that not each cell barcode contains a cell. Due to ambient RNA in solution from cells lysed before reverse emulsion partitioning, droplets without an intact cell will have some reads. We next need to determine which cell barcodes have cells and which do not. Initially, droplets containing cells were called using the second derivative of the log-log UMI counts by barcodes plot (see figure 1.4). More recently, a method using the RNA content of the confidently empty droplets called EmptyDrops[207] was developed to compare that RNA content (which is generally an average of the RNA content of all cells assuming each cell type lyses with equal probability) with the RNA content of the cells to determine an appropriate cutoff where the transcription profile from the average is dramatically different. This particularly helps in situations where the sample contains some cells with a large amount of transcripts and another cell type with many fewer transcripts. Both of these cell types will likely still have a very different transcriptional profile than the empty droplets. This algorithm has now been implemented in cellranger. The raw cell barcode by genes UMI counts matrix is then filtered to retain only cell containing cell barcodes.

Fig. 1.4: Cell barcode detection knee plot old vs new algorithm



Log-log barcode by UMI count knee plots showing which barcodes were determined to contain cells under the **a)** old method using 2nd derivative and the **b)** new EmptyDrops method (image credit: 10x genomics website[4]).

1.3.4.5 Quality control

Many of the following steps are done in software packages downstream from cellranger that aim to implement many types of analyses for scRNAseq data. The most popular of these software packages is Seurat[286][119] and while alternatives such as Monocle[269] exist, a comparison is out of the scope of this thesis.

In the process of cell dissociation, liquid handling, and partitioning, some cells may be damaged. For this reason, many researchers use different criteria to remove these poor quality cells. Some also use different criteria for different cell types and so a less stringent global filter may be applied prior to cell type detection and further filtering. These criteria include the number of genes per cell, percent mapped reads, percent reads that map to spike in controls, percent mitochondrial reads, and percent of reads that are PCR duplicates. While these are reasonable markers for dead or dying cells[251][142], it is my personal opinion that this type of quality control should be limited and determined at the experimental design stage to prevent unintentional bias in the results due to how these thresholds are chosen. Of course, there will always be a trade-off between unbiased data and clean data for downstream analysis and each application may require differing levels of quality control.

1.3.4.6 Normalization

As previously discussed, individual cells have extremely small amounts of mRNA and require methods to amplify this material in order to be made into a cDNA library and sequenced. These methods, along with the innate difficulties of measuring such a small amount of starting material inevitably result in some technical artifacts. Genes that are expressed to a lesser degree than other genes may show zero counts or lower than true counts in the experiment for several reasons[31]. Capture rate of mRNA in the reverse transcription step will never have complete yield and may vary from cell to cell and gene to gene. Additionally, genes that are expressed might be made into cDNA and amplified, but not sampled in the sequencing step as transcripts that begin with a higher copy number get amplified more in the exponential PCR step. Differences in cell size, and thus mRNA content, may result in sampling of genes in one cell type not sampled in other cell types even if both express them. To address this, many normalization methods have been developed[327][209]. Spike in controls can be used to improve this normalization[319] but takes up valuable sequencing. Another solution is imputation[131], but this can introduce unwanted false positives[12]. In comparison papers, differential expression analysis has been shown to be the downstream application most sensitive to these methods[210], with scran[206] performing the best of those tested.

1.3.4.7 Visualization

In order to visualize this high dimensionality data, we must project it into two or three dimensions in a way that preserves the biologically interesting structure at multiple scales. First, a principal component analysis (PCA) of the filtered cell by gene counts is done to find the most meaningful features and reduce the dimensionality from cells by genes to cells by M where M is a user settable value[19]. For most experiments, the complexity of the transcriptional profile cannot be easily gleaned by looking at the first two-three principal components visually, so the next step is to use non-linear dimensionality reduction techniques to bring the data into a visually informative two or three dimensional space. The two most popular methods for this are the t-Stochastic Neighbor embedding (t-SNE)[320][321][127] and the Uniform Manifold Approximation and Projection (UMAP)[227]. Both of these methods aim to preserve pairwise distances in the final projections, but are parametric and non-deterministic (without a fixed psuedorandom number generator seed). There is an inherent trade-off between how well distances should be preserved at different scales, which the parameters can help guide. However, due to the randomness and parametric nature of these algorithms, it can lead some researchers to use them to bias the results toward the expected outcome of their hypothesis. Nonetheless, these are powerful techniques to understand highly dimensional data such as single cell RNAseq. Recently, UMAP has grown in popularity over t-SNE because it has been shown to better preserve pairwise distance due to the improved initialization strategy employed in the primary implementation and is more computationally efficient than t-SNE[23][171]. These projections are often fed into downstream analysis such as clustering and lineage reconstruction. It is not clear that clustering in this space is better than clustering on the raw data, PCA, random projection[151], or other dimensionality reduction space, but they are likely to look more visually correct in the UMAP or t-SNE space when both the clustering and visualization is in the same projection. An analysis of this observation is out of the scope of this thesis.

1.3.4.8 Cell type clustering and annotation

It is useful to group similar cells together for the purposes of cell type annotation and cell state detection. Because individual cells may not have enough UMIs sequenced, these analyses may not be possible on an individual cell basis whereas grouping similar cells together will pool enough data to conclusively do so. Clustering is typically done on the dimensionality reduced data (either PCA or UMAP/t-SNE) and many methods have been used including K-means, hierarchical clustering, graph based methods and meta-heuristics have been applied including consensus clustering, cluster trees among others[13][35][167][349]. These methods are reviewed in [13] and otherwise a comparison of these methods is out of the scope of this thesis.

Once cells have been clustered into similar groups, we can try to understand what each of these groups of cells represent. Marker genes have been studied for many decades to identify and differentiate different cell types. One can visually display the expression values of these marker genes versus the cell clusters to find the cell types of interest. Increasingly more popular are automatic methods which use annotated cell atlases to match cell types. These include scMatch[130], cellHarmony[69], Garnet[261], scPred[9] with some using prior knowledge of marker genes and some not. A comparison of these methods found that they work fairly similarly with scPred performing the best overall. Interestingly, they find that prior knowledge of marker genes does not improve performance. Other methods allow you to project cells from one dataset onto another[168].

For systems which have robust prior annotated datasets, these are powerful and accurate tools for automatic annotation.

1.3.5 Downstream analysis

Many further analyses on single cell experiments are possible and a comprehensive review of these is out of the scope of this thesis. Pseudotime analysis can order cells along some cell state change such as differentiation or cell cycle[280][102]. Gene regulatory networks may be inferred using the correlation of genes indicating they may be under similar regulatory control[6]. Somatic mutations in the mitochondrial genes can be used to discover cell lineages[72]. And many more analyses are possible especially if the experimental design is non-standard such as multi-Omic single cell sequencing or CRSPR-Cas9 screening[73] is added to the mix.

1.3.6 scRNAseq error modes

1.3.6.1 Batch effects

Due to the manner in which scRNAseq data is created, it naturally has certain noise and error characteristics. In section 1.3.4.6, we discussed intra-dataset technical artifacts. Naturally, inter-dataset technical artifacts are larger in magnitude and more diverse. If any of the laboratory protocols were changed, or the experiment was done by a different person, on a different day, at a different temperature, it may introduce inter-dataset differences that may be even larger than the biological differences we wish to measure. There are a multitude of computational methods to correct these batch effects such as scanorama[125], mnnCorrect[116], BBKNN[262], Harmony[177], Seurat[306], and LIGER[336]. Each of these either finds matching cell populations or overall data correlations to then create a projection to bring the datasets into a common space. A comparison study tested 14 of these methods and evaluate the adjusted rand index of cell type clustering, average silhouette width of cell type clustering, and two other metrics and found that Seurat, LIGER, and Harmony performed best. While these are powerful tools for correcting these technical variations, it is possible that in trying to correct for variation due to batch effects that some biological differences will also be erased or biased.

1.3.6.2 Doublets

In 10x Genomics scRNAseq, the loading of droplets with cells is a random process that follows a Poisson sampling distribution. The experiments are designed with a cell suspension concentration to produce a poisson distribution with a mean much less than one. This results in an experiment in which most droplets sample zero cells and some sample one cell. But in order to collect enough cells, that mean still must not be vanishingly small. So some droplets will sample more than one cell. These cell barcodes associated with more than one cell are generally called doublets though multiplets might be more appropriate as some may have more than two cells. Another way that these arise is if the tissue dissociation of cells was not complete or if there is any cell adhesion causing cells to travel together in suspension. Again, a number of computational tools have been developed to find and remove these doublet cell barcodes from further analysis. These include doubletCells[208], DoubletFinder[225], Scrublet[341], DoubletDecon[70], scDblFinder[231], and Solo[26] among others. Many of these use simulated doublets by combining *in silico* the UMI counts of putative singleton cells to identify what the transcriptional profile of a doublet would look like. Validating these methods is somewhat problematic as in real datasets, the doublets are not known, and simulated doublets may not exactly match what data would look like from true doublets. In a recent benchmark of these methods, many of these performed similarly with scDblFinder scoring the highest overall[343]. Once again, these are powerful methods, but do not work flawlessly and in particular may remove cells in an intermediate state transition between two more distinct cell types in the sample so may remove cells of potential interest.

1.3.6.3 Ambient RNA

Another aspect of scRNAseq data that biases our view of the transcriptional landscape is ambient RNA. Before the cells are partitioned, some cells may have lysed, or there may be other cell free RNA in solution. This RNA will be delivered to all droplets including droplets that contain a cell and that RNA will be sequenced with the same cell barcode as the reads that truly came from the cell. This is alternately called ambient RNA or the ‘soup’. Ambient RNA can be analyzed by looking at the reads from non-cell barcodes and is generally an average of all of the RNA in the experiment, but this is not always true such as when some cell types are more prone to lysing than others or samples such as necrotic tumor. The amount of ambient RNA in the system is generally small, but may be increased in some samples such as tissues that requires harsh detergent agents to dissociate the cells into a cell suspension. SoupX was developed in order to estimate the amount of ambient RNA and remove it[348], but requires prior knowledge of gene expression in different cell types as it uses measurement of genes known to not be expressed in certain cell types to measure the ambient RNA.

1.3.6.4 Mixtures

One experimental design that promises to solve all three of these error modes in scRNAseq are mixture experiments. If cells from multiple samples are mixed together, you limit the number of technical artifacts between them to differences in how the cells were treated prior to being mixed. If you can distinguish which cells came from which samples, you can use that same signal to determine which cell barcodes represent cross-sample doublets. And depending on the sample features, this may also aid in measuring ambient RNA. Several experimental methods have been developed to tag cells by sample prior to mixing. Cell hashing uses oligonucleotide tagged antibodies attached to cell surface proteins as a sample signal[304][100]. MULTI-seq uses lipid and cholesterol modified oligonucleotides which incorporate into any lipid membrane to generate a sample read-out[226]. CellTag uses heritable genetic material as a sample index for tracking cells through passages after mixing to better study sample interactions[113]. These are powerful methods for reducing batch effects, detecting doublets, and reducing costs through both multiplexing and giving more scope for overloading the number of cells per experiment. As the number of cells per experiment is increased, the number of doublets grows as well. If one has a robust method of identifying and removing doublets, one can load more cells and recover more singletons even after removing the doublets. However, there is a limit to this. 10x Genomics reports that the number of doublets is roughly 1% per thousand cells recovered. This is of course an oversimplification. If the cell suspension is fully dissociated, the cell loading is a Poisson process. In the range of two to ten thousand cells, this generates roughly 1% doublets per thousand cells. Outside of this range, this rule falls apart. One can, however, fit a Poisson to a number of different experiments with differing number of cells for a better model. Over some number, the marginal increase in singletons recovered as more cells are added decreases and at some point actually diminishes. This is also worsened by the fact that the doublets and multiplets take up valuable sequencing only to be discarded. Additionally, doublet detection methods are not perfect and the remaining doublets will bias your experimental results.

But these methods require additional experimental work and are not always possible. Some mixture samples are naturally occurring such as at the maternal/fetal boundary, transplant patient tissue, or complex infections. If the mixed samples have distinct genotypes, one can use the genetic variation between samples to demultiplex them. Demuxlet was first developed for this purpose, but requires prior knowledge of the genotypes of each sample and its rigid model based system can make errors especially as the amount of ambient RNA in the system increases[291][154]. In chapter 2, I present souporcell, a method for clustering cells by genotype without prior knowledge of each sample, cross sample doublet detection, and ambient RNA measurement. We compare our system against Demuxlet and two other methods, scSplit and vireo[344][138], across a wide range of challenging datasets. Since then, freemuxlet was developed as another such method, but is not compared to as it came out later and is unpublished outside of a thesis[352].

1.4 Genome assembly and scaffolding

Since Mendel's discovery of the laws of heritability[229], it has been a goal to link the micro to the macro to explain evolution in a quantitative fashion[140]. The discovery that DNA encoded the hereditary information of organisms[18] and subsequent discovery of its chemical structure[331] made clear the nature of information storage in a linear polymer and mechanism for stable replication. Even before the discovery of mRNA[32][111], Francis Crick hypothesized that nucleic acids direct the synthesis of proteins[59] and later elucidated what is now known as the central dogma of molecular biology[57]. In brief, this states that the information flow of an organism is through the DNA being transcribed into RNA and the RNA translated into proteins which perform most of the functions of the cell. With the information source being the DNA, this made clear the importance of reading the sequence. And for several decades, our ability to read DNA sequence has dramatically increased in both amount and accuracy.

1.4.1 DNA sequencing

The history of DNA sequencing is generally thought of as having three waves—Sanger sequencing, next generation sequencing, and third generation sequencing. Sanger sequencing is highly accurate and produces relatively long reads at 500bp to 1kb but is not highly scalable. Next generation sequencing produces high accuracy short reads (initially 35bp, but now can be up to 250bp) and is massively scalable. At the same time, many other sequencing technologies came along without much success. In the third wave, the ability to sequence long reads of single molecules without amplification has transformed genome assembly. PacBio and Oxford Nanopore Technologies (ONT) developed single molecule long read sequencers that produced low accuracy (85% and ≈90% respectively) long reads limited mostly by the input DNA length and stability. More recently, PacBio utilizes circular consensus sequencing (CCS) to produce reads in the 5-25kb range with high accuracy. Due to the vast differences in application performance long accurate reads bring, it could be argued that this represents the fourth generation of DNA sequencing.

1.4.1.1 Sanger sequencing

The ability to sequence proteins and certain RNA molecules came before the ability to sequence DNA due to proteins being made of more diverse monomers and RNA not being complicated by a complementary strand[282][122]. In 1965, Robert Holly and Frederick Sanger developed two related methods for sequencing RNA[128][283]. These were labor intensive and Sanger's method employed dangerous radioactive material. This method was then extended to DNA in 1973 and used to sequence 50 bases of the phage f1[284]. Eventually, the use of polyacrylamide gel electrophoresis, chain termination chemistry with dideoxynucleotides, the use of flourescence instead of radiolabeling, and automation brought us to what is know known as Sanger sequencing[285]. This technology was automated and became the most popular method of sequencing for many years[139]. These sequences are highly accurate as each base signal is the result of the termination of a many molecules and have read lengths from 500bp to 1000bp which is limited by reaction efficiency requiring a fraction of chain terminations at every base of the sequence. This method requires clonal DNA and thus laboratory methods were developed for creating libraries for sequencing. BACs and YACs [233] were developed and each end could be sequenced creating a mate pair read spanning hundred of kilobases giving long genetic distance linking information.

1.4.1.2 Short reads

In the poorly named "next generation" phase of DNA sequencing, there were many technologies created, but ultimately one became the dominant one and discussing each in detail is out of the scope of this thesis. In the late 90s, Solexa (later acquired by Illumina) created the Genome Analyzer in which DNA attaches to a primer on a flow cell surface and is amplified into clonal arrays of single stranded DNA[314]. This is achieved by what is known as bridge amplification. The DNA has two different primers attached during library preparation that correspond to two oligos on the flow cell. Initially, a polymerase creates the reverse strand and the double stranded DNA is denatured and the forward strand is washed way. The reverse strand then bends over (aka bridges) to attach to both oligos and the forward strand is synthesized. This bridge is then denatured resulting in forward and reverse single strands attached to the flow cell. This is then repeated multiple times and in the end one of the oligos on the flow cell is cleaved and those strands are washed away leaving many copies of only one strand of the DNA[24]. Fluorescently labeled dNTP are added and each DNA colony is then imaged. The terminator group is then removed[40] and the process is repeated for each base. Because this method uses the synthesis of the second strand as the mechanism for sequencing, it is often referred to as 'sequencing by synthesis'.

Initially the read length was limited to 35bp but over the years this has increased to 150bp on the high throughput sequencers and 250bp on the lower throughput MiSeq. The read length is limited by reagent stability as well as phase problems. If not every molecule gets extended at every step, eventually the signal will degrade until eventually it is impossible to tell which base is the correct one. Despite the short read length, paired-end reads are made possible by having a longer DNA insert than the read length and after reading one end, bridging the DNA on the flowcell and sequencing the other end. This technology produces highly accurate reads at roughly 0.1% error rate. While many other sequencing technologies emerged in the same time frame, Illumina's was much higher throughput and was highly accurate and has few context specific errors when compared with the others[8]. Illumina sequencing has made up the vast majority of DNA sequencing to this day, but other third generation technologies are increasing in utilization.

Several methods build on next generation sequencing to add further information. Moleculo[223], contiguity preserving sequencing (cpt-seq)[11], long fragment read (LFR)[224], 10x Genomics linked reads[354], and more recently haplotagging[228] use various methods for making short reads from long molecules and tagging each short read with a barcode specific to the HMW DNA molecule of origin. Strand-seq uses bromodeoxyuridine labeled DNA to degrade one strand of DNA which can be useful for haplotype phasing[88][265][101][263]. In chapter 4, I use 10x Genomics linked reads which I outline in detail in section 1.4.1.4.

1.4.1.3 High throughput chromatin conformation capture (Hi-C)

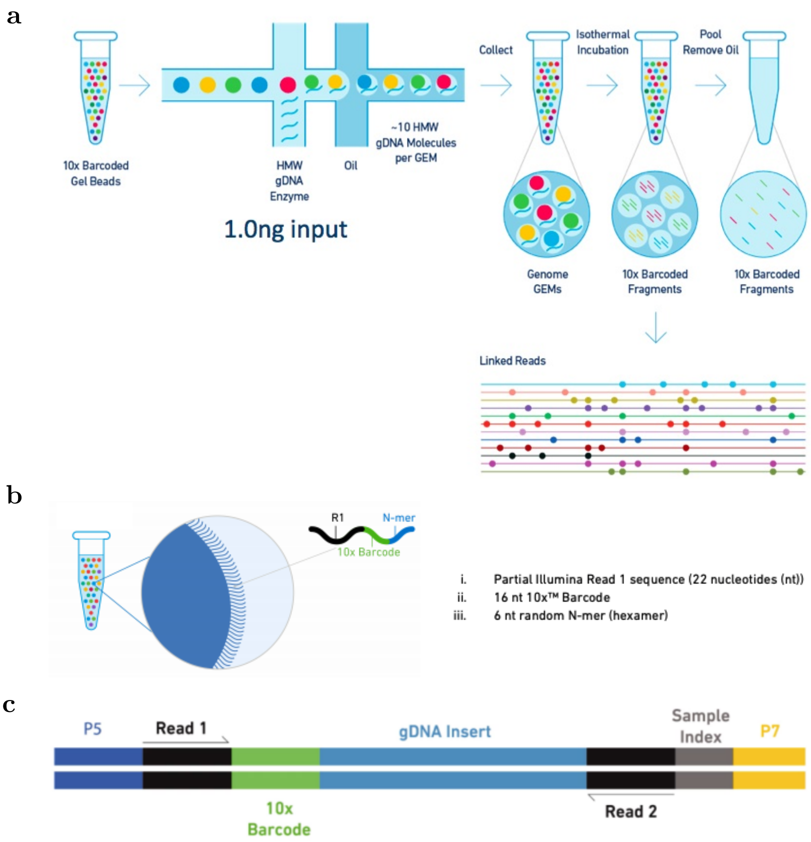
Hi-C crosslinks cells' DNA with formaldehyde, breaks the DNA with a restriction enzyme, and blunt end ligation is used in conditions which prefer joining cross-linked DNA[64][201][271][74]. This produces read pairs which were spatially close in the nucleus but may be far apart in the genome. Because of the 3D structure of the tightly wrapped genome in the nucleus, this means that most links are intra-chromosomal and thus is useful in assembly scaffolding[104]. Hi-C data is used extensively in chapter 4 for haplotype phasing and phasing aware scaffolding.

1.4.1.4 Linked reads

The 10x platform for linked reads uses the same microfluidic system as in scRNAseq. Instead of delivering cells to the droplets, the linked read system delivers HMW DNA from which short reads are created. Whereas the barcode oligo acted as a cellular barcode in scRNAseq, it acts as a long molecule barcode in the linked read system. It starts with high molecular weight DNA input into a microfluidic system that partitions those long DNA molecules into GEMs (Gel bead in EMulsion) with oil surrounding an aqueous solution containing the DNA and reagents with a gel bead housing millions of copies of an oligo containing random primers, Illumina adapters, and the same barcode DNA sequence. Each different gel bead has a different barcode DNA sequence with high probability. Each GEM is Poisson loaded with HMW DNA and on average gets roughly ten long molecules in the standard workflow. Short sequences are then amplified from these long molecules with random priming, creating a construct with the Illumina P5 and P7 adapters, the barcode oligo, and the DNA insert. This is then sequenced using standard short read Illumina sequencing. All of the reads with the same barcode sequence come from the same GEM and thus from a handful of long molecules. When the reads are mapped to a reference genome, the reads from each barcode cluster into a few small regions of the genome associated with their molecule of origin. This long range information can then be used to map into repeat regions of the genome, phase haplotypes, and call structural variation[354][3].

[3] From 2014 to 2017 I worked at 10x Genomics and was the main developer on Lariat, the software used to confidently map into repeat regions of the genome, and had a role in the phasing algorithm including its ability to correct genotypes as well as on the development of the technology as a whole. I am grateful to have been a part of such a talented group of people and to have had the opportunity to learn from them.

Fig. 1.5: 10x Genomics linked reads



a) outlines the microfluidic system to create the Gel bead in reverse emulsion. **b)** shows the gel bead oligo setup and **c)** diagrams the final construct. (image credits to 10xgenomics website)

While 10x Genomics Linked reads are used in Chapter 4 for phasing and phased assembly, the technology is no longer offered by the company. More recently, bead based systems have been developed that do not require a microfluidic system. In solution with microbeads, long DNA molecules tend to wrap around a single bead[353][224]. Separately, Tn5 transposase has been used to insert adapter and barcode sequences at high frequency into genomic DNA[11]. With these ideas combined, Frank Chan's group has developed a technique called Haplotagging that uses microbeads bound to Tn5 transposase with one of 85 million molecular barcodes and Illumina sequencing adapters creating linked read libraries for a fraction of the price in a single tube[228].

1.4.2 Third generation sequencing: long reads

1.4.2.1 PacBio

Pacific Biosciences (PacBio) uses microscopic wells known as zero-mode waveguides (ZMWs) along with single molecules of DNA and DNA polymerase, and optically measures fluorescent tagged nucleotides as they are incorporated by the polymerase. This is known as single molecule real-time sequencing (SMRT sequencing). The DNA template is ligated with hairpin adapter sequences known as the SMRTbell adapters to create a topologically circular template. This allows for multiple passes of the same DNA molecule. Initially, polymerase nucleotide incorporation and optical measurement speed were limited. That combined with the rate at which DNA polymerase dissociates from the molecule limited the number of times long molecules could be sequenced to once or just a few times. This results in long, but noisy reads with roughly 15% error rate[85][42][49] known as continuous long reads (CLR). The PacBio data used in Chapter 3 is CLR data. More recently, advances in the speed of polymerase nucleotide incorporation and optical measurements have allowed for many passes of the same long molecules. This allows for circular consensus sequence calling across these multiple passes creating High Fidelity (HiFi) reads with much higher accuracy (<<1% error rate on average) while maintaining true single molecule sequencing[337]. Over the past few years, PacBio data—both CLR and CCS—has revolutionized genome assembly and is now used routinely for generating high quality reference genome assemblies[4]. In chapter 4, I use PacBio HiFi data for phased assembly.

Fig. 1.6: Circular consensus sequencing

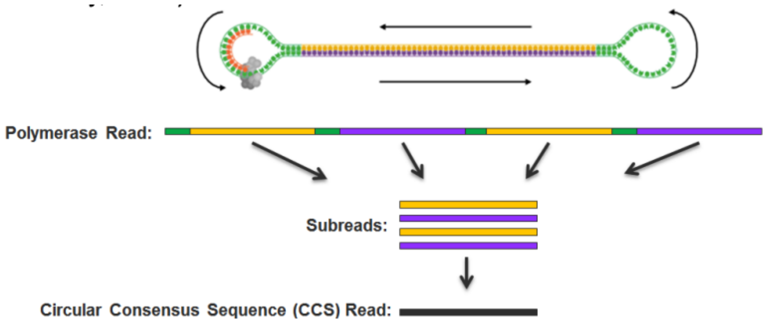


Diagram outlining circular consensus sequencing (image credit: PacBio website[332]).

1.4.2.2 Nanopore sequencing

The idea of reading single molecules of DNA by the current changes as a molecule translocates through a protein nanopore in a bilipid membrane by electrophoresis goes back to the 1980s[156] but took twenty-five years of work before Oxford Nanopore brought the technology to market[61][268]. Because more than one base is inside the nanopore at each timepoint and the past current fluctuations affect the current signal, complex models must be used to interpret these data[60][28]. With these models and improvements in the protein nanopore, sequence accuracy has been reported as high as 92%, but is sequence context dependent. Read lengths for this technology are limited to the input DNA size and reads have been reported as long as 2.3Mb[10][148]. ONT data is not used in this thesis, but is an important aspect of the third generation of DNA sequencing and has been used successfully in the telomere to telomere project[249].

[4] While I would take no credit for HiFi technology, it is poorly known the contribution that my good friend Brendan Galvin had on it. I don't know the full history, but when Brendan was working for PacBio in late 2017, he called me and offered three options for possible improvements to long read sequencing. Among these was longer read CCS data—he estimated the possibility of 10kb CCS reads. He asked me among his three options, which one would have the biggest impact on genome assembly. I responded that 10kb accurate reads would revolutionize the world of assembly. Within a year and a half of that conversation, this became a reality. Of course this was not entirely Brendan's doing either, but between his technical contributions and internal advocation for the technology, I am convinced he played a large role. This is just one of several major contributions to the field he has had.

1.4.3 Reference Genomes

Despite their limitations, reference genomes enable a host of downstream analyses. They provide a common coordinate system by which to say certain genetic variants in one genome are the “same” or different from one another[289]. This allows one to compare multiple genomes and associate certain genetic variants with phenotypes.

1.4.3.1 Resequencing

Reference genomes also allow much more inference to be made from the cheap and high throughput next generation sequencing technology. Instead of generating the entire sequence of each new individual *de novo*, one can create short or long reads and map them onto a reference genome with the assumption that the reference is similar enough to the genome of interest that the mapping is globally correct. Aligning a small sequence with a very large sequence would be computationally expensive with traditional alignment algorithms[301][246]. Many algorithms and data structures have focused on the ability to quickly find all locations in one large reference or database or sequences a new query sequence will match well such as the suffix tree and suffix array, FM-index, Burrows Wheeler index, and minimizers. A full review of these is out of the scope of this thesis[333][213][91][92][198][159]. One of the more recent of these is relevant to this thesis which is the minimizer. In a sliding window of sequence of size W, the minimum (lexicographically or by hash value) sequence of length K is stored for each window[274]. Because adjacent windows often have the same minimum kmer, these can be stored efficiently. And it guarantees a representative kmer at least every W-K bases. These have been used in genome assembly, read mapping, among others[25][197][147]. In a somewhat related idea, if one wants a subset of kmers and does not require the locality guarantee, one can take the kmer hash or two-bit encoding value modulo some number and use only kmers where the resulting value meets some criteria (eg. kmer-hash % 2 == 0 retains half of the kmers)[82]. These are used in chapter 4 in phased assembly and scaffolding to sparsely sample homozygous kmers to cover areas of the genome with large homozygous stretches.

Once one has determined the location to map a read to in the genome, a full Smith-Waterman alignment of the sequence to that region of the genome is done[301]. One can then inspect the differences between the genome of interest and the reference genome to call genetic variants. When one parental chromosome contains one allele and the other parental chromosome contains a different allele, this site is said to be heterozygous and results in roughly half of the reads that are sampled containing each allele.[99][71].

1.4.4 Haplotype phasing

The genetic variants for these individual(s) are generally called in a localized fashion. Some haplotype information may be used, but due to the read lengths of next generation sequencing, they usually are fairly isolated from one another. Determining which alleles occur on the same chromosome and which occur on the alternate parental chromosome was historically termed haplotype assembly but today is called haplotype phasing. Methods for haplotype phasing generally fall into two categories—population level haplotype estimation and individual genome haplotype phasing.

1.4.4.1 Statistical

Because chromosomal recombination is relatively rare, alleles close to one another in the genome will generally cooccur across individuals in the population. Given the genotypes of many individuals, it is possible to statistically determine the most likely set of haplotypes in the population[65] and use this for genotype inference of other variants not sampled in a sparsely sequenced dataset.

1.4.4.2 Direct / Read based

Haplotype phasing across large regions of the genome requires long range genetic information not present in next generation sequencing. If one considers the graph where heterozygous variants are nodes and reads containing two heterozygous variants create an edge between those nodes, it is only possible to phase variants with respect to one another in connected components in that graph. With contiguous reads such as those generated by PacBio or ONT, the potential sizes of regions that can be phased are limited by the length of homozygous regions. Evolution can naturally create large regions of homozygosity through harsh selective pressures sweeping the haplotype landscape in particular regions[165]. This limits the ability of contiguous or localized technologies to generate chromosome scale haplotype phasing. By this reasoning, longer reads generate longer phase-blocks, linked reads generate longer phase blocks due to their longer molecule length, and chromosome scale data such as Hi-C has the ability to infer chromosome scale haplotypes[178]. Many algorithms have been developed for haplotype phasing for different data types or combinations of data types and employ search methods, dynamic programming optimization, graph theoretical approaches, and more. These include HapCut2[84], Whatshap[254], Longranger[354], and many more. In chapter 4 I use haplotype phasing and the phasing consistency of molecules across multiple heterozygous variants as evidence that those variant loci are physically linked in both haplotypes in the genome.

One can also phase individuals based on pedigree information. With the genotypes from the mother, father, and child (or larger pedigree datasets), one can phase variants in the child in all cases except when all three are heterozygous[34]. One can also combine read based phasing with pedigree information[95].

1.4.5 Creating reference genomes

Physical and technical limitations make reading whole chromosomes a practical impossibility. Therefore, to recover the sequence of the genome often requires sampling reads from the genome randomly until one has with high probability covered the genome more than once in almost all regions[184] in a method known as 'shotgun' sequencing. Using the sequence similarity of multiple reads, one can build a consensus sequence of the underlying genome. This process is known as 'assembly'.

1.4.5.1 The old way

Before next generation or third generation sequencing was available, there were several massive efforts to sequence genomes, but they were limited to human, important crops, and model organisms due to their cost and time[185][54][288]. In these projects, BACs and other vectors were used to clonally amplify large sections of the genome. Shorter sequences from these would then be subcloned and Sanger sequencing was used to produce reads from each end of these, and each BAC would be assembled separately. At the end, all BAC sequences would be assembled together, often with the aid of a physical map. While costly, this method had several benefits over the whole genome shotgun method used in the privately funded human genome effort and over the later next generation shotgun assemblies. The ability to deal with 150-200kb at a time means that many repeats that would make the assembly process difficult are absent as they may occur in a separate BAC clone. The clonal nature of this strategy also means that only one haplotype is sampled in a given BAC and the heterozygosity that may make assembly difficult is also not present. And the Sanger sequencing read lengths also spanned small repeat structures that the later next generation sequencing did not. These projects, while costly, produced high quality reference genomes for the most important organisms to the human species and are still being used today.

1.4.5.2 The dark times

In the age of next-gen sequencing, genome assembly was initially popular. The low cost of sequencing promised personal genomes and the ability for every lab to sequence and assemble the genome of their organism of interest. Despite a massive amount of research going into assembly algorithms, *de novo* genomes produced with short read technology were never even close to as high quality as the ones produced by the BAC+Sanger sequencing methods of the previous generation[287][281]. Over time, interest in genome assembly waned both from the perspective of assembly algorithm development as well from people wanting to assemble new genomes as the results were so fragmented and error prone as to be of limited downstream use[83].

1.4.5.3 A new hope

Long single molecule read technologies have been around and commercialized since 2010, but due to cost, scaling, and the computational difficulties of dealing with reads with high error rates took some time for widespread adoption for genome assembly[50]. But as throughput scaled up, costs came down, and tools for assembling these data improved[52][278][173], long read technology rapidly became the de facto for *de novo* genome assembly. Initially this was usually paired with a next-gen short read dataset used to polish the assembly and/or reads prior to assembly, but eventually partial order alignment based multiple sequence alignment[190] was used for polishing by the noisy long reads themselves[323]. More recently, PacBio's improvements to the DNA polymerase processivity along with their circular consensus technology has allowed for many passes of a long single molecule which can then self polish with these same algorithms. This produces high accuracy long reads, which have even further revolutionized genome assembly[337]. This, combined with Hi-C scaffolding now routinely produces reference quality genomes in full chromosome scaffolds. However, some problems still exist for certain genomes. Until recently, long read technologies required more DNA input than could be extracted from single small organisms. But due to recent advances in library preparation, DNA requirements have come down substantially, and in Chapter 3 I describe the first high quality assembly of a single mosquito. Also, many organisms are highly heterozygous and this creates difficulties in the assembly process. In Chapter 4, I describe methods for turning the problem of heterozygosity into an advantage by using phasing consistency between multiple heterozygous sites as a signal for physical linkage in phased assembly and phased scaffolding.

1.4.6 Assembly algorithms

Outside of some enzymatic reaction modeling I did in high school, the problem of genome assembly was the first problem in computational biology that I was introduced to and arguably one of the most important problems to the field. While the problem naturally lends itself to the pure computer scientists and mathematicians, the peculiarities of how genomes evolve tend to defy most basic assumptions. Not only is genomic sequence not random, but many structures seem designed to make the problem harder. Transposable elements and large segmental duplications, viral insertions and trinucleotide expansions, low GC content and homopolymers, centromeric repeats and telomeric repeats are just a few of the challenges the genome poses to the problem of assembly. At its most basic, the problem is simple. Find similar overlapping sequences through pairwise read alignments. Infer that those sequences most likely originated from the same locus in the genome. Create a larger contiguous sequence representing the underlying genome and repeat. The problems arise when this inference is false. These algorithms have evolved over time and with the data types available, cheapest, or most promising at the time.

1.4.6.1 Overlap, Layout, Consensus

The first assembly algorithms were known as “overlap, layout, consensus” algorithms due to their three primary steps. They first do an all vs all alignment of the reads. While this seems computationally costly, due to exact match hashing of smaller subsequences as a filter, only reads very likely to arise from the same genome locus well will be aligned[136]. These overlaps can then be used to create an ordering, or layout, of these reads. This layout was then used to generate a consensus either through multiple sequence alignment or heuristics[157]. This method was used for many years including on large sequencing efforts such as the Human Genome Project[185]. Implementations of this strategy include PHRAP[1], TIGR[307], GigAssembler[158] (used in the human genome project), and the Celera assembler[240]. The Celera assembler was developed for and used in the privately funded human genome project[325] and many other genome projects of this era. In the modern era, HGAP and Canu are overlap, layout, consensus algorithms build for long noisy reads[50][173] and HiCanu for long accurate reads[248] and have generated many high quality genome assemblies[164].

1.4.6.2 de Bruijn graphs

One of the earliest formalizations of assembly as a graph theoretical problem was in the late 1980s in the context of sequencing by hybridization (SBH)[76]. In sequencing by hybridization, one would expose many copies of single stranded DNA of interest to an array of microwells with different oligonucleotides. The unbound DNA would then be washed away and a reporter system was used to determine which wells the DNA bound to. This indirectly created the later microarray SNP-chip technology. While this technology never proved feasibly scalable for both laboratory and information theoretical reasons[267][5], it did motivate Pavel Pevzner to pose assembly as an Eulerian path on a de Bruijn graph of the sequences. In SBH, the oligos used were short due to the maximum number of microwells possible and the total number of possible oligos of a given length (4^k where k is the length of the oligo). A graph was constructed such that kmers (at the time, these were referred to as I-tuples, but I will use the modern terminology) with overlapping k-1 sequences would have edges between them. A Eulerian path on this graph represents a linear assembly of the sequence. Later, with next-gen short reads, this idea was given new life with Pevzner and Waterman[257] creating an assembly algorithm Euler based on this idea. With reads as opposed to short oligo microarray hits, one uses all subsequences of length k and constructs the same type of graph as with SBH[350][298][144]. The obvious downside of these techniques is the loss of information by breaking the read into shorter sequences. However, this can be mitigated by reconsidering the reads and read pair information to further resolve the graph[37][334]. Although Jue Ruan and Heng Li used a fuzzy de Bruijn graph approach for noisy long reads, these methods are generally only applicable to data types with relatively high accuracy reads as the length of a kmer of any length in PacBio CLR or ONT reads would not be a true kmer with high probability.

[5] From 2011-2013 I worked at Nabsys, a company that was attempting initially to create a positional sequencing by hybridization[118] technology where DNA would be tagged with oligos and translocated by electrophoresis through a nanopore and the oligo locations would be read. If this was done for all of the oligos of a certain length, the positional information would potentially solve the information content problem for longer sequences. This failed horribly due to several limitations. The company continued on to create a digital ordered restriction map technology[108] creating similar data to that produced by bionanogenomics but less successful. While the technology was ultimately a failure, my time there was an incredible learning experience.

1.4.6.3 String graphs

The string graph is a data structure representing the idealized assembly graph and was described by Gene Myers in 2005[242]. It uses the full read lengths and overlaps between reads are collapsed into a single sequence. Thus, if there are repeats longer than the read length, these will be collapsed and unique sequence will create loops between repeats. Jared Simpson and Richard Durbin created a compressed version of this dataset and an assembly algorithm based on it using the FM-index[92][91][297]. This allowed for the use of the full length of the reads without complex and costly read pair threading algorithms on the de Bruijn graph and the compression reduced the memory requirements to the point that mammalian genomes could be assembled on commodity hardware of the time. Falcon[52] is a string graph assembly algorithm written for long noisy reads, and HifiAsm represents a phased string graph built on PacBio HiFi accurate long read data[48] and produces some of the highest quality assemblies today.

1.4.6.4 Repeats, Heterozygosity, and Errors

While tremendous progress has been made in genome assembly through improvements in both the data and algorithms, problems still exist. In the process of creating overlaps, one will encounter inexact homology due to either inexact repeats, heterozygosity, or sequencing errors. In the graph methods that only collapse exact matching sequence, these inexact homologous sequences arising from these create complex graph structures that either need to be resolved or the final sequence assembly will be fragmented[43]. Many organisms are much more heterozygous than humans, who went through a population bottleneck in recent evolutionary history[120]. While much of the exact repeat problem has been solved due to long accurate reads spanning such repeats, inexact but highly homologous repeats exist on the scale of megabases[66]. In haploid assembly, any inexact homology is either due to errors or repeats. In diploid and polyploid assembly, inexact homology can come from paralogous sequences, heterozygosity, or sequencing errors. Incorrectly inferring one as another can create misassemblies or retained haplotypes assembled separately which are generally intended to be collapsed in a reference genome for the downstream application of resequencing. Several methods have been created to combat these problems.

1.4.6.5 Trio assembly and trio binning

One way to reduce the problem of heterozygosity versus repeats is to add haplotype phasing information to the process. In section 1.4.4, I discussed haplotype phasing via pedigree genotypes. With this information, one can more easily distinguish between heterozygosity and paralogous sequences. In the age of next-gen sequencing, Malinsky, Simpson, and Durbin, created Trio-SGA, an algorithm utilizing parental and child information in the string graph based algorithm to deliver higher quality assemblies[212]. More recently, with the reduced costs of long read sequencing technologies, instead of embedding the knowledge of the pedigree information into an assembly algorithm, trio-binning[174] separates the long reads by haplotype prior to haploid assembly of each haplotype. Because long reads generally span multiple heterozygous variant sites, trio-binning uses the kmer difference between the paternal dataset and maternal dataset to categorize reads as maternal, paternal, or uncategorized. Each bin of haplotype reads, along with the uncategorized reads, are then assembled independently producing highly contiguous and accurate genome assemblies. However, this requires pedigree data which is not feasible for many species in a large project such as the Darwin tree of life and the Earth biogenome project.

1.4.6.6 Haploid assembly: Hytaditiform moles, seeds

In cases where it is possible to assemble haploid data, it is clearly advantageous. The telomere to telomere (T2T) consortium has used multiple technologies to sequence a human cell line derived from the haploid complete hytaditiform mole (CHM) 13[249]. While this does not represent a viable human genome, it is likely the most complete and accurate sequence of a human genome to date. In some other areas, tissues are clonally haploid with enough material to create a sequencing library. In many conifer species, the seeds within pine cones are haploid and can be used as source material for genome assembly where other material may be polyploid[39].

1.4.6.7 Phased assembly

Another option for combatting the heterozygosity vs paralogous sequence problem is building haplotype phasing into the assembly algorithm and explicitly assembling both haplotypes. Many algorithms have attempted this in the past including Falcon[52] and trio-sga[212], but until recent data improvements, this has proven difficult. One approach used in DipAsm, was to create an initial assembly, haplotype phase that assembly, and then use that phasing to split haplotypes prior to haploid assembly—a process akin to trio-binning but without the pedigree information[97]. Another method is to create a de Bruijn graph or string graph via short reads and align long reads onto that graph to both phase the graph and assemble both haplotypes[96]. And yet another approach employed in HiFiAsm is to create a phased string graph directly from the long accurate HiFi reads[48]. In chapter 4, I present a method using haplotype phasing consistency to create phased assemblies and phased scaffolds.

1.4.7 Post assembly manipulations

1.4.7.1 Polishing

Especially when assembling with long noisy reads, and sometimes with other technologies, a post assembly step of polishing can improve base quality. With long noisy reads such as PacBio CLR or ONT, one can polish with the reads themselves[51][204] or one can use a short read dataset to polish the final assembly[329].

1.4.7.2 Haplotig purging

Because one of the primary downstream applications of reference genomes is resequencing, and it is undesirable for the two haplotypes to compete for read mapping, we generally want to produce an assembly with only one haplotype. If the haplotypes are different enough, assemblers may assemble portions of haplotypes as separate contigs rather than collapsing them. Several tools have been made to remove these alternate haplotype contigs (haplotigs) or "haplotypic" sequence using both sequence similarity as well as coverage of reads mapped to the pre-purged contigs[135][273][112].

1.4.7.3 Scaffolding

In most cases, contigs resulting from the assembly process are not chromosome length, and if further information is available, we would like to order and orient them with or without gaps in their chromosomal context. Paired end reads, long reads, linked reads, Optical maps, and Hi-C have been used for this purpose over the years[272][121][104]. The longer range the data is, generally the better it is for scaffolding to span whatever gaps may exist between contigs. So in the modern era, Hi-C is the preferred data type for scaffolding. In chapter 4, I present a method for phasing and phasing aware assembly scaffolding.

1.4.7.4 Gap filling

Gaps may exist either due to contigs being scaffolded together with a gap between them or through the assembly process itself. These can sometimes be filled by aligning reads across the ends of each contig on either side and creating a consensus sequence for the gap[86]. More recently, ultralong ONT reads have been used to fill gaps in projects like the T2T project[149][148].

1.4.8 Assembly validation and curation

While the modern assembly process is much more automated than it used to be, some validation and curation is still necessary for the highest quality genomes. Today, curators use semi-automated tools to assess haplotig retention, contamination, find and bread misassemblies and order and orientation errors in scaffolding[133]. Assembly completeness and haplotig retention is assessed by orothology of genes to known sets of genes using Benchmarking Universal Single-Copy Orthologs (BUSCO)[330]. Kmer methods such as KAT plots[214] can be used to assess heterozygosity, error rate, and haplotig retention. Hi-C data is visualized on a heatmap and used to correct scaffolding errors, create new scaffolding joins, and find potential misassemblies[80][160][79]. Coverage of reads aligned to the assembly along with G/C content and database searches of sequences are used to find contamination of other organisms in your sample and assembly[183][44].

In the following chapters I will present several methods for using genetic variation to demultiplex single cell RNAseq mixtures and improve genome assembly and scaffolding. I will generally use the word “I” when I have done the work alone and “we” when it was done in collaboration with others.

Chapter 2

Clustering single cell RNAseq by genotypes in mixed samples.

2.1 Background

Cells are the 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[310]. Single cell RNAseq (scRNAseq) is a high precision instrument measuring the transcriptional profile of each cell individually usually by physically separating cells into and delivering distinct barcode sequences to templates generated from the mRNA of each cell[259]. Further advances in nanodroplet and nanowell technologies have made it possible to apply scRNAseq to thousands of cells simultaneously[211][355][106] instead of the plate based strategies that usually were limited to hundreds of cells at a time[259].

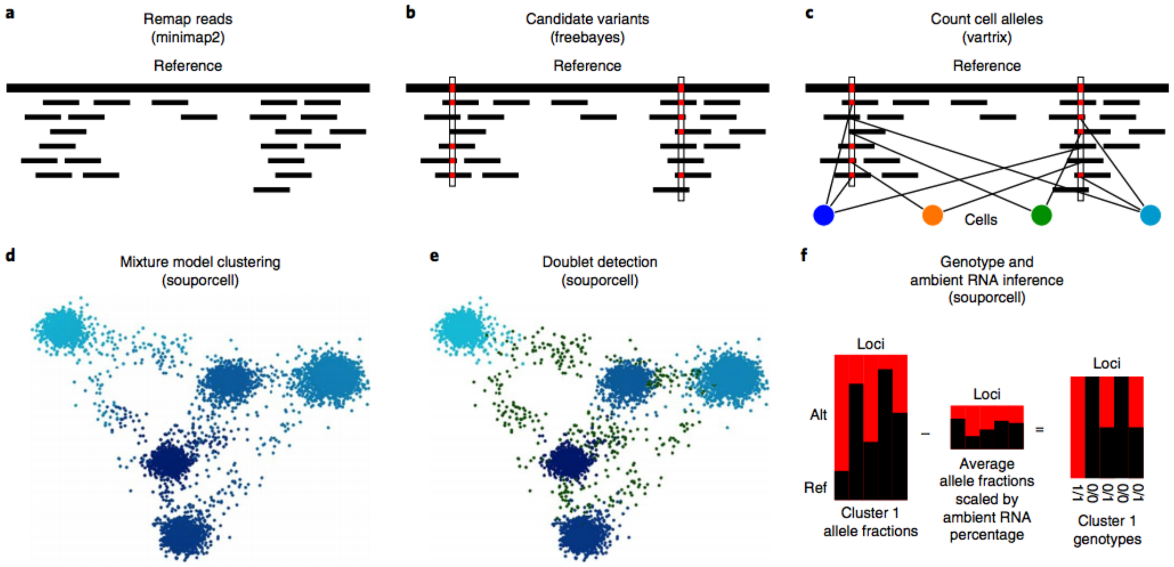
Some samples contain cells of mixed genotypes including those of single celled organisms such as malaria infections, the gut microbiome, and environmental samples as well as intrinsically mixed samples such as maternal/fetal, transplant patient, or tumor samples. Additionally, mixing cells from multiple individuals into a single experiment has become a popular experimental design because it makes the data more comparable between individuals, reduces costs, and can improve doublet detection. In order to properly analyze them, one must first assign each cell to its genotype of origin. Some tools and methods exist for this purpose[154][304][344] but each of them has some downside. Demuxlet requires prior knowledge of the genotypes in the mixture. Vireo and scSplit use clustering initialization and optimization strategies that fall apart when individuals in the mixture are related. And none of these model the ambient RNA in the system, leading to multiple errors including over calling doublet cell barcodes. Ambient RNA in single-cell RNAseq (colloquially 'soup') is a phenomenon in which RNA molecules from cells that have lysed before cell partitioning are included in partitions with cells from which they did not originate[348]. This adds noise to both the transcriptional profiles of the scRNAseq experiment, but also the genotype analysis and demultiplexing of mixed genotype samples. Cell hashing and lipid tagging require additional experimental steps and are not applicable to innate mixtures because the samples must be separate at the time of tagging.

In this chapter, I present souporcell, a tool containing a collection of algorithms to support mixed genotype scRNAseq experiments[123]. To use the genetic variants measured in the RNAseq reads to assign cells to their donor of origin, first I must describe a strategy for calling reliable variants in scRNAseq data (figure 2.1a,b) and assigning allele counts to cell barcodes (see figure 2.1c). I then describe the core algorithm of souporcell, that of clustering cells by their genetic variants in the face of the sparse measurements of expressed alleles by each cell (figure 2.1d). Next I describe the algorithm used for determining which barcodes represent multiple cells instead of a single cell (figure 2.1e). And finally I describe a statistical model and coinference of the cluster genotypes and amount of ambient RNA there is in the experiment (figure 2.1f).

I then demonstrate and benchmark souporcell against a dataset which was mixed *in silico* and thus we retain full knowledge of the ground truth of which barcodes came from which individual, which barcodes represent cross-genotype doublets, and how much ambient RNA was simulated(see figure 2.6). To show that this dataset is a realistic example, I then show the same results on an experimental mixture of cells from those same individuals(shown in figure 2.7).

I compare this method to demuxlet, the previous gold standard method that requires genotype information *a priori*, as well as two new tools that, like souporcell, do not require prior genetic information[138][344] on the *in silico* mixture with various parameters of the data (doublet rate, ambient RNA amount, minority cluster size) swept across a range of values and evaluate clustering, doublet detection, and ambient RNA detection across a wide range of data charactaristics. We sought to test souporcell on more challenging cases and so chose tissue from highly related individuals (maternal-fetal placental and decidual experiments). We also test on mixtures of the malaria parasite, *Plasmodium falciparum*, which is challenging due to the cells having much lower expression levels

Fig. 2.1: souporcell overview



a) First, the reads are remapped using minimap2, retaining the cell and UMI barcode for downstream use. **b),c)** Then candidate variants are called using freebayes **b** and count the allele support for each cell using vartrix **c**. **d)** Using the cell allele support counts, we cluster the cells with sparse mixture model clustering(Methods). **e,f)** Given the cluster allele counts, we categorize cells as doublets (**e**) or singletons and, excluding doublets, the amount of ambient RNA is inferred along with the cluster genotypes(**f**; see example for one cluster). Alt, alternate allele; ref, reference allele.

than the human cells previously tested and without upfront knowledge of the number of genotypes present in the sample.

I show that souporcell not only outperforms the competing methods, but also surpasses the previous gold standard, demuxlet, on both cell assignment and doublet accuracy. Furthermore, souporcell explicitly models and estimates the amount of ambient RNA in the experiment, which is a major confounder of scRNAseq analysis with regard to both expression and genotype. Souporcell is freely available under the MIT open source license at https://github.com/wheaton5/souporcell.

2.2 Methods

2.2.1 Variant calling on scRNAseq data

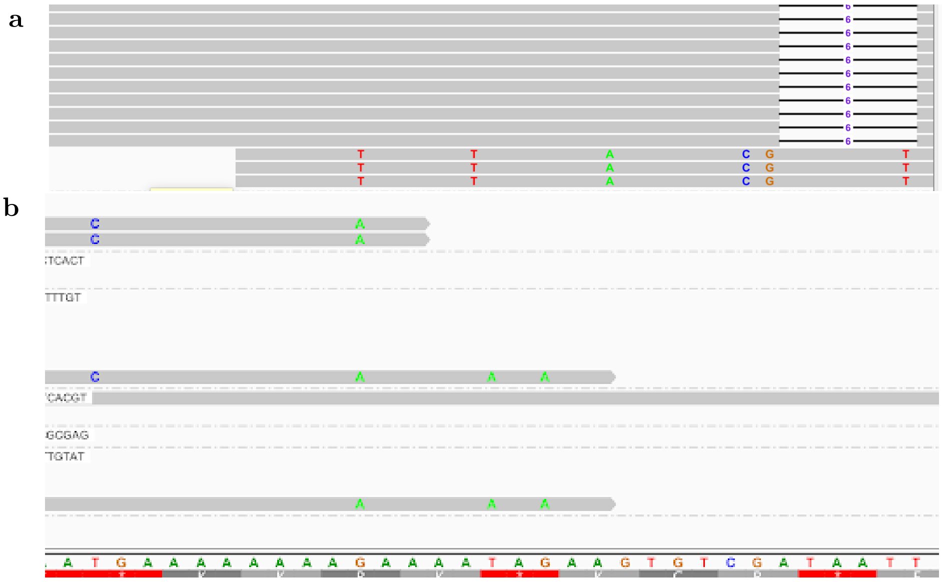
In order to use the genetic variants in the scRNAseq reads to assign cells to individuals, one must first call the variants accurately and assign which alleles were expressed by which cells. Little work has been done on identifying genetic variants in bulk RNAseq[260][346] let alone scRNAseq[94][137].

2.2.1.1 Remapping

Currently, the most popular software for the initial analysis of droplet based scRNAseq data to generate the mRNA expression matrix is cellranger[355]. In the cellranger pipeline, the reads are first mapped to a reference genome. RNA mapping and DNA mapping software differ due to the intended downstream uses. With DNA mapping, accurate variant calling is one of the primary applications and thus much work has been put into providing accurate mapping quality scores and base level alignments making single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) easy to call accurately versus the reference genome used[198][197][187][99][71]. The mapping and alignment operations are relatively computationally costly if accurate variant calling is not one of the desired downstream uses. RNA mapping applications are usually primarily concerned with just counts of reads per gene and sometimes differential RNA splicing. Thus, the RNA community has developed software packages, some of which are faster but do not provide base level alignments[30][253] and others that allow for gapped alignments caused by the introns being spliced out[188][75][163][316]. But in optimizing for the downstream applications of transcript counting and differential splicing, genetic variant calling accuracy can suffer. In cellranger, the mapping component is done with the STAR aligner[75] 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. In addition to this, base level alignments toward the ends of reads can be error prone because there is not a sufficient amount of sequence remaining for the "correct" alignment to be the optimal alignment according to the alignment score. For example, an alignment may prefer to take several single base mismatch penalties rather than a single true indel penalty. This tradeoff is more likely to happen towards the end of the read when there are fewer bases remaining to align (see figure 2.2b). Because of this, mappers built for variant calling such as BWA[198] and minimap2[197] have a relatively small one-time penalty for soft clipping any number of bases from the end of an alignment[195]. 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 (see figure 2.2b). 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, it will make the same error for reads spanning the indel but not make the error for reads not spanning the indel causing those mismatches to potentially appear to be heterozygous genetic variants.

Fig. 2.2: Star alignments' indel and soft clipping problems


a
b
TCACT
TTTGT
CACGT
GCGAG
TGTAT
A T G A A A A A A A A A G A A A A T A G A A G T G T C G A T A A T T

a) Reads that span more bases find a six base indel (upper right), but ones that span fewer bases incur many erroneous single base mismatches without soft clipping. Minimap2 finds this indel in all of the shown reads and soft clips some others with even later start positions. **b)** In the second example, the reads have an adenosine homopolymer and partially match this section of the reference. Minimap2 softclips these reads.

The indel 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 that 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 as low as an eight base segment that should, with high probability, occur by random chance alone. Due to the nature of mapping qualities being assigned to the whole alignment and not each segment of the alignment, these sections are often denoted as having a high mapping quality when in fact the matches should occur by chance. If there is actually an alternative allele in one of these regions to which some reads have spurious matches, those reads, and thus those cells, appear to support the reference allele. These alignments provide one further technical issue, which is that they dramatically slow down the pileup and fetch commands in samtools[199] that are necessary for variant calling.

In order to show that the variant calls from STAR are error prone, we compared the variants obtained with STAR and minimap2 bams versus a single cell experiment done on cells from the Genome in a Bottle (GIAB) consortium cell line NA12878. In the GIAB high confidence regions with a depth of at least 10 and a quality score of over 30, 44,036 variants were called from the STAR bam of which only 15,544 were in the GIAB ground truth data leaving 28,492 likely false positives. In comparison, 25,743 variants were called from the minimap2 aligned bam of which 12,055 were in the GIAB ground truth and 13,688 false positives. Remapping with minimap reduces sensitivity some, but drastically reduces false positives.

For these reasons first the reads are remapped with either BWA, minimap2, or hisat2. I have found good 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 thesis 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. Then, PCR duplicates are removed by identifying reads with the same unique molecular identifier (UMI) barcode, cell barcode, and have the same start and stop position.

2.2.1.2 Variant Candidate Calling

Once the reads are accurately mapped and aligned, one must then proceed to variant calling. I assessed two strategies for calling variants on scRNAseq and assigning alleles to cell barcodes. I first treated the sample as a population of cells and called variants with freebayes in population variant calling mode[99][71][199]. Freebayes is a variant calling tool which looks at the differences between the reads and the reference genome to call genetic variants. In its standard mode, it makes a diploid assumption and attempts to build haplotypes to give more power. In population variant calling mode, each read is labeled with a read group denoting an individual of origin. With this approach I assigned each cell barcode to its own read group in the input bam and the variant caller produces a population VCF with genotype calls for each cell for each locus. I also assessed treating the sample as an unknown mixture of haplotypes. In this mode in freebayes, reads may come from any individual, and variants are called without ploidy or labeled individual assumptions. The output is simply the variants without labeling which cells have which alleles. I then decide whether each cell has which alleles using the tool vartrix[94]. Our analysis suggests these two strategies perform very similarly. 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 which will return a sparse matrix market format indicating how many of the reference allele or alternative allele each cell barcode expressed for each variant locus.

2.2.1.3 Cell allele assignment

Vartrix works by aligning each read to the reference sequence as well as the variant sequence to determine which one it supports. Doing this rather than simply inspecting the base level alignment improves reference bias and alignment end effects. For example, when assessing if a read supports an insertion of an A in a homopolymer of adenosines and the read does not extend past that homopolymer, the read will align without the insertion even if it came from the haplotype with the insertion. Aligning to both underlying sequences will produce the same alignment score and it is ambiguous which allele the read represents.

2.2.1.4 Validation: Genome in a Bottle

We validated our variant calling accuracy by obtaining from 10x Genomics an scRNAseq dataset run on cells from the Genome in a Bottle consortium individual NA12878 cell line for which there are high quality ground truth variant calls available[358]. As the scRNAseq data will only cover a subset of genes and because this dataset was relatively low coverage, I will primarily focus on false positive rate and not sensitivity. Much of this analysis was done by Yichen Wang as part of a rotation project in our lab. Initially we found a false positive rate of 35.6%, dramatically higher than the 1-2% you would have with reasonable coverage whole genome DNA sequencing.

2.2.1.5 RNA editing

We sought to determine the causes of the remaining false positives so identified the false positive SNPs called on the NA12878 scRNAseq data in high confidence regions as defined by the GiaB resource and found that most (80.8%) of them are purine-to-purine or pyrimidine-to-pyrimidine transitions when we considered the reference and observations. A-to-G and T-to-C transitions happened in much higher frequency than the remaining, making up 59.5% of total false positive sites (see table 2.1). Calling variants from bulk RNA sequencing data also displayed a similar pattern, but using whole exome sequencing data did not, linking the excessive purine-to-purine and pyrimidine-to-pyrimidine transition specifically to RNA seq. We hypothesized that this could be due to RNA editing. The most common RNA editing event is the deamination of adenosine to inosine on pre-mRNA[357]. Inosine is then read as guanosine by reverse transcriptase, resulting in a T-to-C event in the cDNA, which can explain A-to-G and T-to-C SNPs in variant calling. Visualization in the Integrative Genomics Viewer[275] validated the existence both reads with the false positive allele and reference allele in SNP loci. Moreover, we observed that the reads that had the same UMI contained the same allele, but not the reads that had the same cell barcode. This further supported the hypothesis of RNA editing, because the the reverse transcriptase reading inosine as guanine would be consistent for PCR duplicates of the cDNA, but not necessary for all reads in one cell. And if these were due to sequencing errors, they would not be consistent across all PCR duplicates. To test the hypothesis of RNA editing, we found an RNA editing database (REDIportal[258]: http://srv00.recas.ba.infn.it/atlas/) and removed the known A-to-I editing sites in our vcf files. Filtering out RNA editing sites considerably reduced the amount of false positive variants (from 2884 to 1937) and kept most true positive variants (from 8093 to 8073), leading to a reduction in false positive rate from 35.6% to 24.0%. We also discovered that remapping with hisat2 could further reduce false positive rate and improve sensitivity (9540 true positive loci, 1065 false positive loci, false positive rate 11.1%). This is due to hisat2 using similar alignment penalty parameters and soft clipping thresholds to the DNA aligners made for variant calling while also being splice aware. However, this work was done after the souporcell paper was published, so the results in this thesis are done with the minimap2 alignments as previously stated.

Despite these results, the extent of RNA editing has been in debate in the literature for some time[200][45]. Many potential causes for these discrepancies have been proposed such as alignment edge effects and systematic sequencing artifacts. I assessed these alignments for alignment edge effects and find that this is not a significant source of these base differences. In our data, variants called from the STAR aligned bam have significantly more of these than variants called by the minimap2 aligned bam. When inspecting these, we find that minimap2 correctly gives these reads low mapping qualities because there is a competing alternate mapping location. When assessing these found in the minimap2 aligned bam, there is no alternative mapping location according to minimap2, blat, or BWA. It could be the case that there is an alternative mapping either not found by any of these tools or the alternative sequence is simply absent from the reference genome. Another potential is that these come from errors in reverse transcription. Potapov et al. used PacBio sequencing to compare errors in first strand vs second strand synthesis to assess errors in transcription versus reverse transcription and found RT errors also have a strong A->G bias[266]. Further research may be done on this via Oxford nanopore direct RNA sequencing vs cDNA sequencing. The fact that these are largely filtered when using the RNA editing database may simply be because false RNA editing sites exist in the database. In any case, whether these are RNA editing sites, mapping errors, RT errors, or from some other etiology, they are likely not germline variants and our system will benefit from filtering them out.

Table 2.1: RNA editing as a source of false positive variant calls

a

obs ref	A	T	C	G
A	0	54	69	867
T	55	0	849	80
C	77	289	0	75
G	326	68	70	0

b

obs ref	A	T	C	G
A	0	54	69	395
T	55	0	376	80
C	77	289	0	75
G	326	68	70	0

a) False positives are primarily purine to purine and pyrimidine to pyrimidine with a notable increase in A->G and T->C caused by the RNA editing adenosine to inosine. The inosine base is then read as a guanine by the reverse transcriptase. **b)** shows the false positive profile after filtering known RNA editing sites.

2.2.2 Sparse mixture model clustering

In order to introduce this method, I 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. In table 2.2 I show some basic statistics about two datasets — one with a mixture of six strains of the malaria parasite *Plasmodium falciparum* 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 a mixture of five human individuals from the human induced pluripotent stem cell project. A filter is used 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 dramatically 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 2.2: Single cell data statistics

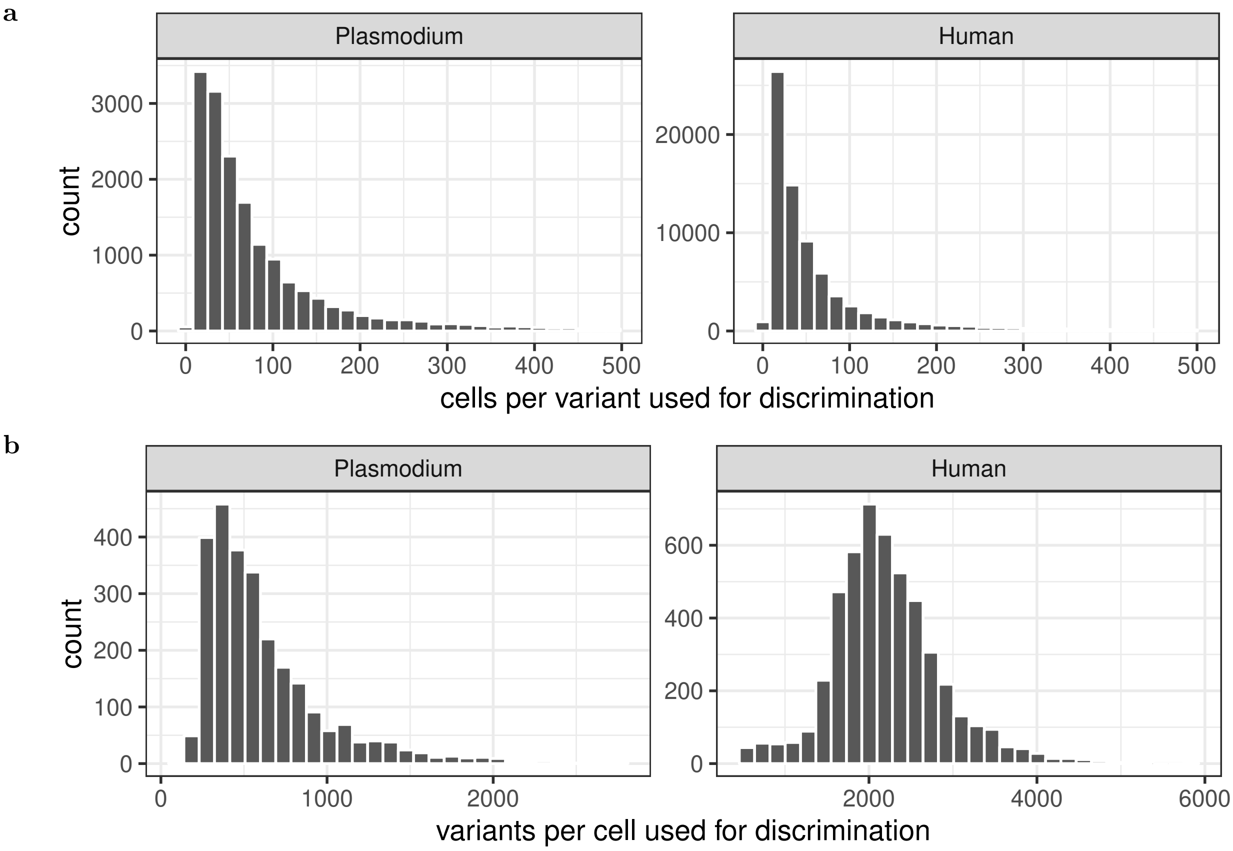
	malaria	human replicate 1
number of cells	2608	4925
median UMI per cell	995	25155
total variants	39487	194079
median cells per variant	24	18
median variants per cell	667	2642
total discriminating variants	16783	77878
median discriminating variant per cell	512	2147
median cells per discriminating variants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

To describe the souporcell clustering algorithm, I will start by making some definitions.

Definitions:

- K: number of genotype clusters to be fixed at the outset. Lower case k will be used for indexing and referring to a specific cluster.

Fig. 2.3: Single cell sparsity



a) Shows the distribution of number of cells that have each variant and **b)** shows the distribution of the number of variants expressed by each cell. Both of these are subset to only consider variants that are used for discrimination. A variant is used for discrimination if it has at least four cells expressing the reference allele and four cells expressing the alternative allele.

- C: number of cells. Lower case c will be used for indexing and referring to a specific cell barcode. This barcode could have 0, 1, or more 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 locus. Only biallelic variants are used. 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 representing the number of alt alleles seen at locus l in cell c.

- $\phi_{k,l}$: cluster center value representing 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. The expected values should 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.
- T: temperature parameter for deterministic annealing process which is described later.

2.2.2.1 Model

A maximum likelihood strategy is used by maximizing $\mathcal{L}(data)$ under a given model.

$$\operatorname*{argmax}_{\phi} \mathcal{L}(data, \phi) \tag{2.1}$$

The likelihood of the data, treating cells independently and marginalizing each cell across the clusters it could belong to, is defined in equation 2.2. At each locus the alternate allele count is modeled by a binomial with n as the reference + alternative allele counts for that cell at that locus and $\phi_{k,l}$ as the cluster center value representing the allele fraction for cluster k at locus l. Each locus is assumed to be a germline genetic variant, but these may also represent false positive variant calls or somatic mutations. In practice, this rarely matters as false positives should be independent between loci across cells thus giving no clustering signal. And somatic mutation would need to represent more genetic variation from the germline than the genetic differences between individuals in order to cause incorrect clusterings.

Cluster model Likelihood function

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

This model deals with sparsity naturally because if a cell has zero reference alleles and zero alternate alleles, a binomial with any probability, zero observations, and zero trials has a probability of one. Instead of uselessly multiplying many ones together, sites for which a cell has no alleles can be ignored. One could then maximize this likelihood with random initialization of cluster centers $\phi \in (0, 1)$ followed by expectation maximization (EM), but there are some problems one can run into.

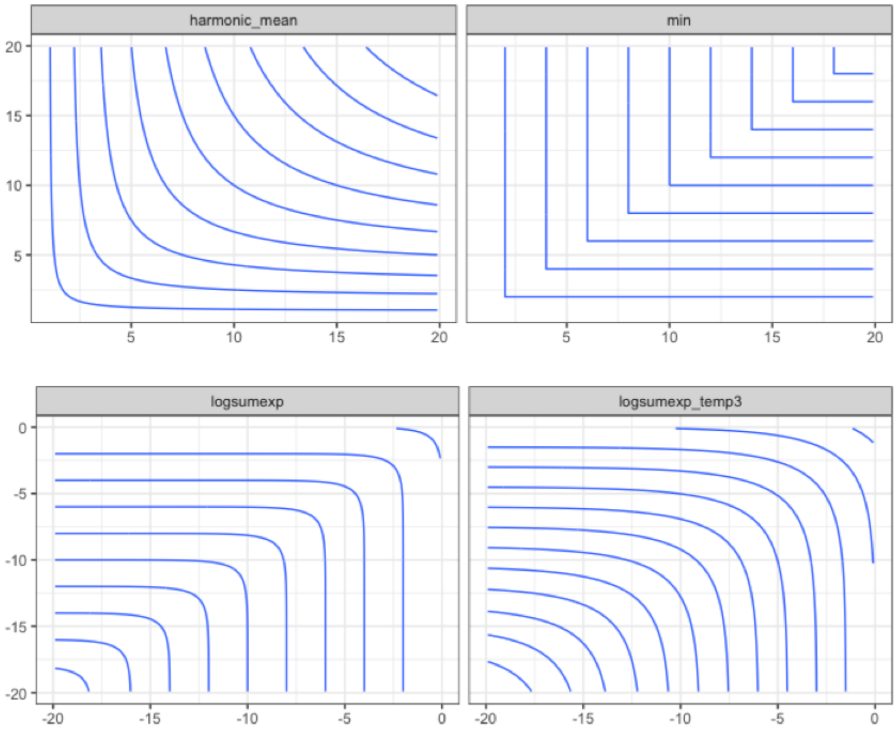
2.2.3 Deterministic Annealing

This method, as is the case with many clustering algorithms, may suffer from local optima in instances of poor initialization of the cluster centers ϕ. This problem increases dramatically as the number of clusters increase. A standard solution to this problem is to have multiple restarts with random cluster center initializations, but as the number of clusters grows, the number of random restarts necessary to obtain the optimal clustering with high probability is unsustainable[324][16]. In addition to this, the sparse nature of the data increases the potential for local optima in the EM process. The two local optima clustering may produce is when one individual is split across multiple clusters and when multiple individuals are assigned to the same cluster. Some cells will express certain loci and other cells will express other loci. This means that given a random initialization of cluster centers, some cells from individual 1 may initially be more similar to one cluster center and other cells from the same individual may be more similar to a different cluster center because they happened to express different loci and the random initialization happened to fall a particular way. This may lead two cluster centers to be optimized for one individual.

Another approach to overcoming local optima in clustering is to initialize the cluster centers intelligently. The most simple initialization strategy is to initialize cluster centers to the values of individual data points (a particular cell in this case) as opposed to randomly in the space, but due to the sparse nature of the data, this would only assign a small minority of the dimensions, the rest of which would need to be random. Other smart cluster center initializations such as kmeans++ are also not particularly applicable to sparse datasets[15]. Not having access to these cluster center initialization strategies is limiting and makes it more likely that at initialization, multiple individuals' cells will match a single cluster.

Without intelligent cluster center initialization as a straight forward option, I turned to methods which are better able to find their way out of local optima. Expectation maximization is known to be less susceptible to local optima as compared to kmeans clustering due to the logsumexp formula which is a smooth maximum (often called softmax) that falls out of marginalizing each data point across all clusters in log space[351]. Another clustering algorithm—K harmonic means—uses a similar technique choosing the harmonic mean of the distances from a datapoint to all clusters rather than K means' distance to the closest cluster center (the min function) as a loss function to be optimized. The harmonic mean can be thought of as a smooth minimum function. Both of these strategies allow a datapoint to partially affect cluster centers that are not their current best cluster center. Over time, this can lead a cluster center that is not currently the best cluster for many data points to drift towards the ones that are closest to it. As it does so, it may reduce the impact those data points have on another cluster. The combination of these effects tends to improve both of the primary error modes of clustering—splitting a true cluster across two cluster centers and assigning data points from multiple true clusters to a single cluster center. These soft maximum and soft minimum functions can be thought of as a continuous spectrum of how soft, or smooth, they are—from min or max to uniform. The shape of these functions between these extremes also matter, but the degree of smoothness tends to matter more. A comparison of these functions can be seen in figure 2.4.

Fig. 2.4: A comparison of smooth minimum and maximum functions



Harmonic mean is smoother than logsumexp, but applying an annealing temperature to logsumexp can make it arbitrarily smooth.

There is another method, deterministic annealing[1], which allows the degree of how smooth the function is across the optimization process[318][276] (see figure 2.4). Deterministic annealing, similar to the older and more widely known simulated annealing[166], takes its namesake from a process in metallurgy in which a metal object is heated to a high temperature and then cooled slowly in a controlled fashion that improves the molecular crystal structures and alters certain properties of the resulting product. They take their mathematical inspiration from statistical mechanics by treating the negative log likelihood as the energy of the system and in an attempt to find the minimum free energy, apply a temperature which begins high and is slowly reduced over time. The temperature dictates the degree of trade-off between exploration of a search space and exploitation of local gradients in the likelihood landscape.

Because the problem lends itself to a simple statistical model and deterministic annealing allows us to vary this tradeoff throughout the optimization process, I chose the deterministic annealing variant of expectation maximization for our clustering algorithm. The annealing process requires us to choose a meta-heuristic which is the temperature schedule. In deterministic annealing, the starting temperature is more important than in simulated annealing. With simulated annealing, a high temperature simply means a uniform search over the space regardless of the likelihood landscape. In deterministic annealing applied to clustering, if the temperature starts too high, it makes the data point's posteriors for each cluster uniform. After a few iterations of expectation maximization, the cluster centers may all be nearly identical making the gradients going forward vanishingly small leading to a symmetry breaking local optima. How smooth the soft max function needs to be is a function of the magnitude of the log-likelihoods of each data point, which, in this application, is largely dictated by how many alleles each cell expresses. Through empirical experimentation, I chose to initialize our temperature to one tenth the average number of alleles expressed by each cell. At each temperature step expecation maximization is run until the change in total log likelihood between steps is

[1] Rather than knowing or finding this method, I rediscovered it. I initially used the binomial loss function and had poor results. I implemented a sum of squared differences loss function and had much better results. I moved on and attempted to publish the work. Reviewer 2 asked why I didn't use the binomial loss function. I considered telling him that I tried it and the sum of squares loss function worked better. But this was unsatisfying. So I dug into why this was the case and the reason was that initially the likelihoods from the binomial loss preferred one cluster over another so much from the first step of random initialization, that they might never change to another cluster. This made me think of simulated annealing with the softening of the likelihood search space. I formulated the equivalent mathematical adaptation that that stochastic process took to this deterministic process which ended up being the same as this method published 22 years earlier. Satisfying, but perhaps I should have just done a more thorough literature search at the outset.

minimal (<0.1) which is used as the criteria of convergence. At each new temperature step, the temperature is halved until it is less than one at which point a final step is run at a temperature of one which reduces to the original likelihood function in equation 2.2. Cluster centers are randomly initialized and the optimization is run 50 times by default and the solution with the maximum total likelihood is chosen as the best solution. At each temperature step, a temperature modified posterior for each cell belonging to each cluster is defined as follows.

$$p_T(c \in k) = \frac{e^{\frac{\log(\mathcal{L}(A_{c,k}))}{T}}}{\sum_{i \in K} e^{\frac{\log(\mathcal{L}(A_{c,i}))}{T}}} \tag{2.3}$$

Which gives our maximization step according to the following equation.

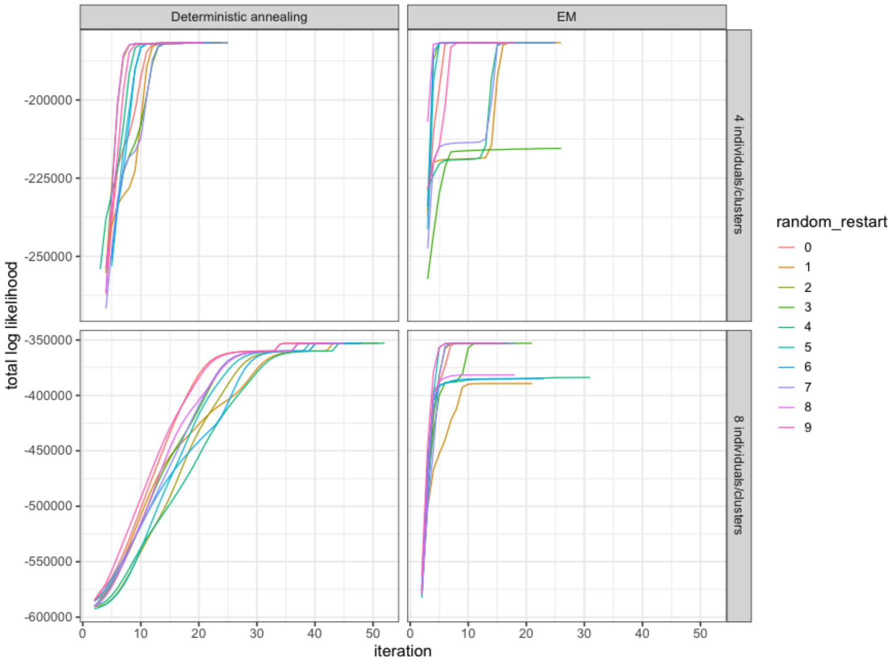
$$\phi'_{k,l} = \frac{\sum_{c \in C} A_{l,c,1} p_T(c \in k)}{\sum_{c \in C}(A_{l,c,1} + A_{l,c,0}) p_T(c \in k)} \tag{2.4}$$

In figure 2.5 you can see that deterministic annealing is better able to find the global optimum likelihood than expectation maximization and that even when it seems to be stuck in a local optimum, as is seen in the log likelihood plateaus in the graph, it is much more likely to find its way out. This becomes much more extreme as the number of individuals, or clusters, there are. And interestingly it also becomes more extreme as the amount of data per cell increases. This may be counter intuitive as you would think the amount of data per cell would make it easier to find the optimal clustering. But instead, the additional data makes the posterior probability for each cell to a cluster be closer and closer to zero or one even at a given random initialization of cluster centers. As the amount of data increases, the log likelihoods are of higher magnitude and the smoothness of the logsumexp function at higher magnitudes is less. The result of this is that it is very common for a given cluster center to be highly preferred over other clusters potentially for multiple individuals' cells and their effect on other clusters to be vanishingly small. This is why it is important both to use deterministic annealing as well as to use the amount of data per cell as a guide for the starting temperature.

2.2.4 Combinatorial experimental design for individual to cluster matching

There has been some concern in the community that it will be difficult to know which cluster corresponds to which individual after deconvolution with multiplexed scRNAseq

Fig. 2.5: EM vs Deterministic Annealing on four and eight individuals



Deterministic annealing finds the optimal clustering (and thus highest likelihood) in all cases whereas EM fails in one random restart with four individuals. With eight individuals, EM fails in several random restarts while Deterministic annealing still finds the optimal clustering every time. This difference becomes much more dramatic with more individuals / clusters.

experiments when genotypes are not known a priori. To address this, I propose an experimental design involving m overlapping mixtures for $2m-1$ multiplexed individuals outlined in table 2.3. Each individual is assigned a binary number $1..2m$, where each bit corresponds to the inclusion (1) or exclusion (0) from each of the mixtures. This gives each individual a unique signature of inclusion/exclusion across the mixtures. Although each sample is in a different number of mixtures, the number of cells per experiment can be adjusted according to the number of mixtures that contain that sample. Souporcell provides a tool to match clusters from two experiments with shared samples.

Table 2.3: Experimental design for matching individuals to clusters

a

Mixture	1	2	3
Individual a	0	0	1
Individual b	0	1	0
Individual c	0	1	1
Individual d	1	0	0
Individual e	1	0	1
Individual f	1	1	0
Individual g	1	1	1

b

Mixture 1	d	e	f	g
Mixture 2	b	c	f	g
Mixture 3	a	c	e	g

This table outlines an experimental design of seven individuals with three overlapping mixtures to allow for clusters to be assigned to individuals. **a)** Shows the mapping of individuals to binary numbers where each digit of the binary number represents inclusion/exclusion from a mixture. **b)** shows the resulting mixtures.

2.2.5 Doublet cell barcode detection

One of the major aims of this work is to detect the barcodes that contain multiple cells with different genotypes. I do not, however, attempt to detect barcodes that contain multiple cells with the same genotype. I make the assumption that the generation of doublet cell barcodes is a random Poisson process and that the rate of this Poisson process is low enough that the chance of droplets with more than two cells are exceedingly unlikely. This is true for the standard experimental design, but is not in the case of super loading cells into the system. As discussed in chapter 1, I advise against this for several reasons. I view this problem as an urn problem in which each cluster is an urn containing alleles expressed by all of the cells assigned to that cluster. Then each cell is inspected to determine if its alleles were more likely to be drawn from the single best cluster or the allele counts of the combination of the top two clusters for this cell.

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.

Allele counts of each cell at each locus are treated random variables drawn from a beta-binomial distribution from either a single cluster or a pair of clusters. The beta-binomial is used to model our uncertainty in the binomial parameter p. For a single cluster the parameters are alpha = 1+alt counts and beta = 1+ref counts. For the singleton case, the likelihood of the data is as follows.

$$\mathcal{L}(c \in K_i) = \prod_{l \in L_c} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,1}} \frac{\beta(A_{l,c,0} + 1 + A_{i,l,0}, A_{l,c,1} + 1 + A_{i,l,1})}{\beta(1 + A_{i,l,0} + A_{i,l,1})} \tag{2.5}$$

Where β is the beta function and cluster i is the best fitting cluster for cell c. The expected allele fractions of a doublet coming from cluster i, and cluster j is the average of the allele fractions of the two clusters. To obtain the pseudocounts needed to parameterize the beta-binomial, the counts of alleles from the cluster with the fewer alleles at this locus are used. That is,

$$alpha_{l,i,j} = 1 + \frac{\frac{A_{i,l,0}}{A_{i,l,0}+A_{i,l,1}} + \frac{A_{j,l,0}}{A_{j,l,0}+A_{j,l,1}}}{2} min(A_{i,l,0} + A_{i,l,1}, A_{j,l,0} + A_{j,l,1}) \tag{2.6}$$

$$beta_{l,i,j} = 1 + \frac{\frac{A_{i,l,1}}{A_{i,l,0}+A_{i,l,1}} + \frac{A_{j,l,1}}{A_{j,l,0}+A_{j,l,1}}}{2} min(A_{i,l,0} + A_{i,l,1}, A_{j,l,0} + A_{j,l,1}) \tag{2.7}$$

The doublet likelihood given those conservative parameters becomes

$$\mathcal{L}(c \in K_i \cup K_j) = \prod_{l \in L_c} \binom{A_{l,c,0} + A_{l,c,1}}{A_{l,c,1}} \frac{\beta(A_{l,c,0} + alpha_{l,i,j}, A_{l,c,1} + beta_{l,i,j})}{\beta(alpha_{l,i,j} + beta_{l,i,j})} \tag{2.8}$$

The posterior for each cell being a doublet is then given by

$$p(doublet_c|c) = \frac{p(c \in K_i \cup K_j)p(doublet)}{p(c \in K_i \cup K_j)p(doublet) + p(c \in K_i)(1 - p(doublet))} \tag{2.9}$$

Where cluster i is the best fitting cluster for cell c and cluster j is the second best fitting cluster for cell c. The prior can be set by the user but have used an uninformed prior of 0.5 for all of our analysis.

The above process is run iteratively removing doublets found until no new doublets are found.

2.2.6 Ambient RNA detection and Cluster genotype coinference

One major goal of clustering scRNAseq by genotypes is calling the genotypes for each individual/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, the prior knowledge of ploidy of the organisms can be used along with our genotype cluster assignments to make a co-inference of both the genotypes and level of ambient RNA in the experiment.

2.2.6.1 Mixture model of ambient RNA and cell RNA

Definitions:

- ρ: probability any given allele is arising from ambient RNA as opposed to from the cell associated with that barcode. This will be learned.
- P: ploidy. Currently, only ploidy of one or two is supported.
- 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. The expected reference allele rate without ambient RNA is g and g is an integer value $[0..P]$. Note that for biallelic variants and ploidy 1 or 2, g is sufficient to uniquely determine the genotype.
- $p(true)$: prior for variant being a true variant vs a false positive. The default is 0.9 which was the value used for all analyses.

Once again, a maximum likelihood approach is taken.

$$\underset{\rho}{\operatorname{argmax}} \mathcal{L}(data, \phi) \tag{2.10}$$

Here, the proportion of ambient RNA in the system, ρ, is the only free parameter and is optimized using maximum likelihood. 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). Each cluster is treated as independent and each locus as independent, before marginalizing across the possible genotypes. The model also considers the possibility of the variant being a false positive. In this case, the variant will not segregate into distinct allele frequencies between different clusters and it will most likely not attain a value close to the standard allele frequencies expected from the diploid or haploid genotypes. Thus, the allele counts in each cluster are modeled as having come from a mixture of ambient RNA (an average allele fraction in the experiment) and from the cells in that cluster. The observed allele fractions are assumed to have been drawn from a binomial distribution with a probability that was skewed away from $p = g/P$ by the level of ambient RNA ρ. Thus, the probability of the binomial from which the allele counts are drawn for true positive variants is the following.

$$p_{tp} = (1-\rho)\frac{g}{P} + \rho\frac{A_{l,0}}{A_{l,0}+A_{l,1}} \tag{2.11}$$

For a false positive the parameter is

$$p_{fp} = \frac{A_{l,0}}{A_{l,0}+A_{l,1}} \tag{2.12}$$

Thus, the full model is

$$\begin{aligned} p(data|\rho) = \prod_{l\in L}\Bigg[&p(true)\prod_{k\in K}\sum_{g=0}^{P}\frac{1}{P}\binom{A_{k,l,0}+A_{k,l,1}}{A_{k,l,0}}p_{tp}^{A_{k,l,0}}(1-p_{tp})^{A_{k,l,1}} \\ &+(1-p(true))\prod_{k\in K}\binom{A_{k,l,0}+A_{k,l,1}}{A_{k,l,0}}p_{fp}^{A_{k,l,0}}(1-p_{fp})^{A_{k,l,1}}\Bigg] \end{aligned} \tag{2.13}$$

2.2.6.2 Inference

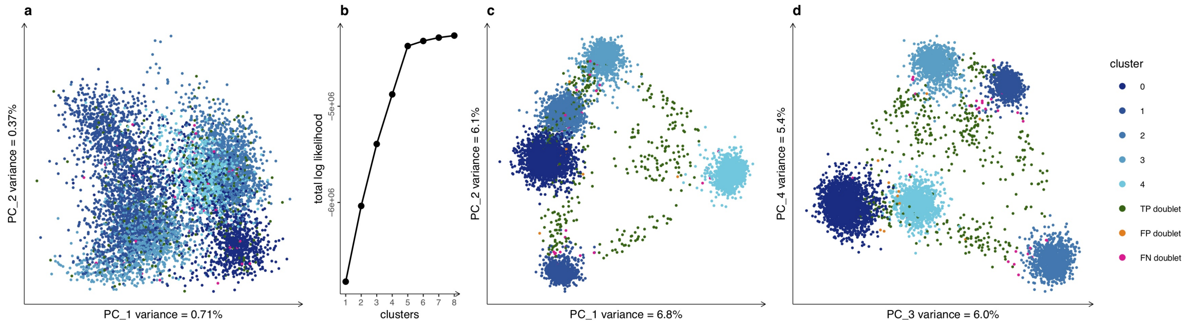
We solve for ρ with gradient descent using the statistical modeling domain specific language STAN. Next, the posterior of the variant being a true variant is calculated for each of the three (or two in the haploid case) genotypes versus it being a false positive. The prior on variants being true positives can be set by the user, but defaults to 0.9 which is the value used in our analyses.

2.3 Results

2.3.1 Benchmarking: Synthetic human cell mixture

Currently, there are no good generative models available for batch effects, allele-specific expression, ambient RNA, and doublets in scRNAseq that can be used to generate *in silico* data for testing methods that cluster by genotype. To generate realistic data with known ground truth we sequenced five lines of induced pluripotent stem cells (iPSCs) from the Human iPSC initiative[305] with the 10x Chromium single cell system, both individually and in a mixture of all five lines (with three replicates of the mixture). Each mixture contained 5-7,000 cells and 25,000 UMIs per cell. We first synthetically mixed 20% of the cells from the 5 individual samples while retaining their sample of origin. To make the synthetic mixture as close to real data as possible, I also simulated 6% doublets by switching all of the reads' barcodes from one cell to that of another cell and 5% ambient RNA by randomly switching cell barcodes for 5% of the reads. A low dimensional representation of the expression matrix reveals relatively little variation, as expected, because there is only one cell type present (2.6a). Indeed, the most significant driver of expression appears to be the donor of origin, but the donor cells overlap in expression patterns and it is not possible to assign a donor to each cell based solely on expression patterns.

Fig. 2.6: Synthetic human mixture

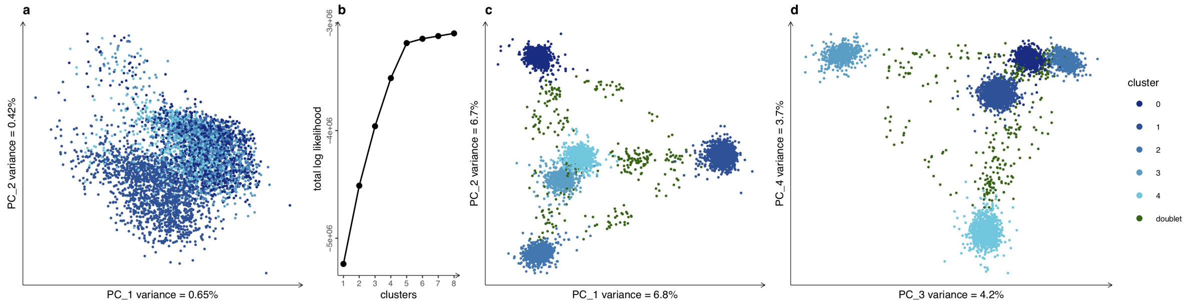


a) Expression PCA of a synthetic mixture cells from five HipSci cells lines (n=7073 cells) with 5% ambient RNA and 6% doublets colored by known genotypes. Because these samples only contain one cell type, the largest remaining source of variation in the expression profile comes from the genotype, although the signal is not sufficient for accurate genotype clustering. **b)** Elbow plot of the number of clusters versus the total log likelihood showing a clear preference for the correct number of clusters (k=5). **c** and **d)** PCA of the normalized cell-by-cluster log likelihood matrix from souporcell (n=7073 cells). As this is a synthetic mixture in which the ground truth is known, the genotype clusters are colored and errors are highlighted in orange (false positive doublets) and pink (false negative doublets).

2.3.2 Benchmarking: Real human cell mixture

Next I compare to a true mixture of human cells in which the ground truth is not known. The results (shown in 2.7) are strikingly similar to those in 2.6 which suggest that our synthetic mixture is realistic.

Fig. 2.7: Experimental human mixture



a) Expression PCA of a single replicate of the experimental mixtures (n=4925 cells) colored by genotype clusters from souporcell. **b)** Elbow plot of the total log likelihood versus different numbers of clusters showing a clear preference for the correct number of clusters. **c** and **d)** PCA showing the first four PCs of the normalized cell-by-cluster log likelihood matrix colored by cluster (n=4925 cells).

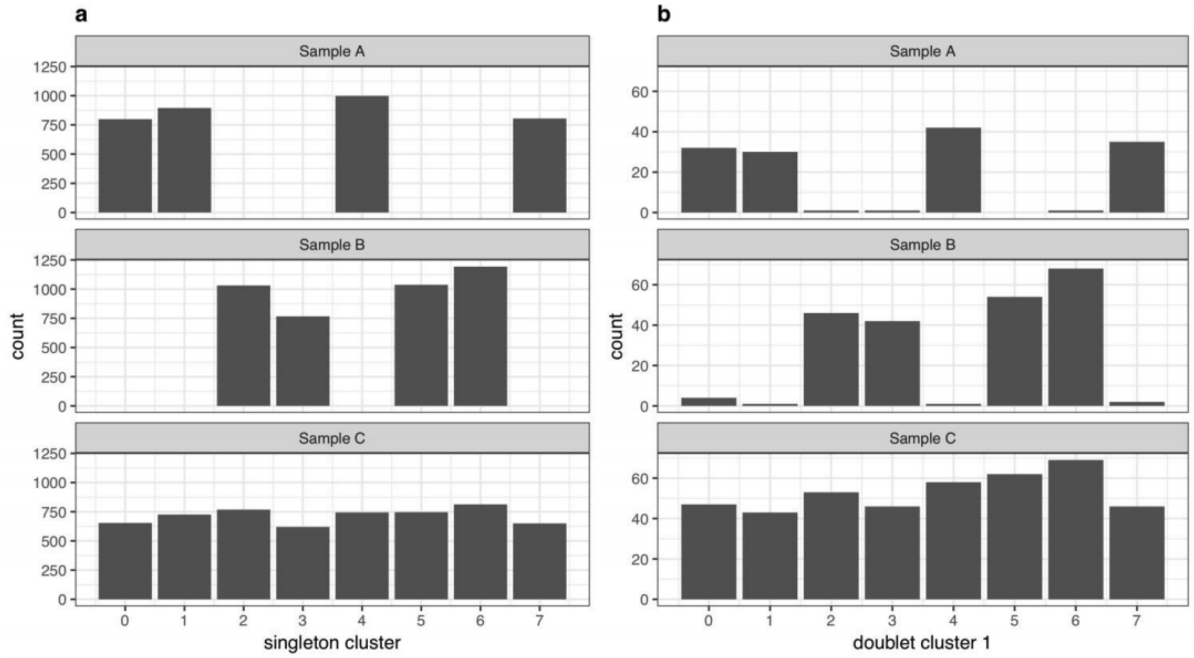
2.3.3 Benchmarking: demuxlet paper dataset

In order to demonstrate souporcell on an external and widely used benchmark dataset, I downloaded the three overlapping mixtures from the demuxlet paper[154] . Sample A contains a mixture of four donors' PBMCs, Sample B contains a mixture of four different donors? PBMCs, and Sample C contains a mixture of all 8 donors' PBMCs. I synthetically combined this data into a single dataset and clustered with souporcell. Figure 2.8a shows that the resulting clusters either contain cells from Sample A or Sample B, but not both as is expected from this experimental setup. I also show that the first cluster of the doublet assignments are also largely consistent with this experimental design (figure 2.8b).

2.3.3.1 Deconvolution of overlapping mixtures

To enable identification of which cluster is which individual using the overlapping mixture experimental design outlined in Table 1, I provide a tool shared_samples.py that takes

Fig. 2.8: Demuxlet data



a) souporcell cluster assignments of singletons for combined dataset showing that Sample A and Sample B are non-overlapping and Sample C contains all 8 samples. **b)** shows the first cluster of the doublet assignment for doublets showing largely non-overlapping assignments between Samples A and B.

as input two souporcell output directories and the number of samples which are shared. It compares the sum of squared differences of the allele fraction of confident (>95% confident genotype call in all clusters) shared variant calls between clusters in the two experiments and outputs the best matches for the number of shared samples. I tested this using multiple synthetic mixtures of 5 HipSci cell lines with 6% doublets and 5% ambient RNA and gave both as input to the shared_samples.py tool and it correctly assigned the clusters in one run to the clusters in the second experiment which corresponded to the same samples. I also ran souporcell on the three demuxlet datasets separately and ran the shared_samples.py tool on Sample A vs Sample C and Sample B vs Sample C and it confidently identified the non-overlapping clusters in Sample C which correspond to A and B.

2.3.3.2 Validation and comparison to other methods

I compare souporcell to demuxlet, vireo, and scSplit. Demuxlet uses the prior knowledge of the individuals' genotypes along with a statistical model to assign cells to genotypes and call doublets. It does not model ambient RNA, and due to the rigid model may interpret ambient RNA and other sources of noise as signal that this cell is more likely a doublet than a single genotype. Souporcell, Vireo, and ScSplit all cluster cells by genotype without requiring genotypes *a priori* and also all use cluster center based clustering methods. Vireo uses variational inference EM. The self regularizing nature of variational inference has a similar effect to souporcell's deterministic annealing by having less confidence in the cluster assignments until the clusters much better match the data. Vireo also over clusters initially by adding additional cluster centers. Then once the variation inference has converged, vireo picks the overclustering clusters that are the smallest and tries to identify which larger cluster these cells could belong to. While this can help with local maxima, this can run into problems when the some true clusters have a very small number of individuals. After some dialogue with the author of Vireo, the later versions pick the clusters which are most different than other clusters as the final clusters rather than throwing out the small clusters. scSplit uses an iterative process to find the most informative SNPs, using them to cluster, and using that clustering to find the most informative SNPs for further iterations. In the clustering process, they use EM and only use multiple random restarts to get around local maximum which will begin to fail when the number of clusters is high. In addition to this, there appears to be a math error in the scSplit paper in which the binomial loss function excludes the binomial coefficient. It is not clear to me what affect this will have on the clustering process, but will weight loci improperly. As discussed previously, souporcell uses a cluster center based approach which is robust to ambient RNA and false positive variants and uses a determinisic annealing approach to overcoming local maxima in the clustering process. These combined produce reliable clustering across many sample types and error modes.

To compare souporcell to vireo and scSplit, the two other tools that do not require prior genetic information, I first ran variant calling and cell allele counting as recommended for each tool. Using souporcell, I clustered cells by their genotypes, and evaluated the correct number of clusters through an elbow plot comparing the total log probability versus a varying number of clusters (2.6b). The clustering output can be viewed as a matrix with cells as rows and clusters as columns with the values being the log likelihood of that cell versus the corresponding cluster. To visualize the five clusters identified by genotype I carried out a PCA of the normalized log likelihood matrix, which reveals a clear separation of the clusters, with interspersed doublets (2.6c and d). For these data souporcell assigned 6612/6622 singletons and 415/451 doublets correctly; four singletons were falsely labeled as a doublet, 35 doublets were misidentified as singletons, and one doublet and four singletons were unassigned. I carried out the same analysis for the three replicates of the experiment mixtures and show results for one (2.7). The expression PCA (Fig. 2e) and normalized cell-cluster loss PCA (2.7c,d) of the experimental mixture were similar to the synthetic mixture indicating that the synthetic mixtures were an accurate approximation of real mixtures. To compare doublet detection between methods, I calculated a receiver-operator characteristic (ROC) curve of the doublet calls (2.9i) on a synthetic mixture with 6% doublets and 10% ambient RNA that showed the area under the curve values of 0.98 and 0.91 for souporcell and vireo, respectively. I also show point estimates for the doublet threshold chosen. Demuxlet's posterior doublet probability output did not have enough significant digits and is 1.0 until it starts varying with 27% false positives. The default doublet probability threshold for demuxlet gives nearly 40% false positive doublets.

Each of the five human iPSC lines has existing WGS data generated as part of the HipSci Project[161]. Therefore, for the experimentally mixed replicates, I compared each tool's clustering to sample assignments obtained from demuxlet using genotypes available from the WGS. Demuxlet significantly overestimates doublets versus expectations based on the number of cells loaded[355] especially as ambient RNA increases (2.9b). Because I could not trust the doublet calls of demuxlet, I allowed scSplit, vireo, and souporcell to exclude their called doublets and then compared the remaining cells to demuxlet's best single genotype assignment. The Adjusted Rand Index (ARI) of the remaining cell assignments versus demuxlet were 1.0 (fully concordant) for souporcell and vireo across the three replicates and an average of 0.97 for scSplit.

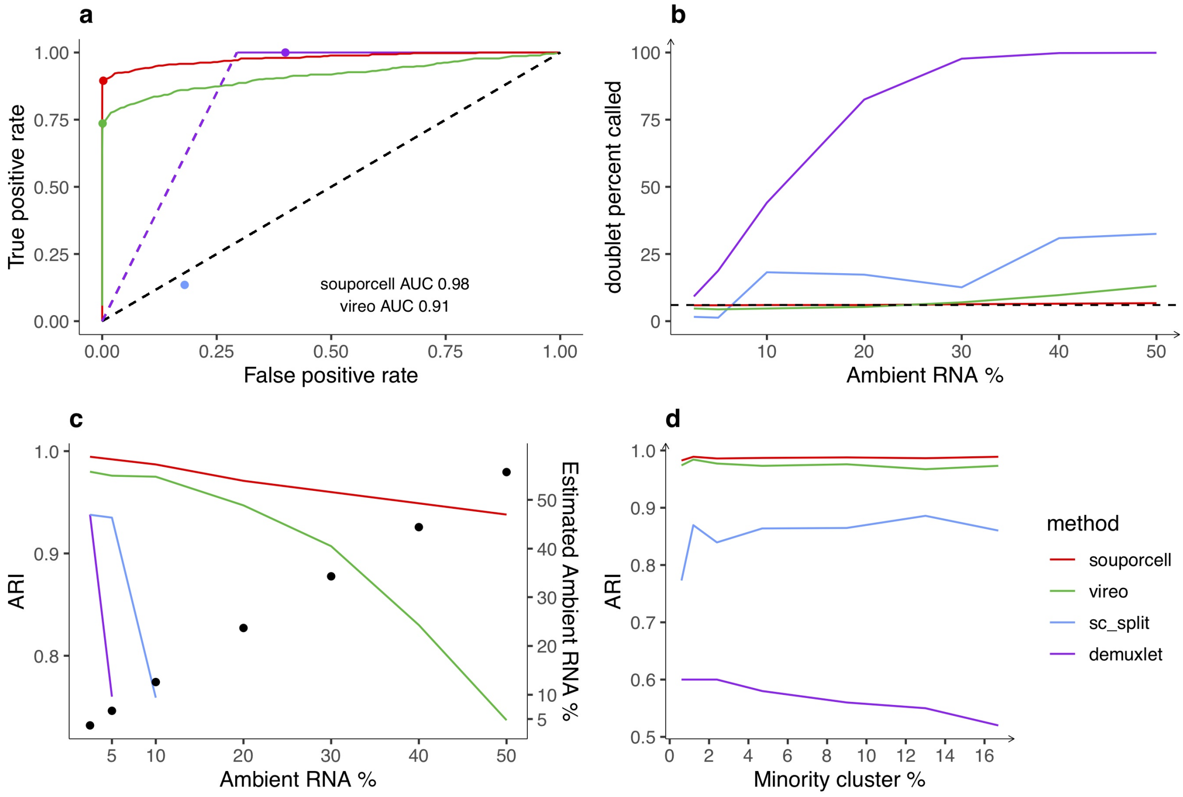
To evaluate the robustness of each tool across a range of parameters, I created synthetic mixtures of the five individual human iPSC scRNAseq experiments to test both the sensitivity to the ambient RNA level (2.9b,c) and the ability to accurately assign cells to a cluster if it is much smaller than other clusters (2.9d). For the ambient RNA experiment, I synthetically combined 20% of the cells from each of the five individual samples and simulated 6% intergenotypic doublets and a range of ambient RNA from 2.5%-50% representing realistic ranges previously reported[348]. I found that souporcell and vireo retain high accuracy with souporcell being more robust at accurately calling doublets in high ambient RNA cases (figure 2.9c). The ARI of scSplit and demuxlet suffered due to poor doublet detection. With these data I also show that souporcell is able to accurately estimate the amount of ambient RNA in the experiment (figure 2.9c). To test robustness to sample skew, e.g., one donor's cells are underrepresented, I created a set of synthetic mixtures with 1,000 cells from each of four individual samples and 25-800 cells for the minority cluster including 8% ambient RNA and 6% doublets (2.9d). I found that all tools performed well down to the minority cell cluster comprising only 1.2% (50 cells) of total cells (Fig. 2m), but only souporcell and vireo were able to correctly identify all minority sample singletons as their own cluster down to 0.6% of all cells. Again, demuxlet's poor ARI was due primarily to extremely high levels of false positive doublets (figure 2.9a).

I then compared souporcell's genotype and ambient RNA co-inference to vireo and scSplit versus the variants called from whole genome sequencing data. In scRNAseq data most variants have very low coverage per cluster compared to what would be generated from WGS data, thus the genotype accuracy is significantly lower than one would attain with genome sequencing. Nevertheless, souporcell surpasses both vireo and scSplit in genotype accuracy on a synthetically mixed sample with 6% doublets and 10% ambient RNA. The most common error mode for vireo and scSplit is calling homozygous reference loci as heterozygous variants, which is expected when ambient RNA is not accounted for, as it is not in these two tools.

2.3.4 Maternal-Fetal data

Next, we considered more challenging scenarios involving multiple cell types, widely varying numbers of cells per sample, and closely related genotypes. The decidua-placental interface plays an important role in pregnancy and birth, and is of importance to several

Fig. 2.9: Comparison to competing methods



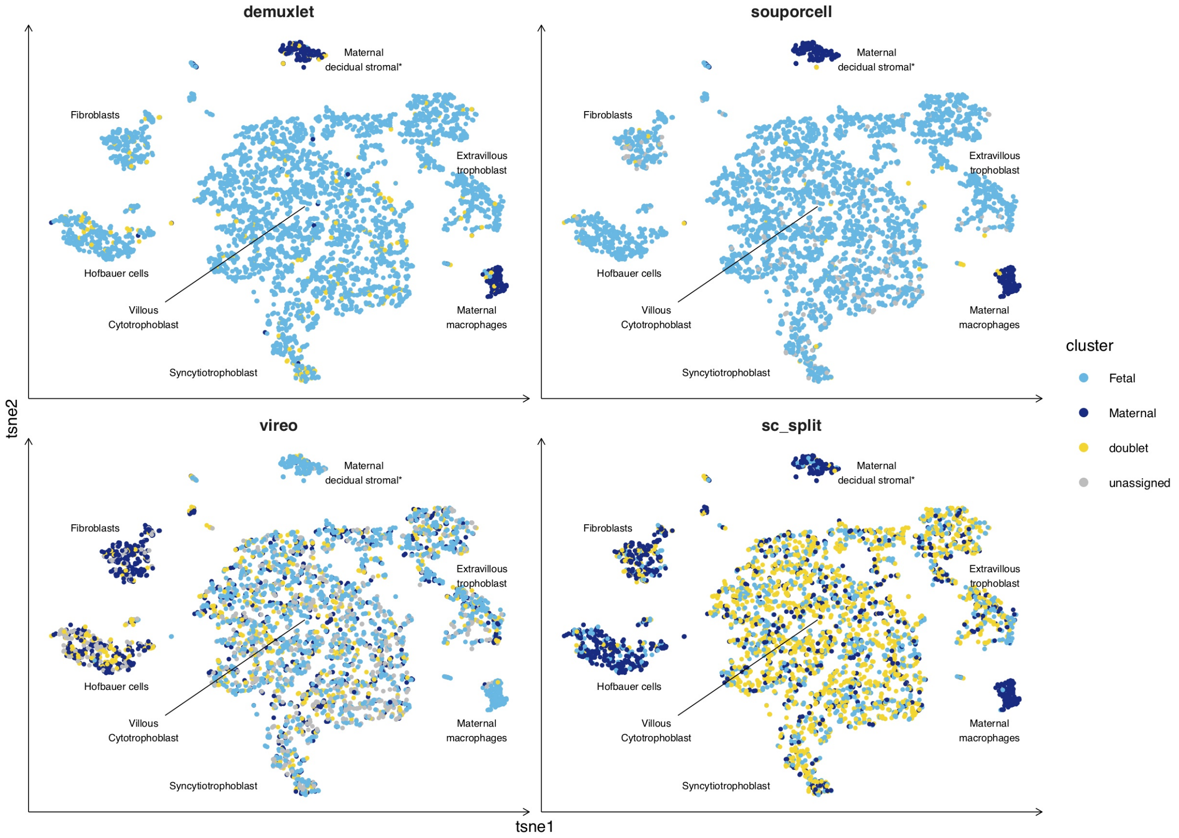
a) ROC curve of the doublet calls made by souporcell and vireo and a point estimate for scSplit (blue dot) for a synthetic mixture with 6% doublets 451/7073 and 10% ambient RNA. I show both the curves and the threshold chosen (points) for each tool. scSplit did not give a score so I simply show the point estimate. Demuxlet's doublet probabilities were all 1.0 until the solid line starts, so I show a theoretical dotted line up to that point. **b)** Doublet call percentages for all tools on synthetic mixtures for varying amounts of ambient RNA versus the actual doublet rate (dotted line). **c)** Adjusted Rand Index (ARI) versus the known ground truth of synthetic mixtures with 6% doublets and a varying amount of ambient RNA. For levels >=10% ambient RNA, scSplit identified one of the singleton clusters as the doublet cluster, which means that the ARI was not clearly interpretable. Right y-axis vs points shows the estimated ambient RNA percent by souporcell versus the simulated ambient RNA percent. **d)** ARI of each tool on a synthetic mixture with 8% ambient RNA and 6% doublet rate with 1,000 cells per cluster for the first four clusters and a variable number of cells in the minority cluster (25-800 cells in the minority cluster).

diseases, including pre-eclampsia[326]. Recently, more than 70,000 cells were profiled by scRNAseq[232] to explore the transcriptional landscape at this interface. The decidua is primarily composed of maternal cells with some invading fetal trophoblasts, while the placenta is largely composed of cells of fetal origin with the exception of maternal macrophages. In the study exploring this interface[326], WGS from blood and placenta was used to genotype both mother and fetus, and demuxlet was used to assign cells to each individual. Here, I applied souporcell, vireo, and scSplit to two placental samples and one decidual sample from a single mother to determine if cellular origins could be established without reference genotypes. I show the expression t-SNE of a single placental sample labeled by cell type annotation[326] and colored by genotype cluster as assigned by each method (2.10). While souporcell clusters agree with demuxlet and segregate with the expected cell type clusters, vireo and scSplit have major discordances with demuxlet. This is similar for the other samples tested. Comparing souporcell to demuxlet, there are 21 cells that demuxlet labels as maternal or fetal but which appear in the other individual's cell type clusters. Based on the position of these cells in the expression t-SNE plot, it is most likely that these are errors in the demuxlet assignments that are not made by souporcell.

2.3.5 Plasmodium

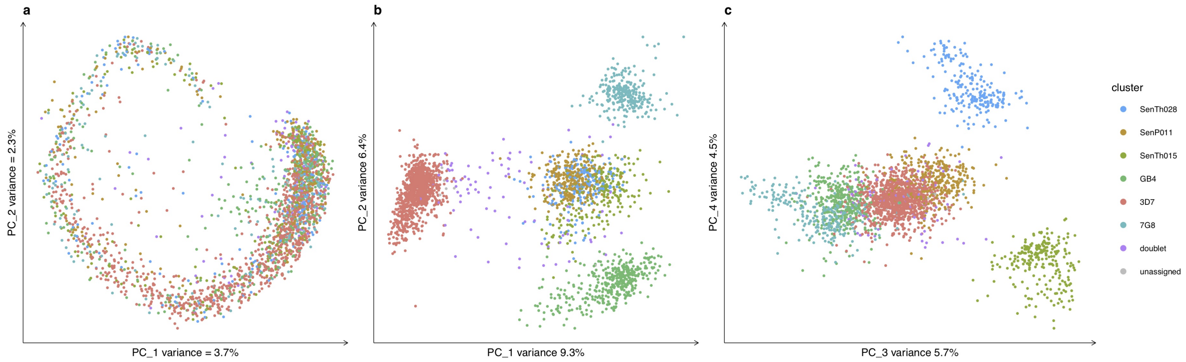
I also tested souporcell on a non-human sample, the single-celled malaria parasite *Plasmodium falciparum*, for which single cell approaches are now used to explore natural infections[134]. Malaria infections often contain parasites from multiple different genetic backgrounds, and it is not possible to separate the strains prior to sequencing. These samples differ from human samples in a variety of ways; they are haploid when infecting humans, the genome is $> 80\%$ A/T, and the transcriptome is only ~ 12 megabases (genome is ~ 23 Mb). We generated three datasets containing six genetically distinct strains of *P. falciparum* sampling 1893-2608 cells with median UMIs of 1000. Analysis of the expression profile of one of these reveals that the genotypes are distributed across the *Plasmodium* intra-erythrocytic cycle (2.11a) while being well separated in normalized loss cluster space(figure 2.11b,c). The ARI for each method on the three *Plasmodium* data sets show superior performance for souporcell across the board, with scSplit suffering on all datasets and vireo performing poorly on one, which had an ARI versus demuxlet of 0.24. This sample was more difficult due to sample skew caused by a clonal expansion of one of the six strains.

Fig. 2.10: Maternal/fetal data



Cell expression t-SNE plots of n=3,835 cells colored by each tool's genotype assignments or clusters for a placental sample. Cell phenotype clusters and cell genotype clusters co-segregate, with the majority of cell types being of fetal origin with the exception of maternal macrophages and *maternal decidual stromal cells, the latter of which (found only in one donor) were considered to be a non-placental artefact arising from the surgical procedure and were removed during data quality control in the original study[326]. Concordance is high between souporcell and demuxlet (ARI 0.96) whereas vireo and scSplit have large discordances with ARI of 0 and 0.03 respectively.

Fig. 2.11: Plasmodium data



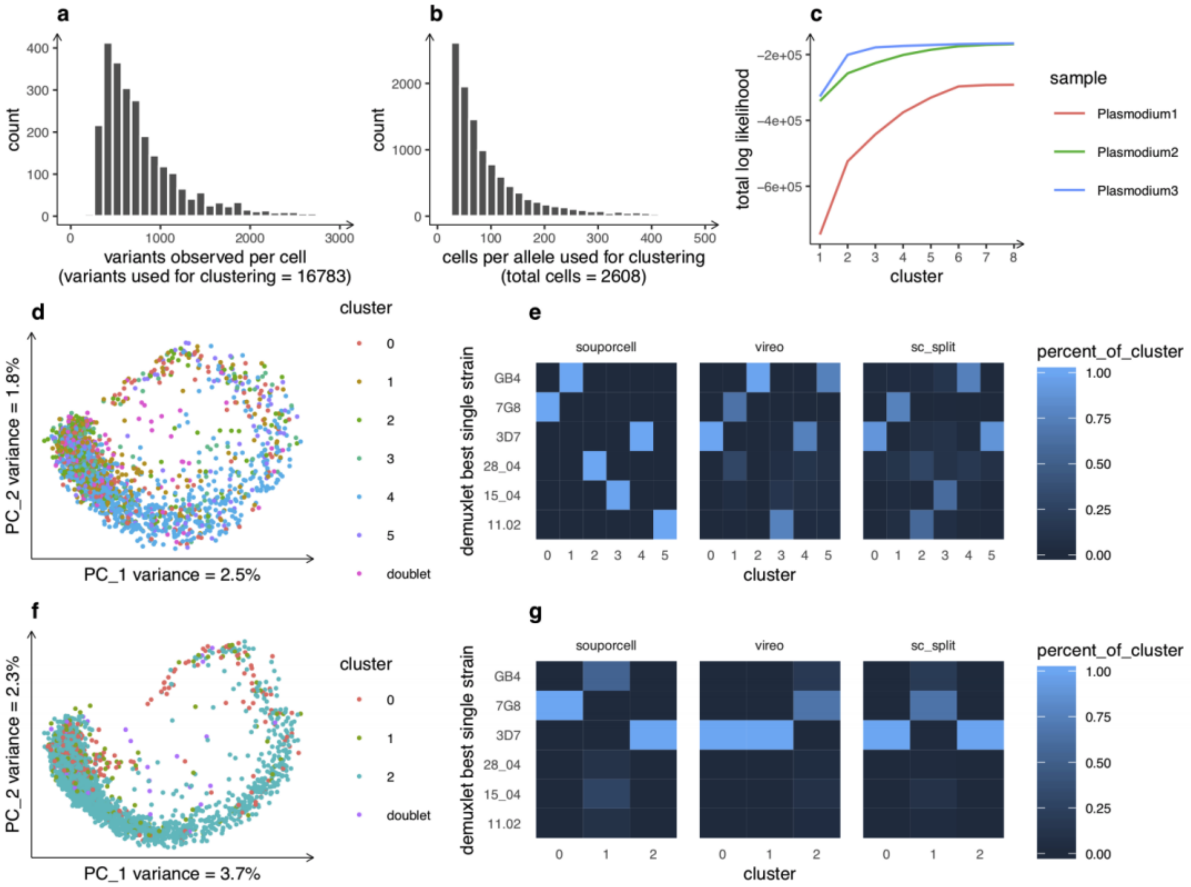
b) Expression PCA colored by genotype clusters for Plasmodium sample 1 (n=2608 cells) showing an even spread of genotypes throughout the asexual lifecycle. **b** and **c)** PCAs of first four PCs of souporcell's normalized cell-by-cluster loss matrix showing good separation of each genotypic cluster (n=2608 cells).

We did three *Plasmodium falciparum* mixture experiments. In the one shown above (Plasmodium1), the cells were mixed and immediately prepared for single cell sequencing. In the Plasmodium2 sample, the cells were mixed and then fixed in methanol before being prepared for scRNAseq. In the Plasmodium3 sample, cells were mixed and then grown in culture for seven days prior to single cell sequencing. Because the initial mixture was not very equal, the majority strain out grew the other strains dramatically. This caused the number of cells from some of the other strains to contain very few cells and be more difficult to cluster. Figure 2.12 shows the results of the 2nd and 3rd mixtures. The Plasmodium3 sample which was cultured for seven days prior to being sequenced did not cluster into six clusters well. The elbow plot seemed to support a K of 3 more than the true number of strains mixed. This shows some of the limitations of souporcell and clustering in general with highly skewed number of cells per sample.

2.3.6 Twenty one individual mixture demonstration

I demonstrate that souporcell is capable of demultiplexing many donors by creating a synthetic mixture of 21 different individuals, which given the current recommendations from 10x on cells per run would be a high-end number of donors to multiplex. To generate this 21-donor mix, I used the 5 HipSci samples described in figure 2.6 and added to them 16 PBMC samples obtained from the Human Cell Atlas Census of Immune Cells. From each dataset I randomly selected 1000 cells with at least 4000 UMIs and simulated 10%

Fig. 2.12: Plasmodium data replicates



a) Distribution of number of variants observed per cell used for clustering (with at least 4 cells required to support each allele) and the total number of variants used for clustering on the Plasmodium1 sample. **b)** Distribution of counts of the number of cells expressing each allele used for clustering as well as the total number of cells in the Plasmodium1 sample. **c)** Elbow plots for each Plasmodium data set show relatively strong support for the correct number of clusters (6) for Plasmodium1, but less clear results for Plasmodium2, which suffered from higher amounts of ambient RNA, and for Plasmodium3 due to bias towards three genotypes rather than a relatively even mixture. For this reason, I analyzed Plasmodium3 with k=3. **d)** Expression PCA of the Plasmodium2 sample (1893 cells) colored by genotype clusters as called by souporcell. **e)** Confusion matrix heatmap of the demuxlet best single strain (Y axis) versus souporcell, vireo, and scSplit. For souporcell one cluster per strain is seen as expected. Both vireo and scSplit have the majority strain, 3D7, split across two clusters and two other strains combined into a single cluster. **f)** Expression PCA of the Plasmodium3 sample (2293 cells) colored by genotype clusters as called by souporcell. **g)** Confusion matrix heatmap of the demuxlet best single strain (Y axis) versus souporcell, vireo, and scSplit genotype clusters with k=3. Souporcell clusters out the 3D7 and 7G8 strains correctly and puts all other cells into the final cluster while both vireo and scSplit put 3D7 into two clusters and all other cells into the remaining cluster.

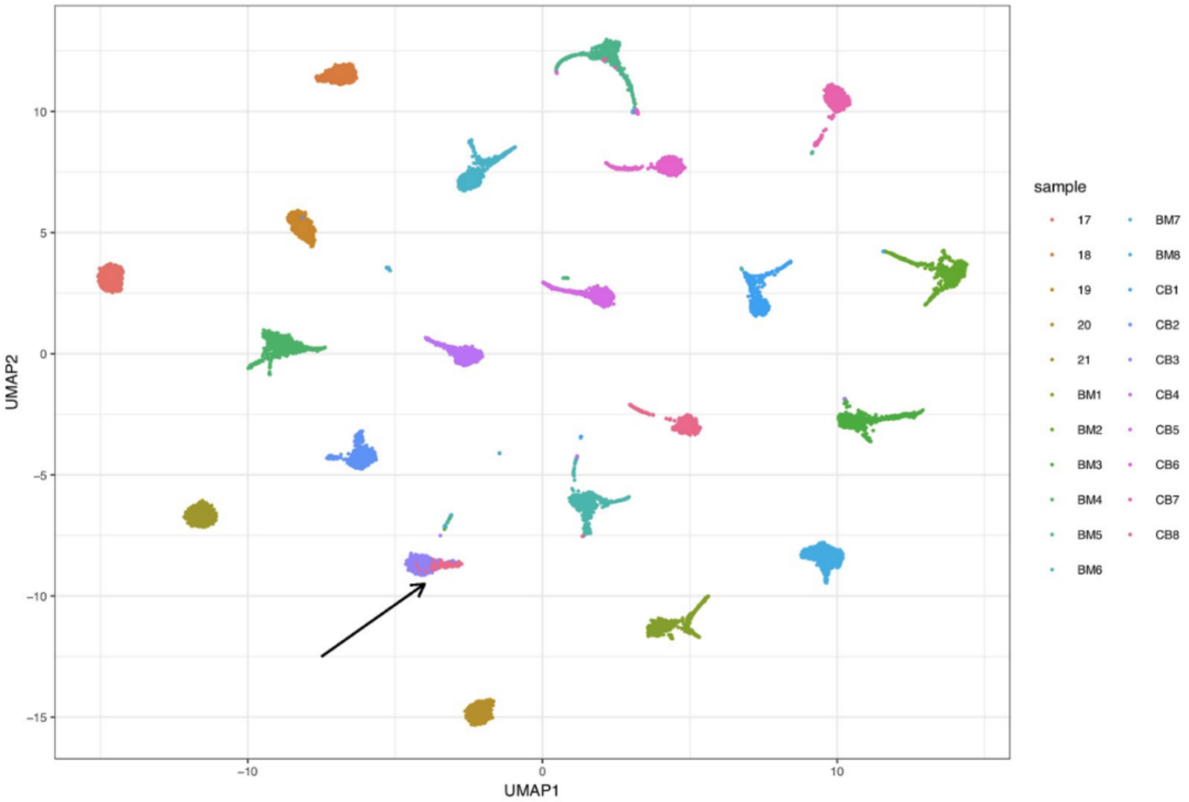
doublets and 2.5% ambient RNA by altering the cell barcodes, as described above. I clustered these with souporcell and the software correctly identifies 1690 of the 2100 synthetic doublets. A further 69 cells were unassigned, and in total has an ARI of 0.95. Excluding all doublets the ARI is 0.98. I found that a total of 134/16800 singletons misassigned where 129 of them are CB8 cells assigned to the CB3 cluster. I show later that this is likely because the CB8 sample is contaminated by another (non CB3) donor. 2.13 shows the UMAP projection of the normalized cluster log likelihood matrix. It is clear that souporcell is able to handle at least 21 distinct donors and accurately assign cluster identities to the majority of cells.

Because the misassignment of CB8 cells accounted for >95% of singleton errors, we suspected this may be due to contamination. I repeated this experiment with several of the replicates of the CB8 donor and found consistent results. I then made a synthetic mixture of CB3 and CB8 in order to determine if this was due to the large number of donors and it was not. I still found that roughly 20% of CB8 cells would cluster with CB3, but if given 3 clusters, all of those cells formed their own cluster. This made us suspect that the CB8 sample was contaminated with cells from a different (non CB3) donor.

2.3.6.1 Contamination revealed

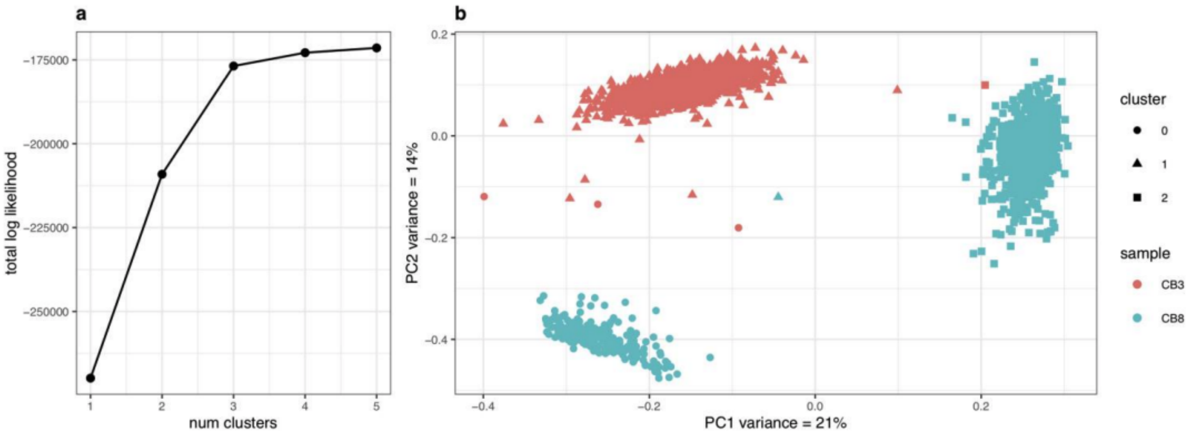
To further test whether the CB8+CB3 “misassignment” was an error or true signal, I created a synthetic mixture of all cells from both the CB8 sample and the CB3 sample. I ran souporcell with a range of K from 1 to 5 and plotted the elbow curve shown in figure 2.14a and the PCA of the cell cluster likelihoods for the clearly optimal number of clusters which was three. This PCA has good separation and gives good evidence that the CB8 sample was contaminated with cells from another donor that were most likely more closely related to the CB3 sample than the CB8 sample. I followed up with the creators of the census of immune cells data resource and they said that they were already aware of a contamination in the CB8 sample. This corroborated discovery was not picked up by the vireo team, which used the same data with which they reported high concordance. This shows further the power of souporcell for detection of unexpected events such as sample contamination.

Fig. 2.13: 21 donor example



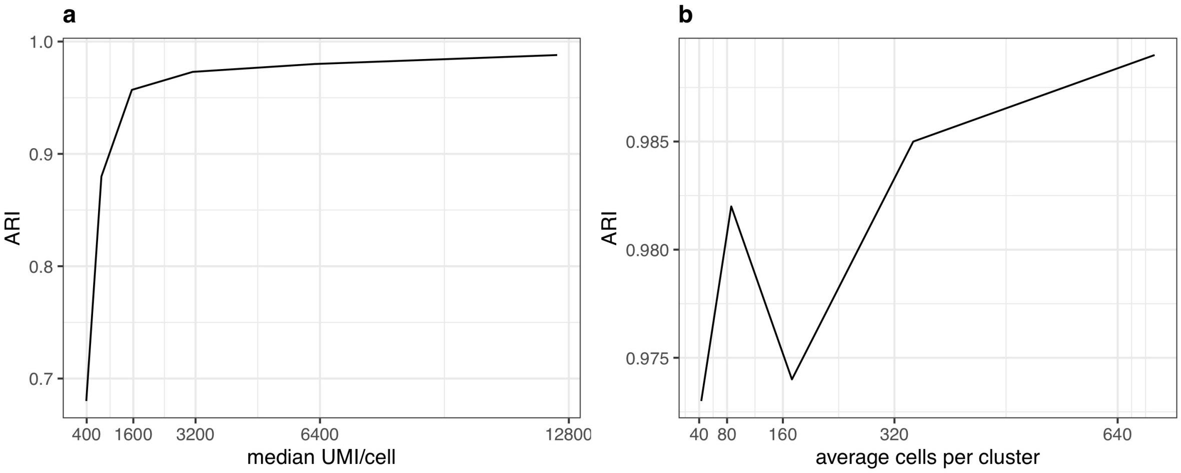
UMAP of the normalized log likelihood cluster matrix for the singletons of a mixture of the 5 HipSci samples and the 16 PBMC samples from the Human Cell Atlas project. The main error is the assignment of 129 CB8 cells to the CB3 dominant cluster indicated by the arrow. I show that this is likely due to contamination (see figure figure:contamination).

Fig. 2.14: Contamination revealed



a) Elbow plot of CB8+CB3 synthetic mixture with 3% doublets shows a clear preference for three clusters rather than the expected two. **b)** Shows the PCA of the normalized cell by cluster log likelihood matrix (n=2716 cells) showing three distinct genotypes.

Fig. 2.15: Performance on low UMI counts



a) The synthetic mixture of 5 HipSci cell lines with 6% doublets and 5% ambient RNA with UMIs downsampled shows predominantly good clustering, but performance drops below 800 UMIs/cell. **b)** The clustering is consistently good with downsampled cells down to an average cell per cluster of 40. The cluster with the fewest cells in the 40 average cells per cluster had 20 cells.

2.3.6.2 Downsampling experiments for cells and UMIs

In order to explore the regime for which it is still possible to accurately demultiplex mixed samples, I used our synthetically mixed 5 HipSci samples and downsampled UMIs (figure 2.15a) and cell (figure 2.15 b) and report the ARI versus the ground truth. I found that while overall clustering remains good, cell assignment accuracy decreases below 800 median UMI per cell and that accuracy remains high down to an average of 40 cells per cluster (see figure 2.15b).

2.4 Discussion

Here I have presented souporcell, a method for clustering scRNAseq cells by genotype using sparse mixture model clustering with explicit ambient RNA modeling. Our benchmarks show that souporcell can outperform all other currently available methods, including those that require genotypes *a priori*. Using more realistic and challenging test cases than previous studies, I show that souporcell is robust across a large range of parameters, and more so than any other currently available method. Moreover, souporcell is highly accurate for challenging datasets involving closely related maternal/fetal samples, and varying mixtures of *Plasmodium falciparum* strains. Limitations of souporcell include low signal to noise due to decreased UMI per cell and high numbers of donors causing increased local maxima. Due to the advantages that mixtures give to scRNAseq experiments in ameliorating batch effects, improving doublet detection, and allowing for ambient RNA estimation, souporcell enables donor multiplexing designs to be used more easily than was previously possible, including in situations when no WGS or genotyping data are available. In addition to reducing cost and allowing for more complex and robust experimental designs, souporcell also enables valuable genotype information to be extracted and ambient RNA estimation at no additional cost.

I believe that mixing individuals will become a more and more popular experimental design due to the advantages it brings. When using genetic variation as the signal for demultiplexing and doublet calling, some general guidelines should be followed. While we have made great advances in the ability to demultiplex mixtures with a large number of individuals, it remains the case that clustering and doublet calling difficulty increases with the number of individuals especially in low coverage datasets. In addition to this, as the number of individuals increases, the number of cells per individual decreases assuming total cell loading remains constant. If a small enough number of cells is sampled from an individual, minority cell types of interest may not be sampled. Increasing the cell loading (superloading) is not recommended much beyond 10k cells recovered as the increase to multiplets it creates makes clustering and doublet detection more difficult and the errors it introduces may corrupt downstream analysis. For this reason, a reasonable guideline might be to limit the number of individuals mixed to around 10 thus sampling 1000 cells on average from each individual with 10k cells recovered resulting in a 10% doublet rate. It should also be noted that sequencing the sample more deeply improves the performance of these methods as long as more UMIs are being sampled. Due to cost, many single cell experiments sample relatively few UMI (<4k, <2k sometimes) per cell. I recommend to sample at least 4k UMI/cell and performance continues to increase as more UMI are sampled. The more individuals per experiment, the more data it will take to accurately cluster and call doublets.

Further research should be done to evaluate demultiplexing and doublet calling on heterogeneous cell types. This should include evaluation of co-expression between cell types and samples with non-overlapping cell types. The maternal/fetal data contained different cell types for different individuals, but an evaluation of co-expression was not done. In the extreme case of non-overlapping expression patterns and non-overlapping cell types between individuals, clustering by genotype will not be possible. In practice, there are so called house keeping genes which are widely expressed by most cell types making this problem moot, but more analysis should be done to quantify this potential problem and its likelihood in various datasets.

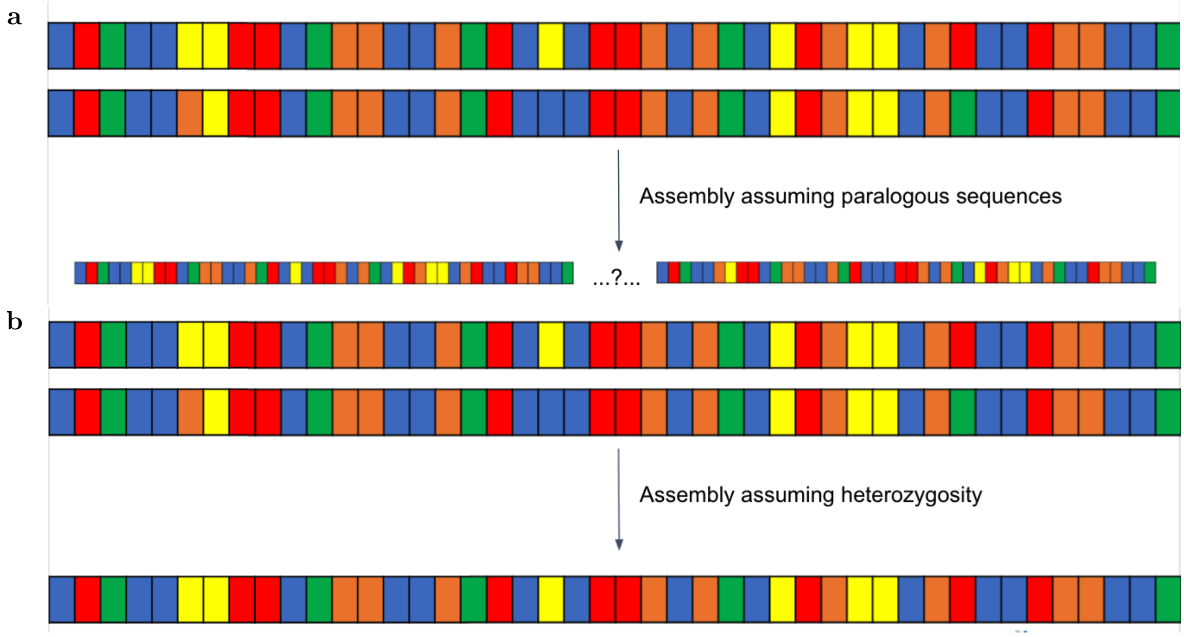
Chapter 3

High quality assembly of a single Mosquito

3.1 Background

Exciting efforts to sequence the diversity of life are building momentum[193] 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. Advances in long read sequencing over the past decade have revolutionized genome assembly and reference genome creation[85][109], but until recently the DNA requirements for these technologies were relatively high. This made long read sequencing of single individuals impossible for many small species due to the amount of DNA that can be extracted even when consuming the whole specimen. In the standard assembly process, when considering sequences which have inexact homology, one must decide whether the differences arose from errors, haplotype differences, or paralogous sequences. If it is determined that the differences are due to heterozygosity, an assembler would collapse the sequence. However, if the assembler decides the sequences are repeats and thus represent different locations (close or distal) in the genome, they should be assembled separately (see figure 3.1). As the haplotype differences increase, it reduces the assembler's ability to distinguish paralogous sequences from haplotype differences for higher divergent repeats. When one cannot distinguish these processes and no reads span the repeat (and if it is due to haplotype differences, no reads will span as the homology is highly likely to continue), the contig must end to avoid chimeric misassemblies. This results in fractious and error prone genome assemblies. One could, of course, pool multiple individuals together to meet the DNA requirements, but this has serious downsides. Using a pool of individuals increases the number of haplotypes being sequenced and increases the expected haplotype differences which reduces one's ability to distinguish paralogous sequences from haplotype variation. Moreover, the structural variation in the pool of haplotypes can cause further problems in assembly. These problems are accentuated in these small species that require pooled long read sequencing, because, while levels of heterozygosity within species vary widely across taxa, intraspecific genetic variation is often highest in small organisms[191].

Fig. 3.1: Assembly of inexact homologous sequences: heterozygosity vs paralogous sequences


a
Assembly assuming paralogous sequences
...?...
b
Assembly assuming heterozygosity

Inexact homologous sequences and how they would be assembled if the differences are due to **a)** paralogous sequences or **b)** heterozygous differences.

To address these problems, 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 with more similar haplotypes. This approach has varied in its success, for example working well for organisms that are easy to inbreed (e.g., many *Drosophila* species[77]), but less well for species that are difficult or impossible to inbreed (e.g., *Anopheles*[244]). 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 sequencing. 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[312].

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[219]. These four haplotypes will have recombined with each other in the cross, but recombinations are fairly rare and do not greatly increase the haplotype differences problems in assembly. Even this may run into problems though. For example, in species that mate multiple times and store sperm in a spermatheca as is the case in many diptera[62][124] it may be difficult to create a pure single cross from a wild caught individual.

However, for an initiative like the Earth BioGenome Project[193] 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 preparation[1][164], 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 still 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.

In this chapter we discuss the process of making the first high quality assembly of a single mosquito and assess its quality and completeness. We first discuss the methods used for high molecular weight (HMW) DNA extraction and resulting length profiles. The extraction was done at the Sanger Institute, but the sequencing was done in California. A DNA fragment size profile is included to show the DNA length degradation from

[1] My friend and former coworker, Brendan Galvin, was the person at PacBio who made these library preparation improvements.

transit. We then discuss the new low input library prep and the sequencing used. We outline briefly the curation steps taken that resulted in changes to the assembly before going through each analysis in detail. Assembly quality is then assessed first through comparison of contiguity of the assembly and curated assembly against the current gold standard reference genome (Agam4 PEST). Next completeness and assembly duplication are inspected via comparing to known orthologous gene sets with BUSCO (Benchmarking Universal Single-Copy Orthologs). Finally, I do a series of genome comparisons to the PEST reference. In this, I am able to identify and correct a misassembly and uncover significant remaining duplicated haplotype assembly sequence and its cause. In this analysis, I am also able to find many improvements our assembly makes over the PEST assembly including placing of previously unplaced genes in their chromosomal context. We also dramatically reduce the number of assembly gaps. I found significant evidence of collapsed complex repeats in PEST that have been accurately expanded in our assembly. I then found an order-and-orientation error in the PEST reference. Finally, I show the contig coverages of the PEST reference aligned to the PacBio assembly and how much of the UNKN contigs are likely haplotigs and visually show the placement of other UNKN sequence.

This work was done over two years ago now, and the field is rapidly evolving. Many of the procedures described in this chapter have become common practice and have been further improved with the advent of the HiFi data type from PacBio. Today, even higher quality genomes are being produced on a regular basis. This work represents the first, but not the last or best genome assembly of a small organism.

The genome we use for comparison was built using bacterial artificial chromosomes (BACs) and Sanger sequencing, the same basic technology initially used to create the human reference genome, which is highly accurate but extremely labor intensive and expensive. Our assembly allows for the use of a single individual, is relatively cheap, and is more accurate and complete than the previous gold standard *Anopheles* genome. With some additional data, or by scaffolding against the PEST reference as shown in this chapter, it would also be more contiguous with fewer gaps.

3.2 DNA Isolation

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

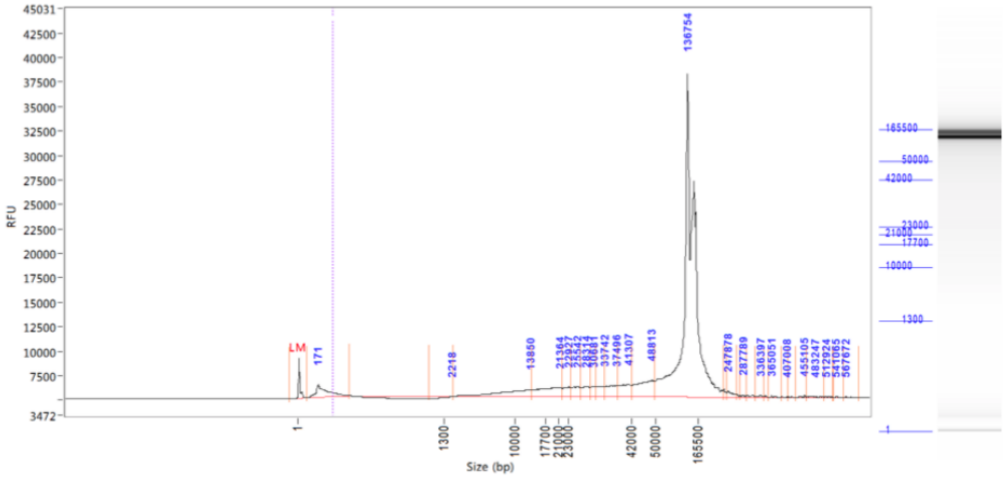
High molecular weight (HMW) DNA was isolated from a single *Anopheles coluzzii* female from the Ngousso colony. This colony was created in 2006 from the broods of approximately 100 wild-caught pure *Anopheles coluzzii* females in Cameroon (pers. comm. Anna Cohuet). Although the colony has been typically 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 homogenization and incubation; incubation time was shortened to 2 h; solutions were mixed by gently flicking the tube rather than pipetting to reduce shearing and maximize extracted DNA length; 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 to examine the molecular weight of the resulting DNA. This revealed a relatively sharp band at 150 kb (figure 3.2). 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 3.3), 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, and 100 ng from the 250 ng total was used for library preparation.

3.3 Library prep and Sequencing

Library prep and sequencing were performed by scientists at PacBio.

A SMRTbell library was constructed using an early access version of SMRTbell Express Prep kit v2.0 from Pacific Biosciences (PacBio). Because the genomic DNA was already fragmented with the majority of DNA fragments above 20 kb (figure 3.3), the sequencing library preparation protocol was modified to exclude an initial shearing step, which facilitated the use of lower input amounts, as shearing and clean up steps

Fig. 3.2: *Anopheles coluzzii* single mosquito HMW DNA extraction

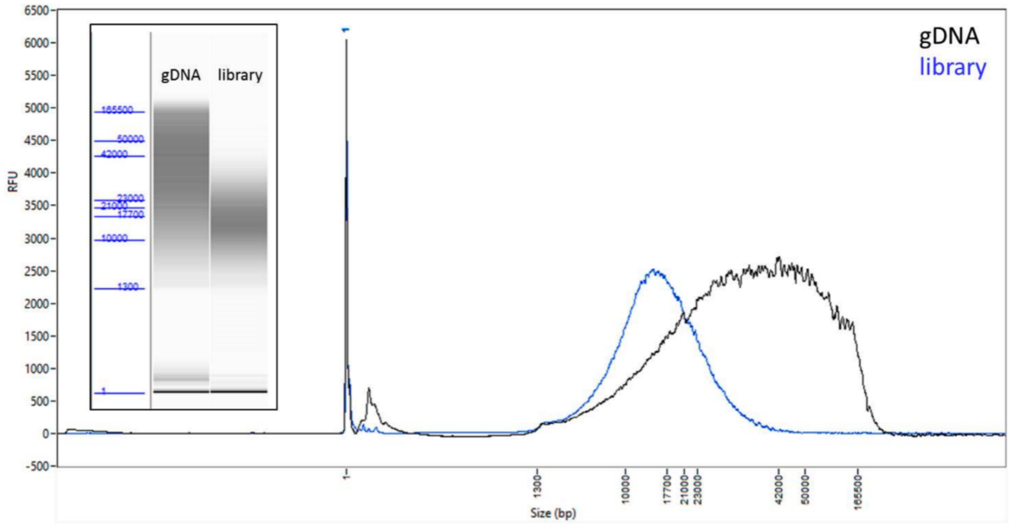


Femto Pulse evaluation of the Modified MagAttract DNA extraction prior to shipment to California.

typically lead to loss of DNA material. After following the Express template preparation protocol, the final clean up step was simplified to just two AMPure purification steps to remove unligated adapters and very short DNA fragments. The size and concentration of the final library (figure 3.3) were assessed using the FEMTO Pulse and the Qubit Fluorometer and Qubit dsDNA HS reagents Assay kit, respectively. This resulted in a final library with a size distribution peak around 15 kb (figure 3.3).

Sequencing primer v4 and Sequel DNA Polymerase 3.0 were annealed and bound, respectively, to the SMRTbell library. The library was then sequenced on the Sequel System with Sequel Sequencing Kit 3.0. 1200 minute movie with 120 minute pre-extension and Software v6.0. A total of three SMRT cells were run generating on average 24.2 Gb of data per SMRT Cell, with average insert lengths of 8.1 kb (insert length N50 $\approx$13 kb, table 3.1). This is double the standard exposure time allowing for more data out of the same sample. However, extending exposure times has diminishing returns. The overall library yield was 59%, which would have allowed for the sequencing of at least 8 SMRT Cells, thereby potentially allowing for genome sizes 2-3 times larger or organisms that yield 2-3 times less DNA in extraction than studied here in conjunction with this protocol.

Fig. 3.3: *Anopheles coluzzii* input and resulting library DNA lengths



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

Table 3.1: Run statistics for Sequel SMRT Cells.

Loading concentration	Gb/cell	Mean Polymerase Read Length	N50 Polymerase Read Length	Mean Subread Length	N50 Subread Length
5 pM	24.1	40290	116615	8185	12978
5 pM	23.6	40077	114807	8254	13132
6 pM	25.0	47177	122898	8012	12751

3.4 Assembly

The assembly was run by Sarah Kingan, Senior Scientist at PacBio. My main role was in quality assessment and comparative genomics.

The genome was assembled using FALCON-Unzip, a diploid assembler that captures haplotype variation in the sample[52]. A single subread per zero-mode waveguide (ZMW) was used for a total of 12.8 Gb of sequence from three SMRT Cells, or 48-fold coverage of the 266 Mb genome. Subreads longer than 4559 bp were designated as seed reads and used as template sequences for preassembly/error correction. A total of 8.1 Gb of preassembled reads was generated (30-fold coverage). After assembly and haplotype separation by FALCON-Unzip, two rounds of polishing were performed to increase the consensus sequence quality of the assembly, aligning the PacBio data to the contigs and computing consensus using the Arrow consensus caller[51]. The first round of polishing was part of the FALCON-Unzip workflow and used a single read per ZMW that was assigned to a haplotype. The second round of polishing was performed in SMRT Link v 6.0.0.43878, concatenating primary contigs and haplotigs into a single reference and aligning all subreads longer than 1000 bp (including multiple subreads from a single sequence read, mean coverage 184-fold) before performing genomic consensus calling. The alignments (BAM files) produced during the two rounds of polishing were used to assess confidence in the contig assembly in regions with rearrangements relative to the AgamP4 PEST assembly for *Anopheles gambiae* (GenBank assembly accession GCA_000005575.2)[129][294]. We referred to the first round of polishing as using unique subreads and the second round as using all subreads.

We explored the performance as a function of the number of SMRT Cells used for the assembly (table 3.2), and found that while a single SMRT Cell was insufficient to result in high-quality assembly, data from two or three SMRT Cells generated a highly contiguous assembly of the correct genome size. We proceeded with the three-cell assembly for all subsequent analyses because it gave the most contiguous and complete assembly results.

3.5 Curation

The contigs were screened by the Sanger Institute and NCBI to identify contaminants and mitochondrial sequence[132]. Windowmasker was used to mask repeats and the MegaBLAST algorithm was run (with parameter settings: -task megablast

Table 3.2: Assembly quality vs the amount of data used.

	1 SMRT cell	2 SMRT cells	3 SMRT cells
Total bases (Gb)	23.6	48.5	72.7
Total unique bases (Gb)	4.46	8.31	12.8
Unique coverage	17x	31x	45x
Assembly size (Mb)	150	265	271
Number of contigs	3,290	815	580
Contig N50 (Mb)	0.066	1.5	3.5

Statistics for *Anopheles coluzzii de novo* genome assemblies as a function of the number of SMRT Cells used for the assembly. One cell failed to assembly the whole genome. Two and three cells assembled the majority of the genome but quality improved with 3 cells.

-word_size 28 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -dust yes -evalue 0.0001 -min_raw_gapped_score 100 -penalty 5 -perc_identity 98.0 -soft_masking true -outfmt 7) on the masked genome versus all complete bacterial genomes to find hits with greater than 98% homology[234][47]. One contig (#20) was identified as a complete 4.24 Mb bacterial genome, closely related to *Elizabethkingia anophelis*, which is a common gut microbe in *Anopheles* mosquitoes[179][44]. It was separated from the mosquito assembly and submitted to NCBI separately. We also identified two contigs of mitochondrial origin that each contained multiple copies of the circular chromosome. Full length copies of the mitochondrial chromosome in the higher quality contig differed by only a single base and the consensus sequence was reported as the mitochondrial genome. One of these copies was discarded.

In addition, I screened the primary assembly for duplicate haplotypes using Purge Haplotigs[273] with default parameters and coverage thresholds of 20, 150, and 700. While FALCON-Unzip resolved haplotypes over 30% of the genome, 110 genes appeared as duplicated copies in the BUSCO analysis, indicating that highly divergent haplotypes may be assembled as distinct primary contigs as has been observed in other mosquito genome assemblies[220][105]. The presence of duplicated haplotypes can result in erroneously low mapping qualities in resequencing studies and cause problems in downstream scaffolding. Using the Purge Haplotigs software[273], I identified 165 primary contigs totalling 10.6 Mb as likely alternate haplotypes, although there remains a possibility that some may be repeats. These contigs were transferred to the alternate haplotig set.

In the process of comparing the assembly to the PEST reference (described later), I found one large potential heterozygous interchromosomal rearrangement between 2L and 3R (see figure 3.5). Upon further exploration, this was not supported by any subreads mapping across the breakpoint (figure 3.5). The putative breakpoints were identified by aligning the PacBio contigs to PEST with minimap2 (asm5 setting)[197], and the start and end position of each aligned subread was determined using bedtools bamtobed[270]. This 4.9 Mb contig had no reads spanning the putative breakpoint when either unique or all subread alignments were examined and thus I designated this a chimeric misassembly, and split the contig into two.

3.6 Assembly quality assessment

Using the FALCON-Unzip assembler[52], the resulting primary *de novo* assembly consisted of 372 contigs totaling 266 Mb in length, with half of the assembly in contigs (contig N50) of 3.5 Mb or longer (table 3.3). FALCON-Unzip also generated 665 alternate haplotigs, representing regions of sufficient heterozygosity to allow for the separation of the maternal and paternal haplotypes. These additional phased haplotype sequences spanned a total of 78.5 Mb (i.e., 29% of the total genome size was separated into haplotypes), with a contig N50 of 223 kb (table 3.3).

Table 3.3: Assembly statistics

		Initial Assembly	**Curated Assembly**	**PEST reference**
Primary Assembly	Size (Mb)	266	251	224
	Number Contigs	372	206	27,063
	Contig N50 (Mb)	3.52	3.47	0.025
Alternate Haplotigs	Size (Mb)	78.5	89.2	unresolved
	Number Contigs	665	830	N/A
	Contig N50 (Mb)	0.22	0.199	N/A

3.6.1 BUSCO analysis: completeness and duplication/haplotig retainment

To evaluate genome completeness and sequence accuracy of the currated assembly, we performed alignment analyses to a set of conserved genes. Using the diptera set of the BUSCO (Benchmarking Universal Single-Copy Orthologs) gene collection, we observed 98% of the 2800 genes were complete and >95% occurred as single copies (table 3.4). By comparison, the previous assembly had 87.5% complete BUSCO alignments, indicating that a fraction of the genome was missing in that assembly. The percentage of duplicated genes was reduced from 3.9% to 2.4% after curation. Additional analyses are required to distinguish true gene duplication events from incomplete purging of duplicated haplotypes (see discussion below and figure 3.6). In addition, we evaluated assembly completeness against a curated set of genes (AgamP4.10 gene set) from the *Anopheles gambiae* PEST reference, using a previously described script[175]. We aligned to the primary assembly a closely related species gene set (the most recent *Anopheles gambiae* (AgamP4.10) gene set), resulting in 14,972 alignments (99.5%) and an average alignment length of 96.6%, and with >96% of alignments showing no frame shift-inducing indels.

Table 3.4: BUSCO analysis.

Gene count (%)	**Initial**	**Curated**	**PEST**
	Assembly	**Assembly**	**Reference**
Complete	2745 (98.0)	2747 (98.1)	2448 (87.5)
Complete Single Copy	2635 (94.1)	2680 (95.7)	2446 (87.4)
Complete Duplicated	110 (3.9)	67 (2.4)	3 (0.1)
Fragmented	25 (0.9)	25 (0.9)	190 (6.8)
Missing	29 (1.1)	28 (1.0)	160 (5.7)
Total	2799 (100)	2799 (100)	2799 (100)

Analysis of single copy conserved genes using BUSCO v3.0.2 and the diptera gene set. Initial assembly: primary contigs from the 3-cell *de novo* FALCON-Unzip assembly. Curated assembly: Primary contigs after removal of bacterial contaminants and duplicated haplotypes. Previous reference from[129] GCA_000150765.1.

3.6.2 Comparison to *Anopheles gambiae* PEST reference

The *Anopheles gambiae* genome, published in 2002, was created using BACs and Sanger sequencing[129]. Further work over the years to order and orient contigs improved this reference[294][295] and to date, AgamP4 (https://vectorbase.org/organisms/anopheles-gambiae/pest/agamp4) remains the highest quality Anopheles genome among the 21 that have now been sequenced[245]. However, there are many problems with this reference genome. AgamP4 PEST still has 6302 gaps of Ns in the primary chromosome scaffolds ranging from 20 bases to 36 kb, including 55 gaps of 10 kb that the AGP (A Golden Path) file on Vectorbase annotates as contig endings. The AgamP4 genome was generated from a lab strain known as PEST (Pink Eye STandard) that is long deceased and also was an accidental mixture of two incipient species, previously known as ‘M’ and ‘S’. To address this, the genomes of pure ‘M’ and ‘S’ from new colonies established in Mali were sequenced using only Sanger sequencing[189]. Since then, the 'S' form has retained the name *Anopheles gambiae* sensu stricto, and the 'M' form has acquired species status and a new name, *Anopheles coluzzii*[53]. It is important to note that while these species show assortative mating, they can hybridize in nature and their hybrids are fully fertile and viable[5]. Given this fact, and the fact that both pure species assemblies remain highly fragmented, I compared our assembly to the best available *Anopheles gambiae* genome (i.e., AgamP4 PEST) to evaluate contiguity and to help order and orient the contigs.

To assess the quality of contig assembly and concordance with existing assemblies, the curated primary contigs were aligned to the PEST *Anopheles gambiae* reference genome [129][294] using minimap2 with the map-pb settings[197]. For the purpose of comparison, contigs were ordered and oriented according to their median alignment position and majority alignment orientation on the chromosome to which they had the most aligned bases. A python script was used in conjunction with ggplot using geom_segments to generate alignment plots. This software is an alternative to the commonly used nucmer/mummer[181] and is available at https://github.com/wheaton5/assembly_comparison_scripts. One important difference is that these only show the single best alignment for a given span of contig sequence and is not a dotplot which would show all similar sequences above some threshold.

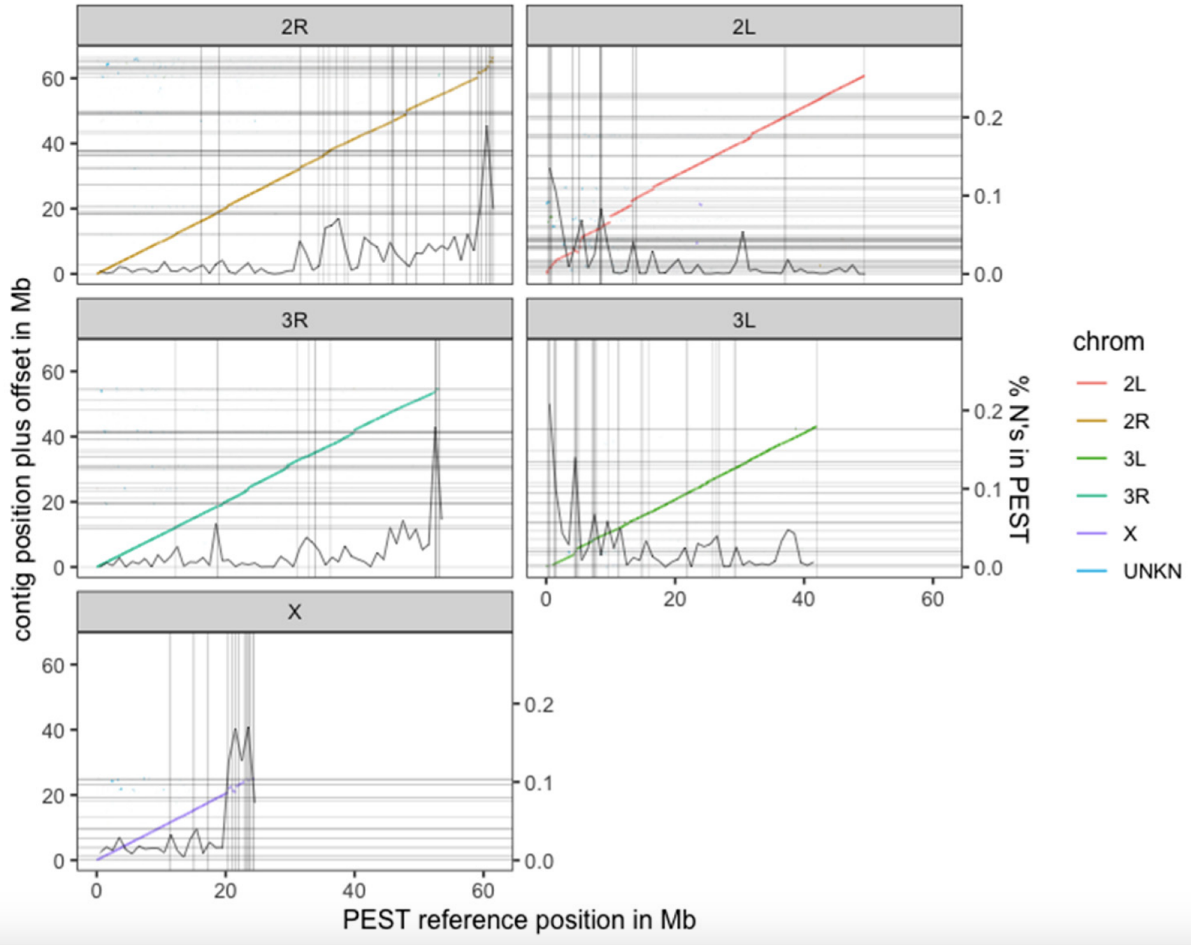
The new PacBio assembly is highly concordant with the AgamP4 PEST reference over the entire genome, allowing the placement of the long PacBio contigs into chromosomal contexts (see figure 3.4). In addition, the high contiguity of the PacBio contigs allows for the resolution of many gaps in the chromosomal PEST contigs. Note that the only gaps in the PacBio assembly are at contig ends, whereas there are many gaps in PEST that are not annotated as contig breaks so the percent Ns per megabase of PEST is overlaid in the graphs in figure 3.4.

3.6.3 Identification and correction of misassembly

Large regions (>200 kb) of discordant alignment of the contigs to the PEST reference were inspected further. Discordant alignments were categorized into one of three cases. 1. Large portions of a single contig aligned discordantly (e.g., to multiple PEST reference chromosomes). 2. Large regions in PEST where multiple assembly contigs aligned to the same reference region. 3. Large region in PEST where assembly contigs did not align.

First, I considered discordant alignments where large portions of a single contig aligned in very different locations of the PEST reference. Using the alignment plots

Fig. 3.4: Comparison of the assembly with the PEST reference



Alignment of the curated PacBio contigs to the AgamP4 PEST reference. 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, but there may be gaps between the PacBio assembly contigs.

as described above, I colored each alignment by that contig's primary chromosome. Immediately one cross-chromosome contig alignment stuck out (see figure 3.5). I then evaluated the evidence for the assembly join at this breakpoint by aligning the subreads to the assembly and inspected the breakpoint region in IGV (figure 3.5). I found a repeat sequence that reads from each end would align into, but found no reads that spanned the repeat. This indicated to us that this was a chimeric misassembly, and I split the contig into two.

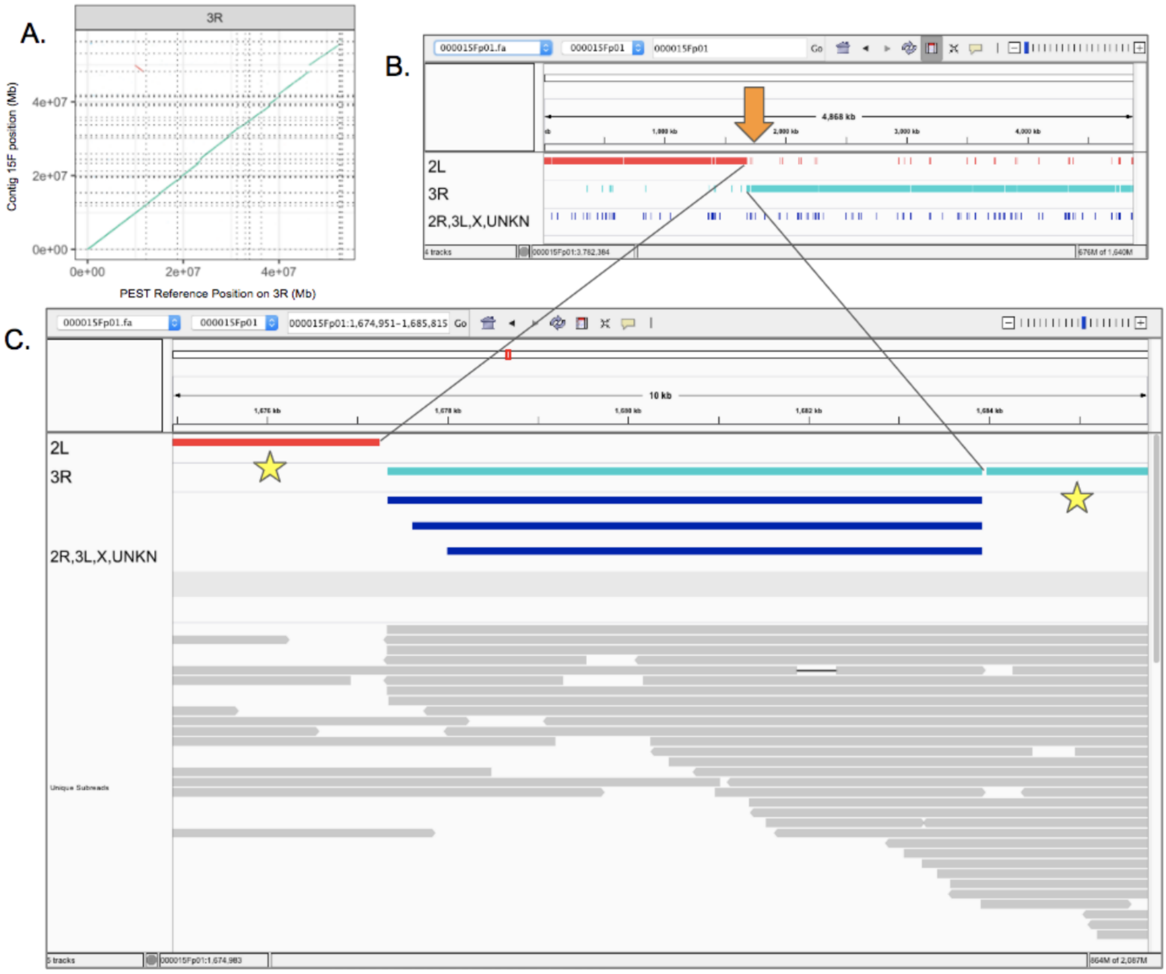
I also noted many smaller cross-chromosome alignments between contigs which primarily aligned to one of the five of PEST's chromosom-arm scaffolds and the UNKN (unknown) scaffold (not a true scaffold, just a collection of unplaced contigs). This is discussed further in section 3.6.8 by using these alignments to place genes and other genomic sequence in its proper chromosomal context which were previously unplaced.

3.6.4 Remaining haplotig sequence on ends of contigs

Next, large discordant regions where multiple contigs align to the same region in PEST were identified and evaluated. I found these regions by running samtools depth on the contig-PEST alignments, compressing to a bed file of contiguous regions of the same coverage, and then plotting to visually see where large sections are discordant (see figure 3.6). Using this, I found several very large segments of coverage two. Further inspection was done by zooming in to those regions specifically in the alignment plots. I observed that ends of contigs were aligned to the same position in the PEST genome. I then mapped the subreads to the assembly and assessed coverage in these regions. As expected if these were haplotig regions, the coverage was roughly half in the overlapping alignment regions. This clearly revealed that contigs were assembling the two haplotypes separately and were not removed by purge haplotigs. This is because purge haplotigs looks for contigs that are fully contained by another contig and it keeps the longer contig and removes the shorter haplotig. This results in regions where the assembly has assembled both haplotypes separately but one contig does not fully contain the other contig, the haplotig sequence remains (see figure 3.6). This realization spawned another project in our lab to improve haplotig purging by combining coverage of reads mapped to the assembly with sequence similarity to identify and remove haplotig sequence even when not fully contained by another contig[112] improving on the two previously available methods[273][135].

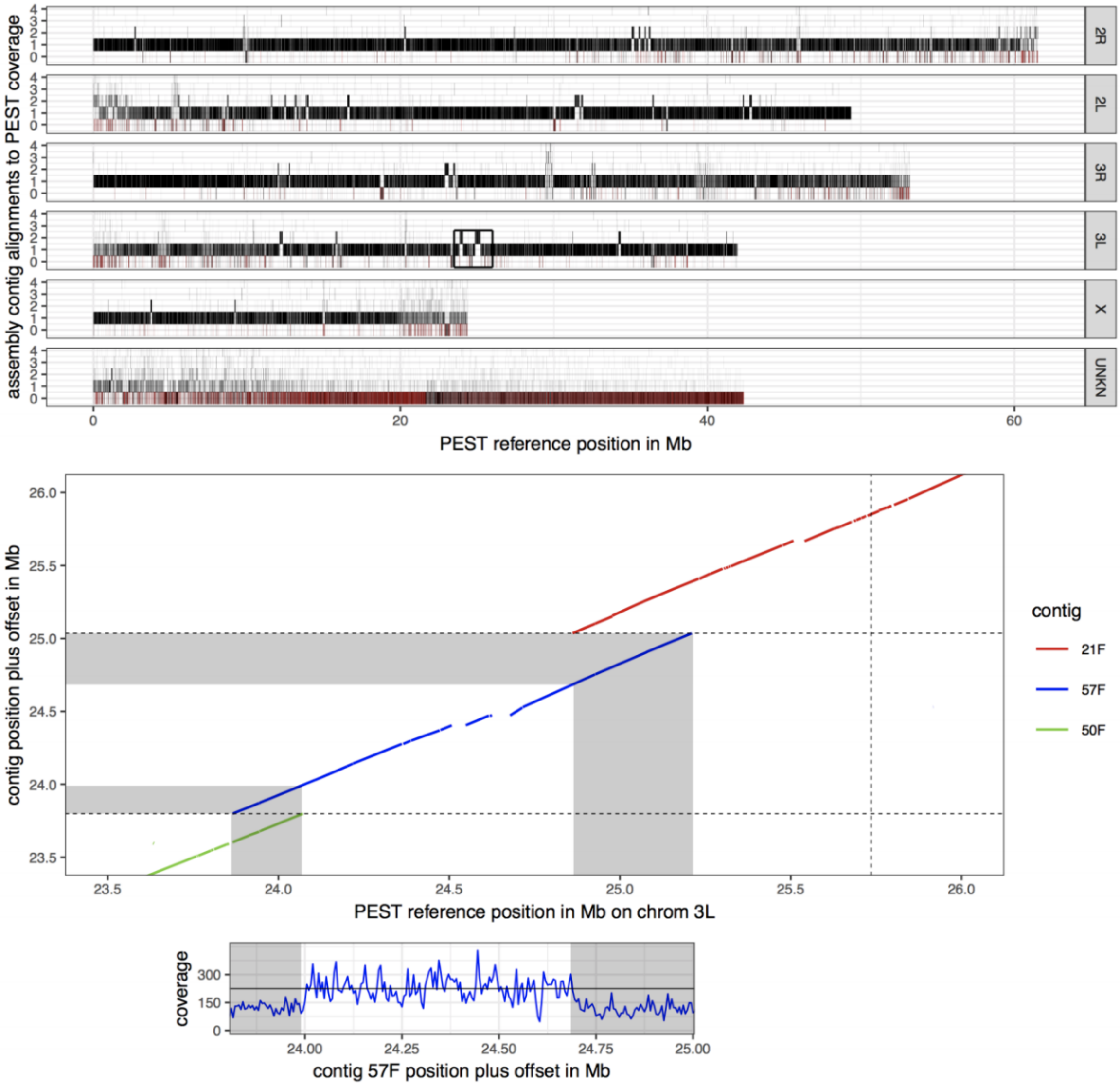
And finally, I considered large discordant regions in which no assembly contigs align to the PEST reference. In general this is rare in the chromosomal scaffolds. Most of the

Fig. 3.5: Chimeric assembly



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*.

Fig. 3.6: Evidence of remaining haplotig contig ends.



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 Ns 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).

zero contig-coverage areas of the contig on PEST alignment file are location in which the PEST contains Ns. And the large majority of zero contig coverage areas are in the UNKN scaffold. I explored the largest of the zero contig coverage region in the chromosomal PEST scaffolds (figure 3.7). I saw that this region is flanked by sections of Ns and that very few PacBio subreads map to the non N regions between indicating that this sequence may be low quality, possible derived from contamination, or be a biological difference (large insertion in PEST) between the two species.

3.6.5 Expansion of previously collapsed repeat

The new PacBio assembly makes many improvements when compared to the PEST assembly. For example, a single contig from the new PacBio assembly expanded a tandem repeat region on chromosome 2L that in PEST was collapsed, while also filling in many Ns (gaps) in PEST, and also spanning a break between PEST scaffolds set to 10,000 Ns (see figure 3.8).

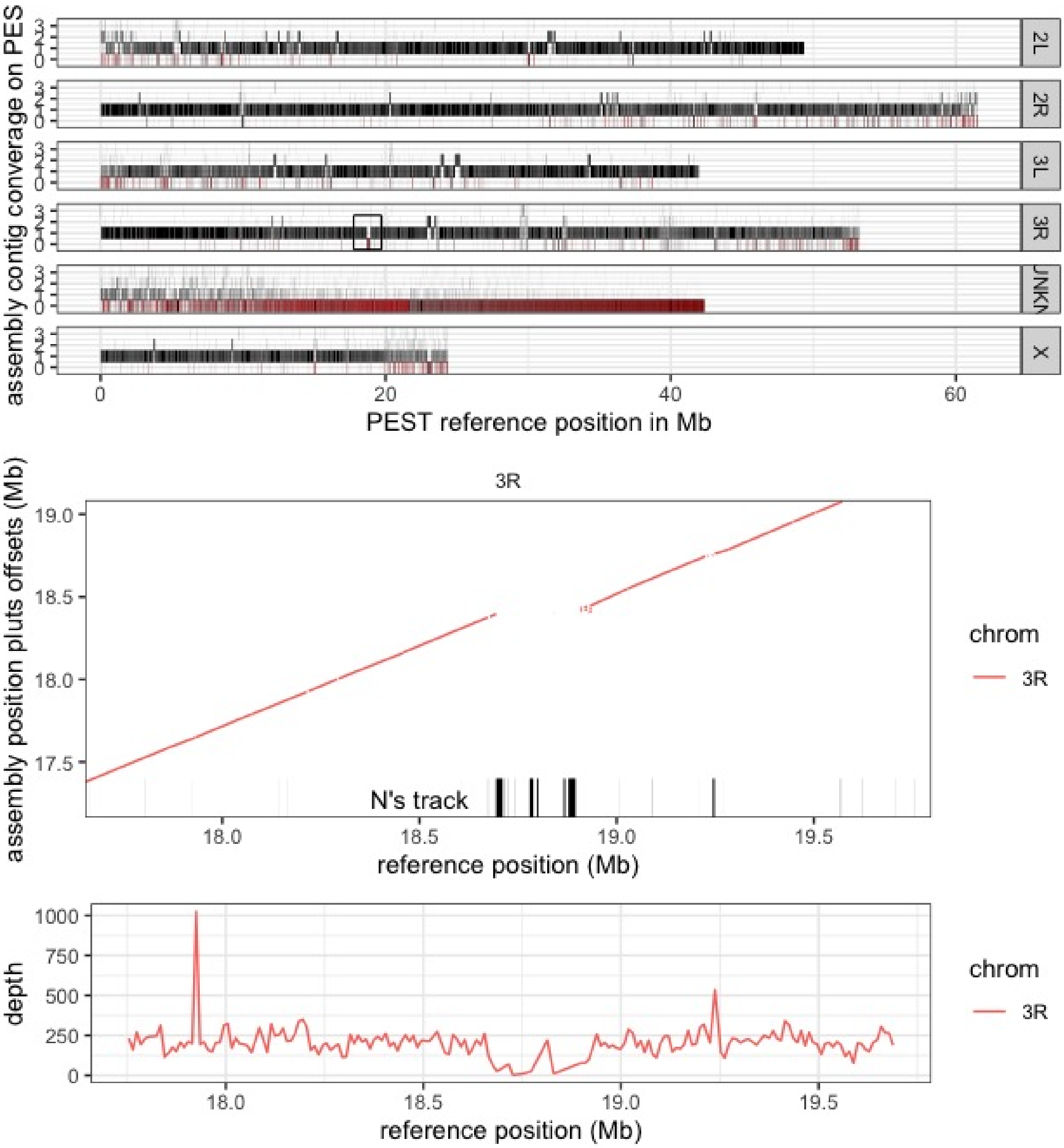
3.6.6 Corrected order and orientation vs PEST scaffolding

I also identified several potential rearrangements in the 20-22 Mb region of the X chromosome (see figure 3.9). PEST has contig breaks at the putative breakpoints relative to the assembly, however, given that a single PacBio contig spans the full region and that potential breakpoints relative to PEST are supported by multiple reads, the most likely explanation is an order and orientation issue in PEST, perhaps combined with a potential inversion difference between *Anopheles coluzzii* and the PEST reference. In addition, the contig contains a relatively large region (380 kb in total) of PacBio sequence corresponding to several pieces in the UNKN section of PEST that can now be assigned to the X chromosome.

3.6.7 Identification of some UNKN PEST sequence as haplotigs

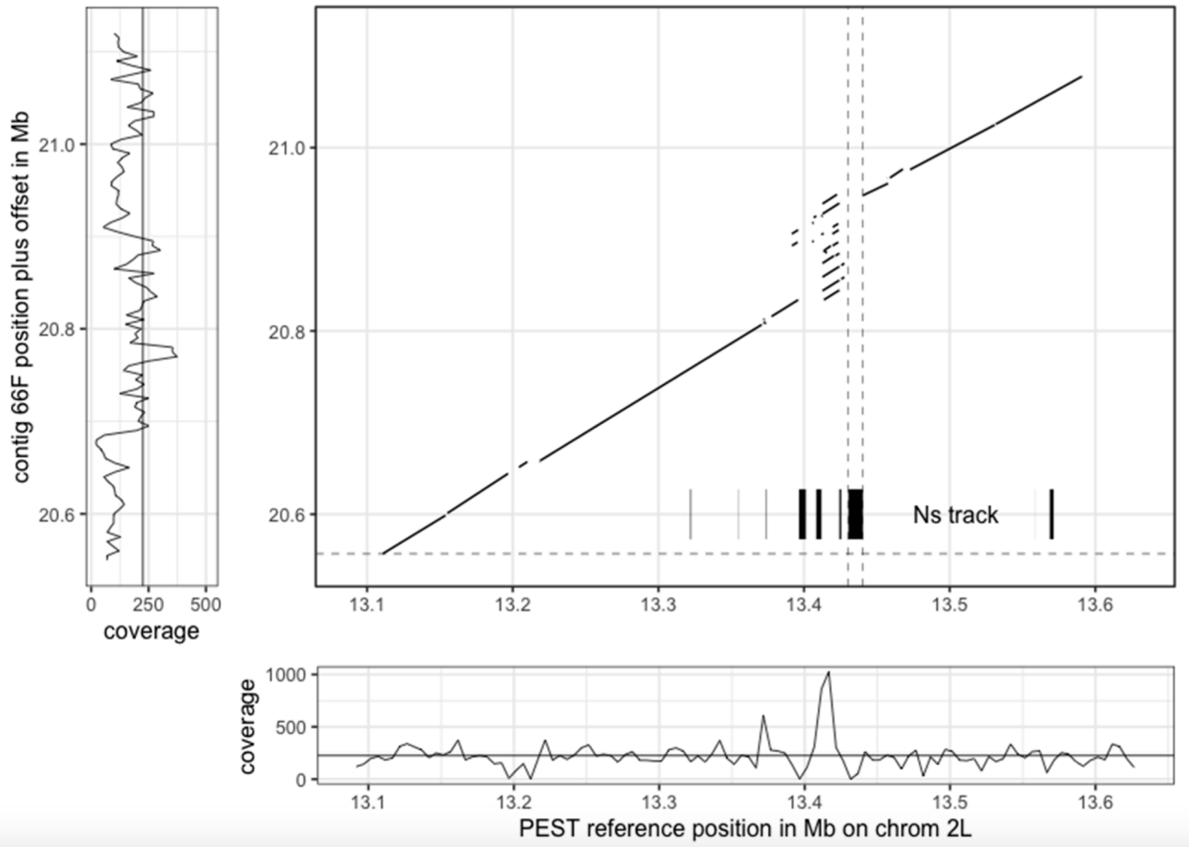
The PEST annotation also retains a large bin of unplaced contigs (27.3 Mb excluding Ns) designated as the UNKN (unknown) chromosome. Previously, I mapped either assembly contigs onto the PEST reference or subreads against the assembly. Now I show the reverse of the former and map PEST contigs onto the assembly. If an UNKN contig alignment overlaps a chromosomal contig alignment versus the assembly (both with

Fig. 3.7: No contig coverage region



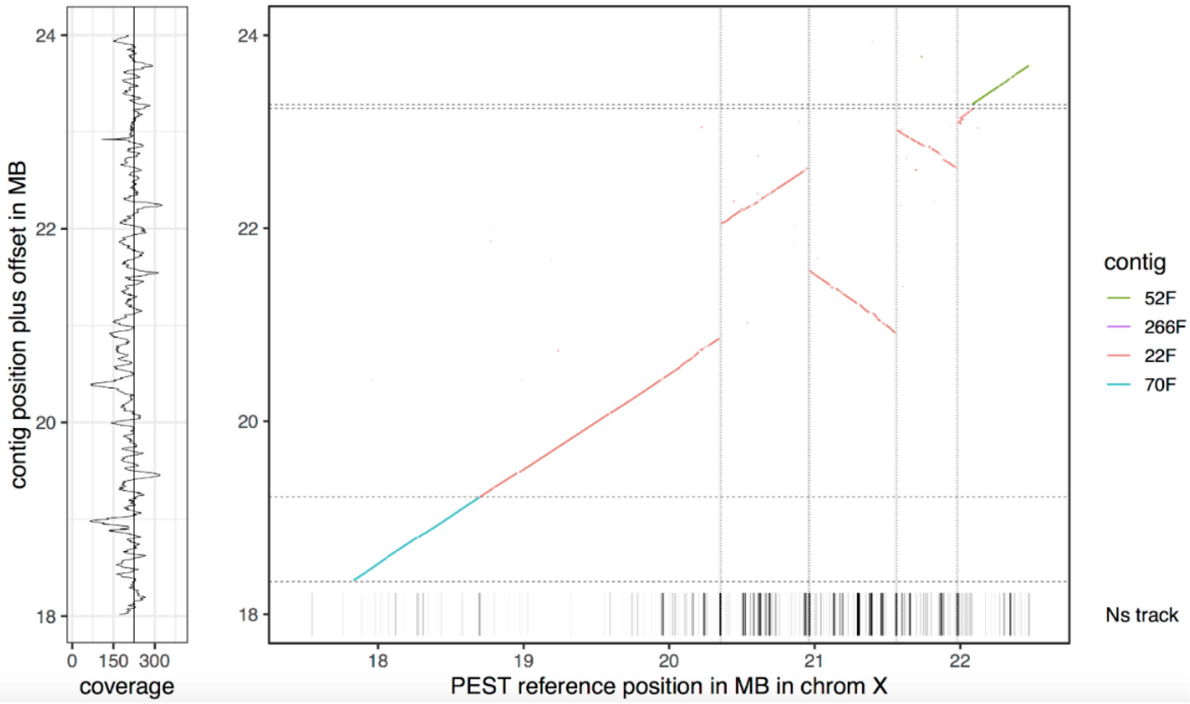
Top: outlines large region with zero coverage in chromosomal scaffold 3R on PEST reference. Middle: Zoom in of alignment plot in that region with Ns track showing regions of Ns flanking this sequence with another section of Ns in the middle. Bottom: shows subread depth when aligned to the PEST reference showing decreased mapping in this region.

Fig. 3.8: Example of expansion of previously collapsed repeat



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.

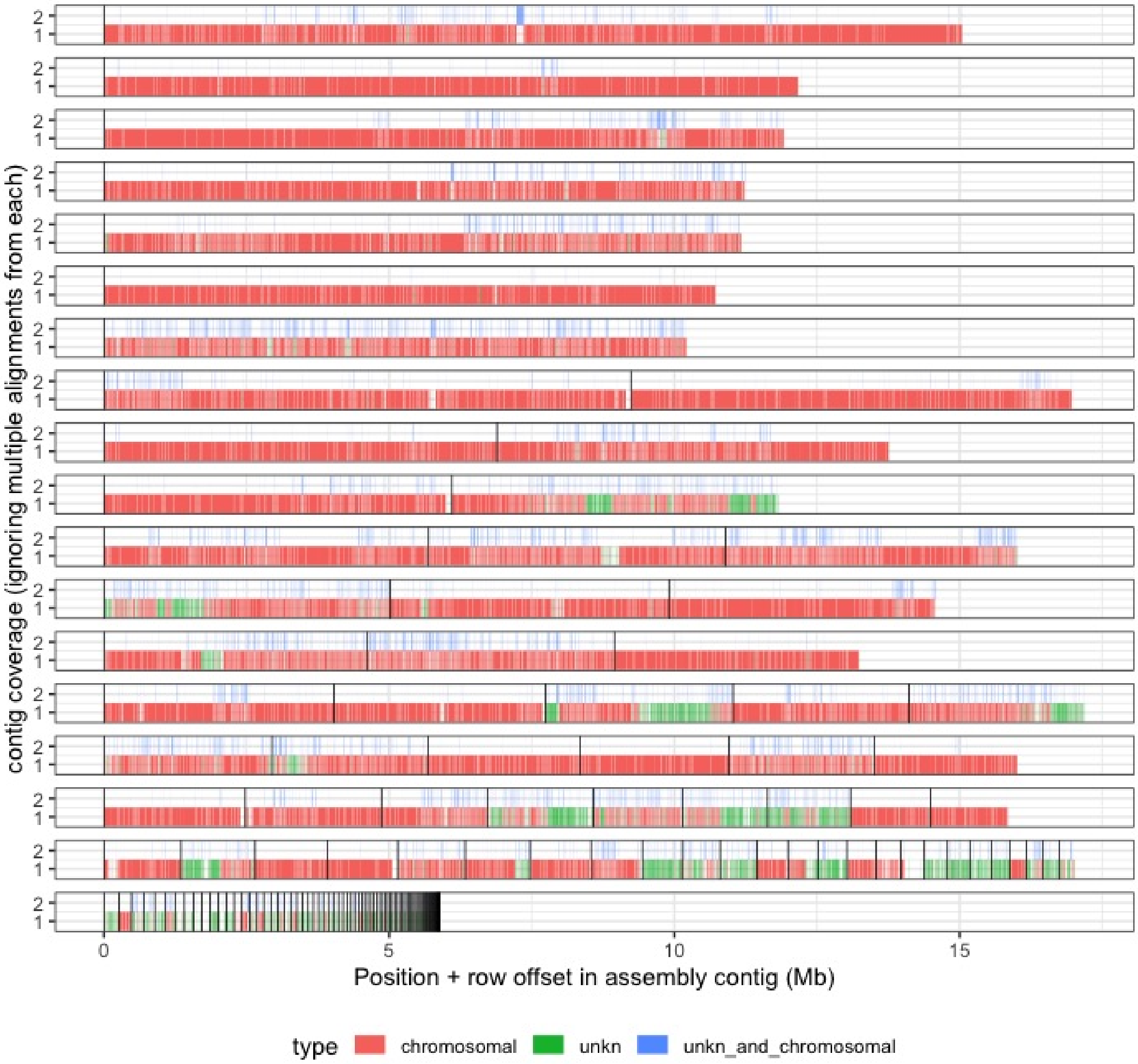
Fig. 3.9: Resolved order and orientation error in PEST scaffolding



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.

mapping quality (mapq) 60), it is likely to be a haplotig in the UNKN (see figure 3.10). In total, I find that 7.27 Mb are haplotigs (i.e., also have high quality PEST chromosomal alignments to the same location in the assembly).

Fig. 3.10: UNKN placement and haplotig identification



Contig coverages (ignoring multiple alignments for each chromosomal and UNKN sequence) of PEST aligned to the curated assembly showing placement of previously unplaced sequence (green on contigs that also have red (chromosomal) alignments). I also note the locations where both UNKN and chromosomal sequence align to the same location in the curated assembly which are likely haplotigs in the PEST UNKN scaffold.

3.6.8 Placement of previously unplaced genes

In addition to the UNKN haplotig sequences, I found another 10.9 Mb of the alignments are newly placed sequence that do not overlap with PEST chromosomal alignments but are in contigs that have a large amount (>100kb) of chromosomal alignments meaning these contigs are confidently ordered and oriented in their chromosomal contexts. 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[164]), which together with their flanking sequence comprise 8.9 Mb of sequence. Altogether, this means that 32.6% of the UNKN chromosome is now placed in the genome and 26.6% is determined to be haplotigs, along with 90% of the genes that were contained within it. Much of the remaining sequence do not have high mapping quality subreads when aligning subreads to the PEST reference meaning they are either repeats or junk sequence.

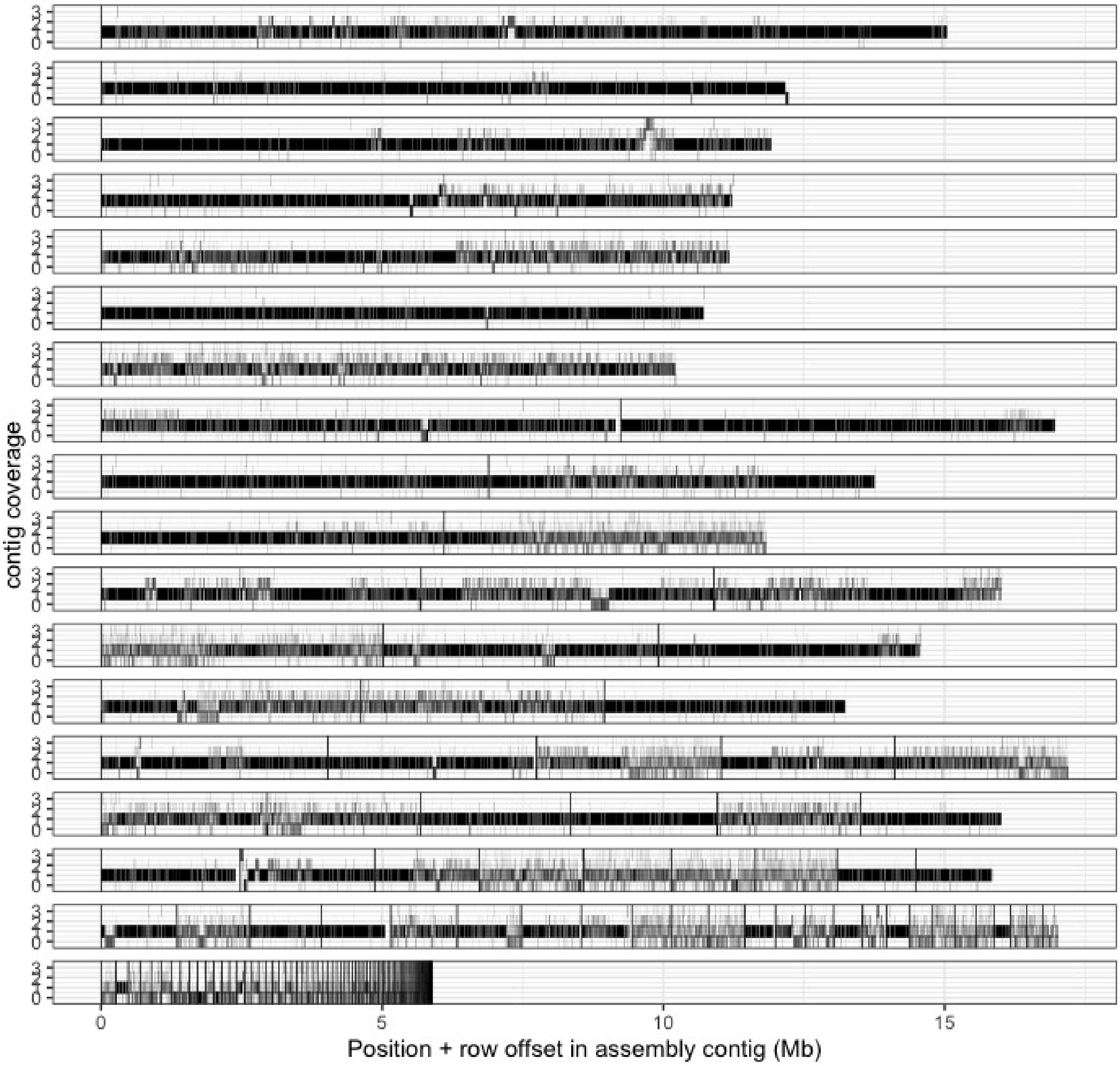
3.6.9 PEST contig coverage on PacBio curated assembly

In addition to looking for the two contig coverage areas and confirming that they primarily come from haplotigs in the UNKN scaffold, other abnormalities such as zero contig coverage regions are looked for. In figure 3.11 there is significant zero coverage regions. However, most of these are intermittently zero and one likely just indicating sequence divergence rather than an incomplete PEST reference. There are a few more solidly zero coverage regions, but they are not dramatically long (<200Kb). Still, the BUSCO 3.4 analysis shows that there are fewer complete genes and more fragmented and missing genes in the PEST reference as compared to the PacBio assembly. These regions likely account for some of that difference.

3.7 Discussion

Long-read PacBio sequencing has been utilized extensively to generate high-quality eukaryote *de novo* genome assemblies, but because of the relatively large DNA input requirements, it has not been used to its full potential for small organisms, requiring time-consuming inbreeding or pooling strategies to generate enough DNA for library preparation and sequencing. Here we present, to our knowledge, the first example of a high-quality *de novo* assembly from a single insect. This assembly, using only

Fig. 3.11: PEST contig coverage when aligned to PacBio curated assembly



Contig coverages of PEST aligned to the curated assembly. When comparing to figure 3.10, most of the two-coverage areas likely come from haplotigs in the UNKN scaffold. But there is some significant zero coverage areas as well.

one individual and one sequencing technology, exhibits a higher level of contiguity, completeness, accuracy, and degree of haplotype separation than any previous *Anopheles* assembly, demonstrating the impact of long reads on assembly statistics. While the assembly did not achieve independent full chromosomal scale assignment of contigs, its mega-base scale contiguity without gaps immediately provides insights into gene structure and larger-scale genomic architecture, such as promoters, enhancers, repeat elements, large-scale structural variation relative to other species, resolution of tandem repeats (figure 3.9), and many other aspects relative to functional and comparative genomics questions.

About a third of the genome for this diploid individual is haplotype-resolved and represented as two separate sequences for the two alleles, thereby providing additional information about the extent and structure of heterozygosity that was not available in previous assemblies, which have been constructed from many pooled individuals. In contrast with approaches requiring multiple individuals, the ability to generate high-quality genomes from single individuals greatly simplifies the assembly process and interpretation, and will allow far clearer lineage and evolutionary conclusions from the sequencing of members of different populations and species. Further, if parental samples are available, the recently developed trio binning assembly approach [174] can be used to further segregate alleles for a full haplotype-resolved assembly of both parental copies of the diploid offspring organism.

The assembly presented here provides an excellent foundation towards generating an improved chromosome-scale reference genome, using the previous PEST reference, scaffolding information from genetic maps, technologies such as Hi-C (e.g., [178]), or alignment of the contigs to closely related species' references. These approaches can also be used to highlight areas of potential improvements to the FALCON-Unzip assembler and to Purge Haplotigs, or other packages used to identify haplotypic contigs. As one example, we noticed in the context of the incomplete haplotype purging described above that some neighboring contig ends exhibited overlaps relative to the PEST reference (figure 3.6). The interpretation of such haplotype contig overlaps was corroborated by the observed halving of average sequencing depth over the regions of overlap. These methods could incorporate adjustments to try to account for haplotypic regions in the ends of contigs rather than complete contigs being fully haplotypic.

We noted the importance of the initial DNA size distribution in conjunction with this protocol. Since neither shearing prior to library construction nor size-selection thereafter were employed, the starting high-molecular weight DNA should contain fragments at greater than 20 kb on average, and without the significant presence of short (smaller than ≈5 kb) DNA fragments. Further research into suitable DNA extraction, storage and transportation methodologies is needed to fulfill these requirements for a broader spectrum of different species and environments, in order to allow for the preparation of suitable DNA samples from wild-caught samples originating in sometimes remote areas with limited sample preparation infrastructure.

The new workflow described here has now become standard procedure for creating high quality reference genomes for small organisms. And the more recent advent of HiFi data has made even higher quality genome assemblies of insects and other small organisms routine. This represents an important prerequisite in view of large-scale initiatives such as i5K and the Earth BioGenome Project [194][313]. In addition, other research areas with typically low DNA input regimes can benefit from the described new workflow, e.g., metagenomic community characterizations of small biofilms, DNA isolated from needle biopsy samples, minimization of amplification cycles for targeted or single-cell sequencing applications, and others.

Chapter 4

Haplotype phasing consistency as a signal for physical linkage in scaffolding and assembly

4.1 Background

Reference genomes have enabled a range of genomic analyses by providing prior knowledge of the sequence and a common coordinate system by which to compare multiple genomes[2][289]. Assembling reference genomes is complicated by repetitive sequences, heterozygosity, and sequencing errors. As discussed in Chapter 3, when an assembler encounters inexact homologous sequences, it must determine from which of these cases the sequences arose. If the assembler cannot distinguish between heterozygosity and repeats and no reads span the homologous sequence into unique regions, the contig must end to avoid assembling sequences from different regions together. Historically, reference genomes were created by sequencing an overlapping set of large haploid bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and fosmid clone libraries [186] (100-200kb, 100-1000kb, and 30-50kb respectively). These methods overcame much of the problem of resolving repeats because their length allowed them to read through all but the longest segmental duplications. However, because each of these BACs were sequenced with Sanger sequencing and assembled, they were still subject to problems in repeats longer than the Sanger reads (500bp-1kb) that were close enough to one another to occur in the same BAC clone. They overcame the problem of heterozygosity because the clone libraries were inherently haploid. But most importantly, these methods are far too costly to apply to many genomes. The human genome project took 13 years to complete and cost approximately 3 billion dollars[185]. Despite the labor intensive and costly process, the resulting reference genomes still had errors and were not complete[114].

More recently, the cost reductions and improved accuracy profiles[317][337] of long read sequencing[85][109], as well as the emergence of other long range genetic information technologies[354][202][150], have converged to make the production of high quality, cost effective reference genomes relatively straight forward. Efforts have begun on the Earth Biogenome Project (EBP)[194], a global project to sequence the entire diversity of multicellular eukaryotic life. In the UK, the Sanger Institute and partners have started to sequence 60,000 species from the British Isles in the Darwin Tree of Life (DToL) project. These projects aim to provide a scientific resource for the next generation of biological science, to serve as a store of data for environmental conservation, and to study evolution at a much broader scale than ever before. For the human genome, we continue to make progress through new technologies and computational methods[338][264][339]. The telomere to telomere project uses multiple technologies on a cell line derived from a haploid genome from the CHM13 hydatidiform mole to create the most complete human genome to date[249][230].

As discussed in Chapter 3, one of the primary remaining difficulties in assembling reference quality genomes is high levels of heterozygosity such as found in many of the non-model organisms included in the EBP and DToL projects. While Chapter 3 focused on going from a pool of individuals—and thus many haplotypes—to a single individual (and thus two haplotypes), this chapter focuses on the problems encountered with high levels of heterozygosity within an individual and the improvements that can be made computationally to alleviate these problems. For the newer technologies mentioned above, there are now assembly algorithms that deal with each data type[52][335][296] as well as combinations of multiple technologies[311][237][78][103]. These methods try to assemble both haplotypes together arriving at a haploid consensus[279][173] or assemble both haplotypes in a diploid assembly[178][335][48][97] and use one of these haplotypes as a reference genome.

As mentioned in Chapter 3, one method for dealing with the problem of heterozygosity is inbreeding organisms to a point of low heterozygosity[241], but this is not feasible for most organisms. Trio-sga used pedigree information in the assembly algorithm[212] but was built exclusively for short reads. Recently Koren et al. described trio binning which uses the kmer differences in a mother-father-child trio to separate long reads into their haplotype of origin prior to haploid assembly[174]. While this method is very effective, creating the necessary crosses would be practically infeasible and too costly for the vast number of reference genomes these large projects intend to produce. Assembly algorithms have also been developed to help overcome the problem of heterozygosity. HiCanu uses accurate long reads and applies sequence context masking techniques such as homopolymer compression to ignore regions that are more likely to contain errors. With these potential errors masked, HiCanu can require nearly identical sequence similarity to extend the assembly graph[248]. This results in the haplotypes being assembled separately in all but long stretches of homozygosity in a genome. HiCanu then relies on haplotig purging software purge dups[112] to remove one of the haplotype assemblies. This has the downside of not matching the haplotypes to make comparisons, but that could be done as an additional analysis step. This method does not explicitly phase the haplotypes and may have long phase switch errors in the contigs especially across long regions of homozygosity. In another method, DipAsm does a first pass assembly followed by haplotype phasing and separation of the reads by haplotype prior to haploid assembly[98]. And Hifiasm produces a diploid assembly with a diploid string graph algorithm[48]. While these methods have made much progress, heterozygosity still injects complexity over a haploid assembly process.

Despite the incredible advances made over the past several years in both sequencing technologies and assembly methods, we still cannot assemble whole chromosomes or chromosome arms with a single technology for most organisms. After assembly we are left with some level of fragmentation of chromosomes into contigs that we wish to scaffold together. While the PacBio HiFi technology has many beneficial properties including continuous, highly accurate reads, it does not produce long enough reads (10-25kb) to span all repeat or low sequence complexity regions in most genomes. 10x Genomics linked read technology, however, will create barcode linked short reads across much longer molecules (50kb+) with some molecules reaching well over 250kb[354]. Optical map technology in which the DNA is linearized and flourescent markers are attached to sequence specific loci via restriction digestion and optically inspected can give sparse data for pieces of DNA about as long as can be isolated with modern high molecular weight extraction methods[292]. And high-throughput chromatin conformation capture sequencing (Hi-C) data creates links between sequences physically located close to one another in the 3D nucleus. While there will be cross-chromosomal links, the large majority of links are intra-chromosomal and can create links of almost any length[63][201]. Each of these technologies have been used both to break misassemblies in contigs and scaffold contigs[121][256][78][104][216][22].

While high levels of heterozygosity make these problems harder for traditional methods, haplotype phasing consistency (the consistent separation of heterozygous alleles into distinct groups corresponding to the haplotype they belong to) can be used as a signal of physical linkage in both assembly and scaffolding and as a method to differentiate inexact repeats from haplotype differences. While the conventional thinking is that heterozygosity makes assembly more difficult, we turn this around and use the phasing consistent property of heterozygous sites as a powerful way to simplify and add statistical power to the physical linkage of sequences. This is made possible by the advent of long read and other long genetic range information technologies, as short reads do not span enough distance to consistently link heterozygous sites. In this chapter I present a toolkit for phasing, phasing aware assembly, and phasing aware scaffolding called phasstools (Phasing and Assembly tools). The code is open source and available at https://github.com/wheaton5/phasstools. The use cases of this package are *de novo* phased assembly, phasing aware scaffolding of that assembly, phasing aware scaffolding of an existing assembly, as well as direct read haplotype phasing.

First I outline phasing consistency using heterozygous single nucleotide polymorphism (SNP) kmer pairs as a mechanism for *de novo* haplotype phasing. This concept is critical for each of the methods I present. I then describe the phased assembly process. I first create an assembly graph using pairwise phasing consistencies, but then show several error modes that this process can encounter. This led me to need to use more than pairwise phasing consistency to overcome these problems. To do this, I need to consider kmers in a particular order. I show how the algorithm recruits each subsequent kmer pair, decides whether to add it to the graph, and how it updates future potential kmer pair's phasing consistency counts. After the phased assembly graph is created, it is used to separate HiFi reads into haplotype bins and haploid assemblies are created from each of those read bins. I show this process on the butterfly dataset of *Vanessa atalanta* and compare our assemblies to the HiCanu assembly of the same data. For diploid chromosomes, I show high concordance with the HiCanu assembly but with lower contiguity. While some problems remain, I believe this work shows the power of using phasing consistency as a signal for physical linkage in diploid chromosomes.

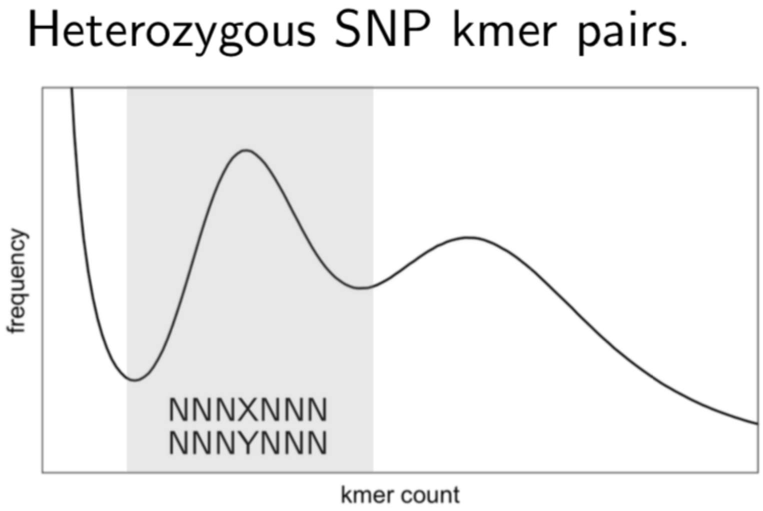
Next, I consider phasing aware scaffolding. If scaffolding an existing, unphased assembly, one first needs to phase it in order to take advantage of that information in the scaffolding process. I introduce an algorithm for haplotype phasing using sparse *Bernoulli* mixture model clustering—an algorithm very similar to the clustering algorithm used in Chapter 2. This algorithm has several benefits including being robust to non-heterozygous input variants and being able to correct those genotypes as well as naturally extending to polyploid genomes. I show promising results for this phasing algorithm, but problems remain if one does not have a large amount (300x) of Hi-C data. I have plans to address this short coming, but in the mean time I created another phasing algorithm that could be used for scaffolding immediately. I show results for both phasing algorithms on a dataset from the DToL project from the butterfly *Vanessa atalanta*.

Next I consider haploid sex contig detection, because these contigs will need to be treated differently in the steps going forward. I then discuss breaking contigs that are incorrectly joined in a chimeric misassembly using Hi-C linkage information and the phasing consistency of those links. Finally, I discuss phasing aware scaffolding and the results on *Vanessa atalanta* HiCanu assembly versus Salsa, the most commonly used scaffolder currently. The inputs required by the scaffolder are among the outputs of both the phased assembly algorithm as well as the haplotype phasing algorithm, making the system modular for scaffolding an existing assembly or a phased assembly from phasstools.

4.1.1 Heterozygous kmer pairs and detection

In order to use phasing consistency, we must find paired heterozygous sequences. If using an existing assembly, one could map the reads to the assembly and call variants to find heterozygous sites as is the common workflow for resequencing efforts. In order to be general to a *de novo* assembly process, I tackle the subject of identifying heterozygous variants in a reference-free manner. I do this using a kmer approach, as many reference-free methods do. Many people have focused on identifying kmers that occur at roughly half counts in short read data[214] and various software exists to count kmers[215] and to model the mixture of expected distributions (errors, haploid, diploid, duplication kmers)[328]. Identifying heterozygous kmers in this way suffers from multiple problems. 1. Many of these identified as half counts will be either randomly high count error kmers or randomly low count homozygous kmers due to the fact that the count distributions are generally not fully separated. 2. It ends up with $K-1$ overlapping kmers for a given variant which is needlessly redundant information that will both slow down any phasing algorithm and break key independence assumptions. And 3. while it identifies heterozygous kmers, one doesn't know which kmers are alternative alleles of each other. This information is powerful and key to using my phasing consistency approach. I instead identify pairs of kmers that vary only in the center position that are also both roughly at half counts. These heterozgyous SNP kmers are much more robust and have the benefit of knowing that one is very likely to be the alternative allele of the other. The kmer count spectrum is generated with a fast disk backed kmer counter KMC[67][68][172]. Figure 4.1 shows an example kmer count histogram. The kmer size chosen will determine how unique these kmers are in the genome as well as how likely they are to be correct in the reads. For CLR data, shorter kmers must be used because kmers of any reasonable length (15+) are unlikely to be correct in the reads. But using kmers of this length mean that many of the chosen kmers not be unique in the genome and many of the one-off kmers will also exist in the genome. For this reason, I limit this project to HiFi data and use a kmer size of 31 throughout.

Fig. 4.1: Kmer count spectra and heterozygous paired kmers



An example of kmer count spectrum showing error kmers on the left, heterozygous kmers in the first peak and homozygous kmers in the next peak.

The heterozygous range of the kmer spectra is calculated with code from purge dups[112]. The kmers are then dumped in alphabetical order and pairs of kmers that vary in only the middle base and fall into the heterozygous counts range are identified. A futher restriction is made that the counts of the kmers with the other two possible bases in the middle are not high (in practice >5, although my reasoning would require hundreds to cause the described phenomenon). This is because a very high repeat count kmer may produce, through sequencing errors or mutations, two lower count kmers differing in the middle base. It should also be noted that while I often refer to these kmer pairs as heterozygous kmer pairs, they may also represent paralogous kmer pairs. The phasing consistency of these kmer pairs with others is used to determine if they are more likely to be from heterozygous or paralogous sequences. Each produces a characteristic pattern in the pairwise kmer consistency counts.

4.1.2 Read data kmer information

I use the kmers in the reads to determine if heterozygous paired kmers are phasing consistent with other paired kmers. For each read of each technology (Hi-C, PacBio, linked reads) the position and ID of each paired kmer is stored on disk in a custom binary format for later use. Of note here is that the linked read technology may have multiple molecules per barcode whereas in other technologies, reads represent single physically linked molecules. Richard Durbin has developed a method to *de novo* assign reads from barcodes to molecule groups using shared kmers across barcode sets to cluster reads into their molecules of origin[81]. In this work, I chose not to use this as it is not extensively tested. Instead, the distance in the assembly graph or contigs (if using an existing assembly) is used to determine which reads came from which barcode. With the recommended DNA input, high molecular weight (HMW) extraction sizes, and number of partitions, the Poisson loading process results in an expected number of molecules per barcode of roughly ten. Because the total amount of DNA per partition is a small percentage of the total genome sizes we work with, the chances of a partition having molecules that arose from nearby or overlapping locations is rare. Thus one can deduce that reads from a barcode which map close to one another on a reference or assembly arose from the same HMW molecule with high probability.

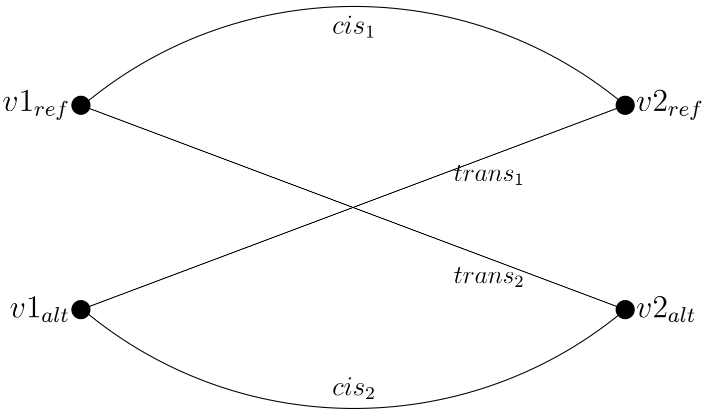
4.1.3 Phasing consistency

For each kmer pair, I refer to one as the reference allele and one as the alternate allele arbitrarily without loss of generality. The read kmer data is used to create phasing consistency counts between different paired kmers. If the read contains the reference version of paired kmer k_1 and reference version of paired kmer k_2, the cis1 count gets incremented. We can do this with any of the data types for different purposes, but we do not combine counts across data types as the error modes are different. For example, the Hi-C data will have some spurious connections across chromosomes due to the 3D conformation of the chromosomes in a particular nucleus. For the linked reads, we may also stipulate that a version of paired kmer k_1 and paired kmer k_2 be within some distance of each other on the assembly graph or contig.

4.1.3.1 Pairwise haplotype phasing consistency

The phasing consistency between pairs of kmer pairs is considered before moving on to higher order phasing consistency. With pairwise phasing consistency, there are only four potential combinations that a read can have of the two pairs (see figure 4.2). With more than two kmer pairs, the number of potential combinations increases exponentially. In order to deal with that, multiple kmers pairs will be phased with respect to one another before considering their consistency with other kmer pairs or groups of phased kmer pairs, always reducing the problem to four possibilities.

Fig. 4.2: Pairwise phasing consistency counts



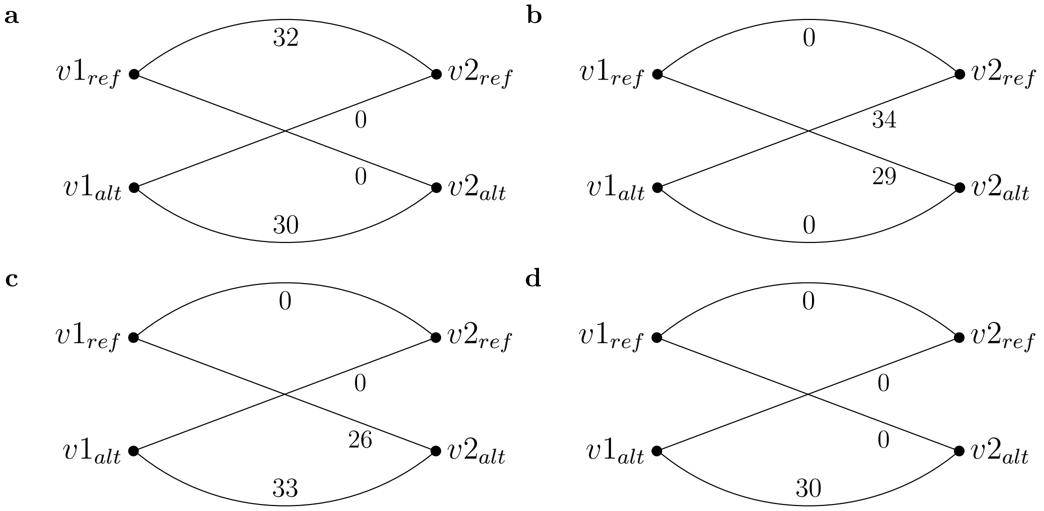
I denote one of each kmer pair as the reference or alternative arbitrarily. Molecules that have the sequences of one of the kmers from each of two kmer pairs will fall into one of four cases represented here by the four edges in this graph. The molecules falling into each of these four categories are tabulated. Phasing consistent heterozygous kmer pairs will have counts predominantly on both cis edges or both trans edges.

4.1.3.2 Phasing consistency and error modes

There are several distinctive signals from pairwise phasing consistency counts and further insight can be gained when looking at those manifestations across the multiple paired kmers to which a given paired kmer is linked.

Figure 4.3 shows the main phasing consistent and inconsistent phenotypes for pairs of paired kmers. It is uncommon to get very mixed signals especially when working with HiFi data. When working with Hi-C data, pairwise counts are not of much use as it is rare for many Hi-C read pairs to link the same two heterozygous kmer pairs.

Fig. 4.3: Pairwise phasing consistency counts



Example of a phasing consistent pair of heterozygous paired kmers **a)** in cis and **b)** in trans. **c)** shows an example of phasing inconsistency because the $v2$ kmer pair is not a heterozygous kmer pair, but likely due to of paralogous sequence. The alternate version of $v2$ is likely homozygous close to $v1$ and the reference version of the $v2$ pair probably exists elsewhere in the genome. If this were due to a tandem repeat, we might expect all four edges to contain counts, perhaps with two edges having fewer counts due to fewer reads reaching into the 2nd repeat. **d)** shows a somewhat ambiguous case. From this, we do not know if $v1$ or $v2$ are heterozygous or not. This could arise from both being homozygous or one of them being heterozygous, but its pair not being the other haplotype.

Due to ambiguous cases expected from data outlined in figure 4.3d), it is not always possible to categorize kmer pairs as heterozygous or paralogous from phasing consistency counts with one other kmer pair. The phasing consistency counts that a single kmer pair has with all of the kmer pairs with which it shares enough counts can be used to find some kmer pairs that are phasing consistent with some others and inconsistent with another set due to those other kmer pairs arising from paralogous sequence. Another set of kmer pairs are inconsistent with nearly all other kmer pairs because they were not truly heterozygous.

4.2 Phasst a: phased assembly

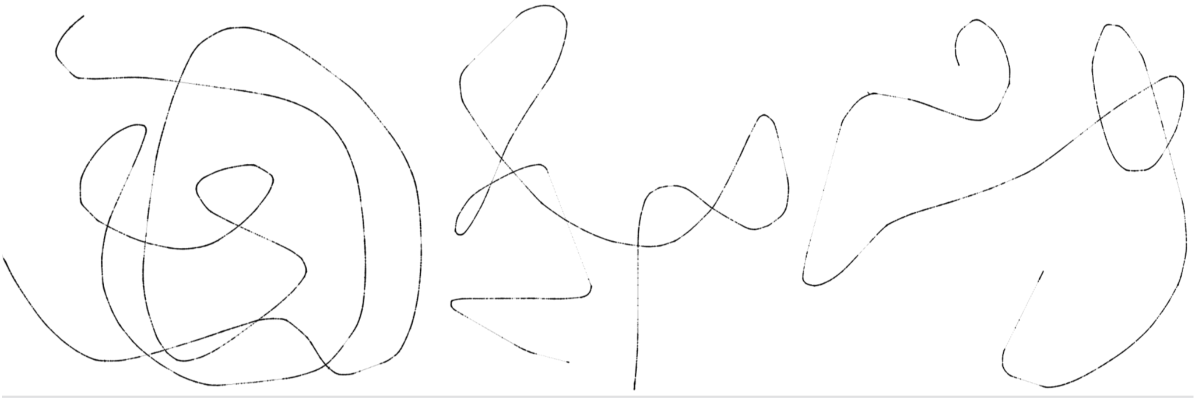
In my first attempt at phased assembly, I used pairwise phasing consistency of kmer pairs. I created a graph where kmer pairs are nodes and edges exist if the two kmer pairs are phasing consistent. I made an edge between two kmer pairs if the phasing consistency was >90%, minor edge of the major phasing was >15% (to avoid problems such as shown in figure 4.3d), and there were at least 10 total counts. This creates megabase scale graphs that are linear on a global level (but not on a local level as kmer pairs will be phasing consistent with many nearby kmer pairs). Figure 4.4 shows examples of these graphs visualized in Gephi with graph layout force atlas 2[21][146].

I applied this method to *Vanessa atalanta*, the Red Admiral butterfly, for which we have ≈40x PacBio HiFi data and ≈90x 10x Genomics linked read data and ≈300x Hi-C data for the same individual. The heterozygosity is roughly 1.1%, which is roughly an order of magnitude higher than the average heterozygosity of humans, which is ≈ 0.1%.

The idea is to then use the connected components of these phasing consistency graphs as psuedo contigs from which kmer pairs can be phased, and then bin reads into haplotypes within those contigs and do haploid assembly of those reads.

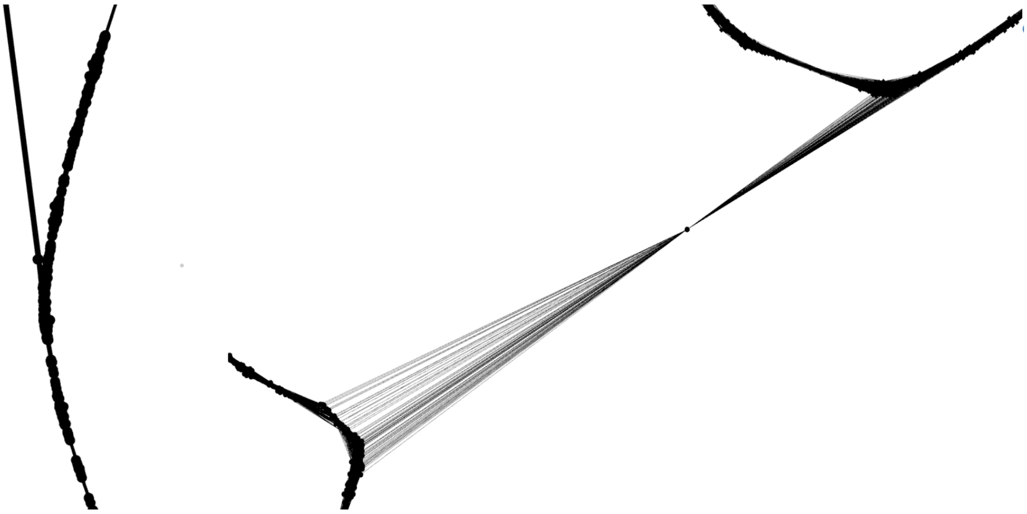
Unfortunately, not all of these graphs were globally linear. This meant that some edges were incorrect. Increasing the stringency of the phasing consistency thresholds did not get rid of all of these spurious connections. If we proceeded with this plan, this could lead to chimeric misassemblies. Figure 4.5 shows two examples of these misassembly causing connections.

Fig. 4.4: Example phasing consistency graph contigs for *Vanessa atalanta*



Graph of kmer pairs with edges between them if they are phasing consistent.

Fig. 4.5: Example of errors in phasing consistency graph for *Vanessa atalanta*



Errors in assembly graph. On the left there is an error in which a single kmer pair is phasing consistent with another single kmer pair elsewhere in the genome. This type of error reduces as the phasing consistency thresholds are made more strict, but the resulting graph becomes more and more fractured. On the right, a single kmer pair consistent with many kmer pairs in two locations. This is because this kmer pair is truly heterozygous in both locations in the genome and randomly had low enough counts to pass our coverage thresholds. This error mode does not reduce with further stringency.

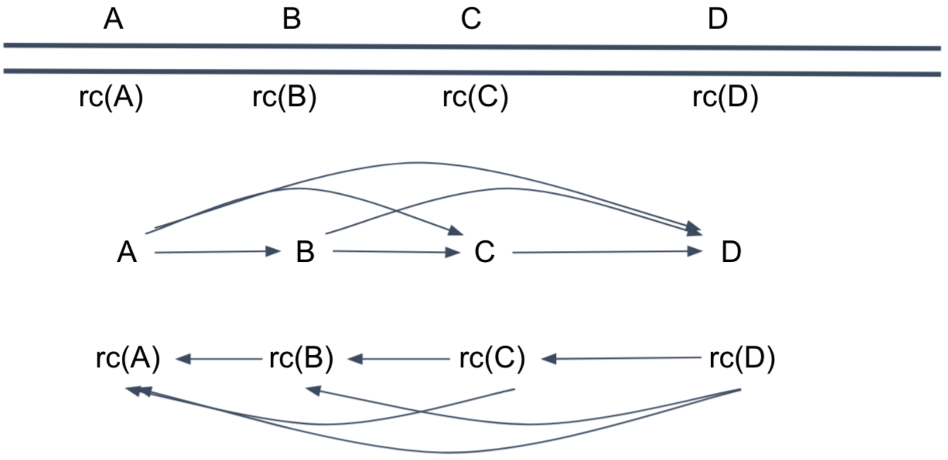
4.2.1 Phasing consistent heterozygous kmer recruitment

To overcome these problems, each new kmer pair should be phasing consistent with multiple prior kmer pairs. To create an ordering of kmers, I begin with a seed kmer pair that has sufficient other kmer pairs it is phasing consistent with on a pairwise basis. Then a graph of kmers is made using the HiFi reads taking into account directionality by also building the graph and the reverse of that graph for the reverse complement of that kmer (see figure 4.6) adding edge counts if the edge already exists. From the seed kmer, this graph is searched in a breadth first manner to choose an order in which to assess new kmers in a directional fashion. Then, at each step, the front kmer pair is popped off of the priority queue and assessed for phasing consistency. If the counts are sufficiently phasing consistent (>90% cis or trans, minor allele fraction of >15%, >10 total phasing consistency counts), that kmer pair is added to the growing phase block in the appropriate phase (whether its dominant counts were cis or trans). And then all molecules containing that kmer pair, if not marked as already used, update phasing consistency counts for other kmer pairs and then are marked as used. If the kmer pair is not sufficiently phasing consistent, it is marked as used and unphased and the process continues. This proceeds until the priority queue is empty. To build the phase block in the other direction from the beginning, the process is reseeded at the initial seed and kmer pairs are added to the priority queue in the opposite direction, keeping track of phasing consistency counts and building the phase block in the other direction until finished. For this step, both HiFi and linked reads are used, both of which have a very low error/cross haplotype signal. Only the HiFi data is used to build the kmer recruitment graph as the linked reads don't have direction across different reads within the same barcode.

4.2.2 Contig and haplotype read binning

Once this kmer pair phased assembly is done, it is used to assign HiFi reads to contigs and haplotypes within contigs. Each read's kmer content is inspected and kmer counts for each contig/haplotype are calculated. Usually, the match is unanimous, but in the case of conflict, the read is assigned if it favors one contig haplotype by three or more kmers.

Fig. 4.6: Kmer recruitment graph



Kmer recruitment graph from HiFi reads.

4.2.3 Haploid chromosome assembly

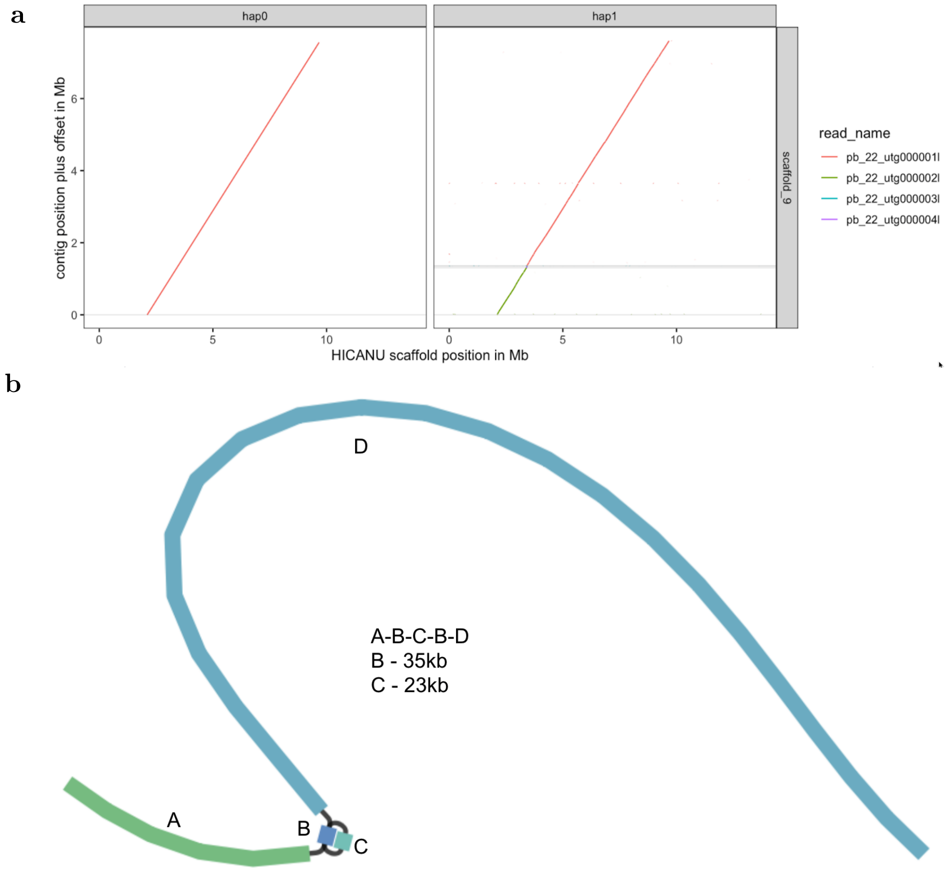
Now that reads are binned to contigs and haplotypes, the next step is haploid assembly. The miniasm assembler is used with parameters suited to HiFi data as opposed to the defaults which were for the more noisy CLR data[196]. This may result in a single contig per assembly phase block or multiple contigs.

To validate, the contigs from these assemblies are mapped to the HiCanu assembly. The contiguity of this assembly is less than HiCanu, with an N50 of ≈900kb. And there are sometimes interesting heterozygous differences between the two haplotype assemblies. In figure 4.7 the haplotype assemblies of one phase block aligned to the HiCanu reference are shown. In one haplotype, the full phase block was assembled as a single contig. The other haplotype assembled into three contigs. The haplotype read sets aligned to the HiCanu assembly were inspected in IGV (not shown) as well as the bandage plot for the GFA from the miniasm assembly of that haplotype, and there is evidence for a 35kb duplication flanking a 23kb unique sequence[340]. This duplication does not appear to exist on the other haplotype.

4.2.4 Assembly contig coverage vs HiCanu assembly

One of the haplotype assemblies from each assembly phase block was aligned to the HiCanu assembly to inspect the overall contig coverage of our assembly (figure 4.8). The first thing one notices is that there are three HiCanu contigs with very little assembly

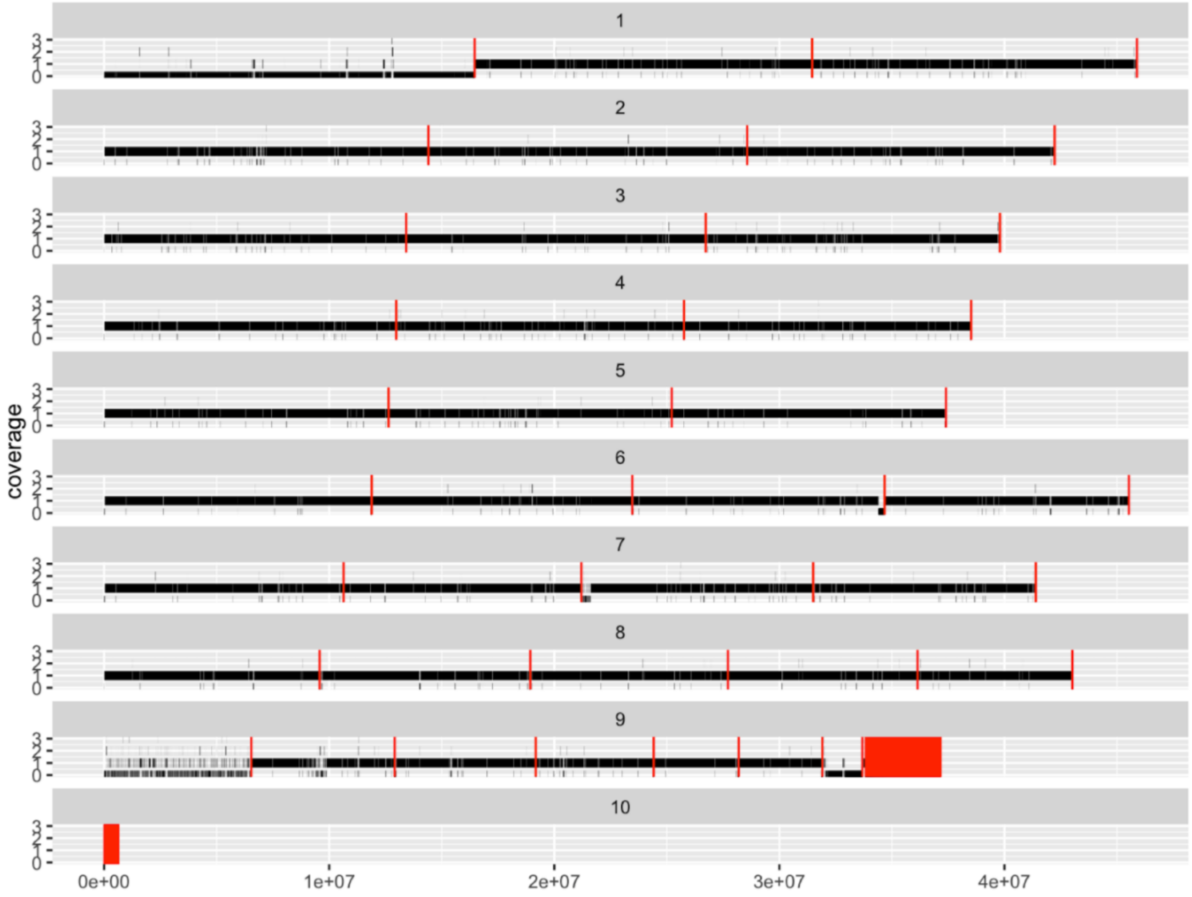
Fig. 4.7: Example of heterozygous structural variation.



a) alignment plots of haplotype contigs vs HiCanu assembly shows one haplotype not assembling into a single contig because **b)** one haplotype has a ≈35kb repeat that is not present in the other haplotype. HiCanu retained the haplotype that was most contiguously assembled.

coverage. These are the haploid sex contigs that do not get assembled by our method as it currently stands. The more complex contig on the lower right sex contig is rich in repeat sequence and our system has assembled some of it. The other notable discrepancy between these assemblies is the zero contig coverage areas on the ends of two contigs in rows six and seven. The HiFi read alignments to the HiCanu assembly were inspected and there is no support for this connection and thus are likely misassemblies in the HiCanu assembly.

Fig. 4.8: Assembly contig coverage



Contig coverage of one haplotype of the phased assembly aligned to the HiCanu assembly. The haploid sex contigs have not been assembled. The drops in coverage on two chromosomes in rows six and seven have been determined to be misassemblies in the HiCanu assembly.

4.3 Phasing aware scaffolding: phasst scaff

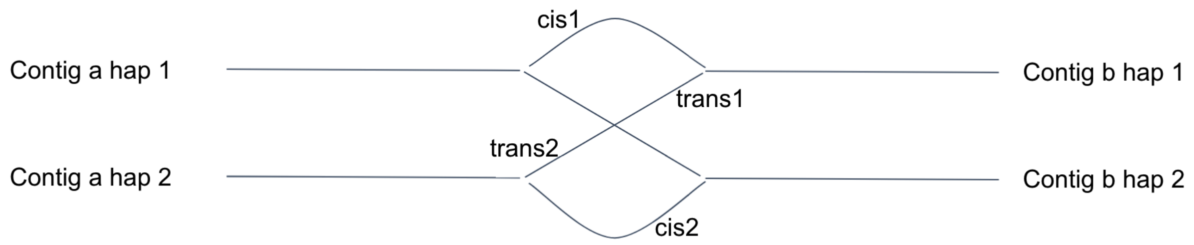
Because the process of assembly rarely creates chromosome length contigs, the order and orientation of these contigs in their chromosomal context is unknown. The process of finding this is called scaffolding. Generally long genetic information, such as Hi-C, linked reads, or optical maps are used for this purpose. The most popular scaffolding method currently is SALSA2, which uses Hi-C paired reads in which one maps to one contig and the other maps to another contig to scaffold. The density of these cross mappings to contig ends are normalized and a greedy approach is used to choose which connections are made[104]. Scaff10x uses the co-occurance of molecular barcodes from linked reads across contig pairs to scaffold in a similar fashion[121]. And Bionano genomics uses optical maps which align to multiple chromosomes to provide order and orientation as well as estimated gap sizes between contigs[296].

Here the goal is to use the phasing consistency of Hi-C reads (and optionally linked reads) that cover heterozygous sequence on two contigs from our phased assembly or an existing assembly to scaffold contigs. If using an existing assembly, the heterozygous sites must be identified first. Then phasing consistency of those sites can be defined.

4.3.1 Multiple heterozygous site haplotype phasing consistency

When using phasing consistency as a signal for assembly scaffolding, all heterozygous kmer pairs in one contig are compared to all heterozygous kmer pairs in another contig. Figure 4.10 outlines what that looks like in the same way the previous phasing consistency diagrams did for pairs of individual kmer pairs. In order to do this, the heterozygous kmer pairs within each contig must be phased with respect to each other.

Fig. 4.9: Phasing aware scaffolding



Using phasing consistency for scaffolding requires all of the heterozygous kmer pairs within a contig to already be phased with respect to one another.

4.3.2 Haplotype phasing

A combination of data types are used to phase. The data type that is required is Hi-C, but long reads or linked reads can be used in addition to them. The reason Hi-C data is required is due to the very long range genetic information it gives us. With Hi-C, it is possible to phase whole chromosomes instead of potentially having to break the phasing into phase blocks that are smaller than the chromosome. While there are existing phasing algorithms that can take multiple data types including Hi-C[84][293], I have developed my own phasing algorithm that has several benefits in general and meets our specific purposes. This algorithm is robust to non heterozygous sequence inputs (which our kmer pairs will have) and can correct these genotypes. It also has the added benefit of being trivially extendable to polyploid genomes.

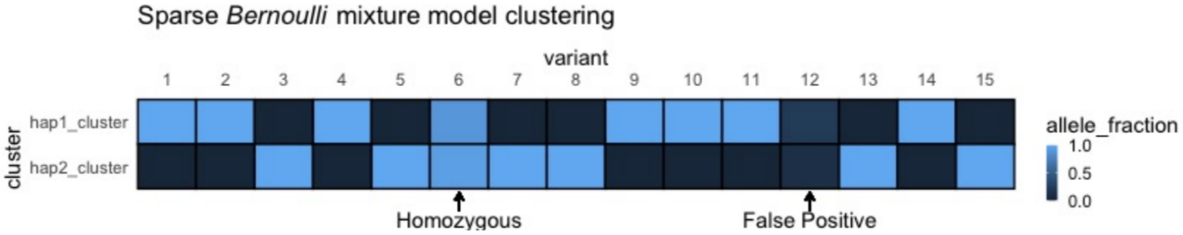
4.3.2.1 Sparse *Bernoulli* mixture model clustering

I treat the haplotype phasing problem as a clustering problem. We first consider the diploid case. By clustering reads according to heterozygous kmer pair alleles they contain, this gives an implicit phasing of each kmer pair. Similar to the clustering algorithm in Chapter 2, I treat the haplotype clusters as vectors of allele fractions. In Chapter 2 this represented the probability parameter of a binomial because the underlying data were allele counts, but here it represents the probability parameter of a *Bernoulli* as I make the assumption that a single read can only have one or the other of two alternate alleles. Of course each haplotype will only have one allele, but treating it as a continuous number instead of a discrete number opens the door to continuous numerical optimization methods that are often faster than discrete combinatorial optimization techniques. The optimization process then can drive the allele fraction numbers to 0 or 1 if the data supports that outcome.

Definitions:

- H: number of haplotypes. Lower case h will be used for indexing and referring to a specific haplotype.
- R: number of reads. Lower case r will be used for indexing and referring to a specific read.
- V: number of kmer pairs. Lower case v will be used to index and refer to a specific locus. V_r will be a list of kmer pairs with observed data in read r.

Fig. 4.10: Sparse *Bernoulli* mixture model clustering haplotype phasing



The phasing problem is treated as a sparse clustering problem in which reads are clustered into haplotypes.

- A: Alleles. $A_{v,r}$ is a Boolean representing whether read r for kmer pair v had the alternative allele or the reference allele.
- $\phi_{h,v}$: cluster center value representing allele fractions of haplotype h at locus v. The expectation is for this to be near 1.0 or 0.0 because each haplotype can only have one allele or the other, but I allow the value to be continuous to allow for continuous numerical optimization techniques and errors.

4.3.2.2 Model

A maximum likelihood strategy is used by optimizing $\mathcal{L}(data)$ under a given model.

$$\operatorname*{argmax}_{\phi} \mathcal{L}(data, \phi) \tag{4.1}$$

The likelihood of the data is defined by treating reads independently and marginalizing each read across the haplotypes it could belong to.

Cluster model Likelihood function

$$\mathcal{L}(A) = \prod_{r \in R} \sum_{h \in H} \frac{1}{H} \prod_{v \in V_r} \begin{cases} \phi_{h,v} & A_{v,r} == true \\ 1 - \phi_{h,v} & A_{v,r} == false \end{cases} \tag{4.2}$$

This likelihood is then optimized by expectation maximization, randomly initializing cluster centers $\phi \in (0, 1)$. Due to the local nature of long reads and linked reads as well as the majority of Hi-C reads, this method can get stuck in local maxima in which for one region, the random initialization favored reads from one haplotype going to haplotype cluster 1 and in another region, reads from that same haplotype preferring haplotype cluster 2. In this case, the optimization would come into conflict at the boundary between these two regions. One could make this solve localized to a window and grow that window over time. Another option is to use the very long scale Hi-C links for an initial solve, thus giving the haplotype cluster centers a rough global initial solution before adding in all of the data and fine tuning the local solutions. This is what I currently do, but this requires plentiful Hi-C data for the initial long range solution. It is also possible that by employing the same deterministic annealing strategy as in Chapter 2, it may be able to get through these local maxima, but I have not tried this.

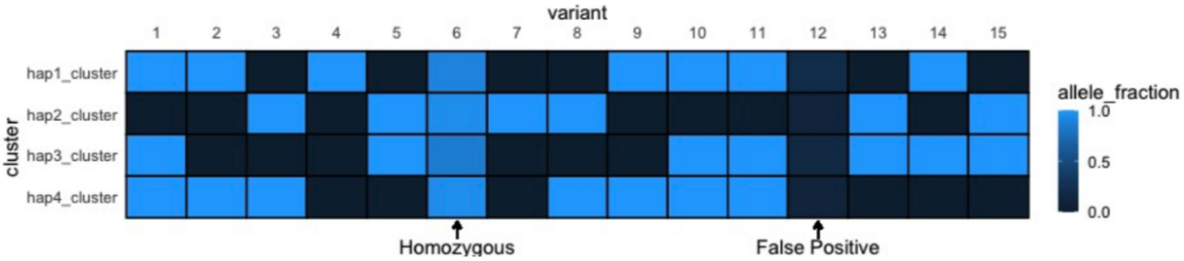
4.3.2.3 Polyploid phasing

This algorithm has several advantages compared to other haplotype phasing methods. Most methods' time complexity scales super linearly, oftentimes exponentially, with ploidy, if they are able to solve polyploid phasing problems at all. In this method, since haplotypes are just cluster centers, one just adds more cluster centers to match the ploidy. It would be untrue to say that this definitely scales linearly for an optimal solution. This algorithm does not guarantee optimality but neither do most modern phasing algorithms. The only modern phasing algorithm I know of that guarantees optimality is WhatsHap, a method that uses dynamic programming to ensure optimality but does not handle gapped data such as linked reads and Hi-C well[254]. While this algorithm is not optimal, and its iterations to convergence may change with ploidy, the computation per expectation maximization step scales linearly with ploidy.

4.3.2.4 Genotype correction via phasing

Additionally this algorithm is highly robust to non heterozygous sites. Most phasing algorithms assume as input a set of heterozygous variants. However, even with standard resequencing, read mapping, and variant calling, some called variants will be false positives (homozygous reference) or falsely called heterozygous when they are actually homozygous for the alternative allele. This algorithm has the valuable property that the inputs are not assumed to all be heterozygous. If input variants are not heterozygous, the haplotype centers should be driven to the correct genotype by the data. If the reads support that this kmer pair is a false variant, the values will be driven to 0 for both haplotypes, and if the reads support the alternative on both haplotypes, the values will be both driven to 1. This type of genotype correction via phasing has been done before, but with a 3rd discrete categorical state on top of the two normal states representing a 0|1 vs 1|0 phasing[182]. In that method, the 3rd state represented either 0|0 or 1|1 by marginalizing across both. If that state was chosen, the posterior for each case was calculated and genotype reassigned accordingly. This does work, but has the dramatic downside of increasing the total solution space from 2^n to 3^n where n is the number of loci.

Fig. 4.11: Sparse *Bernoulli* mixture model phasing allows for polyploid phasing and genotype correction



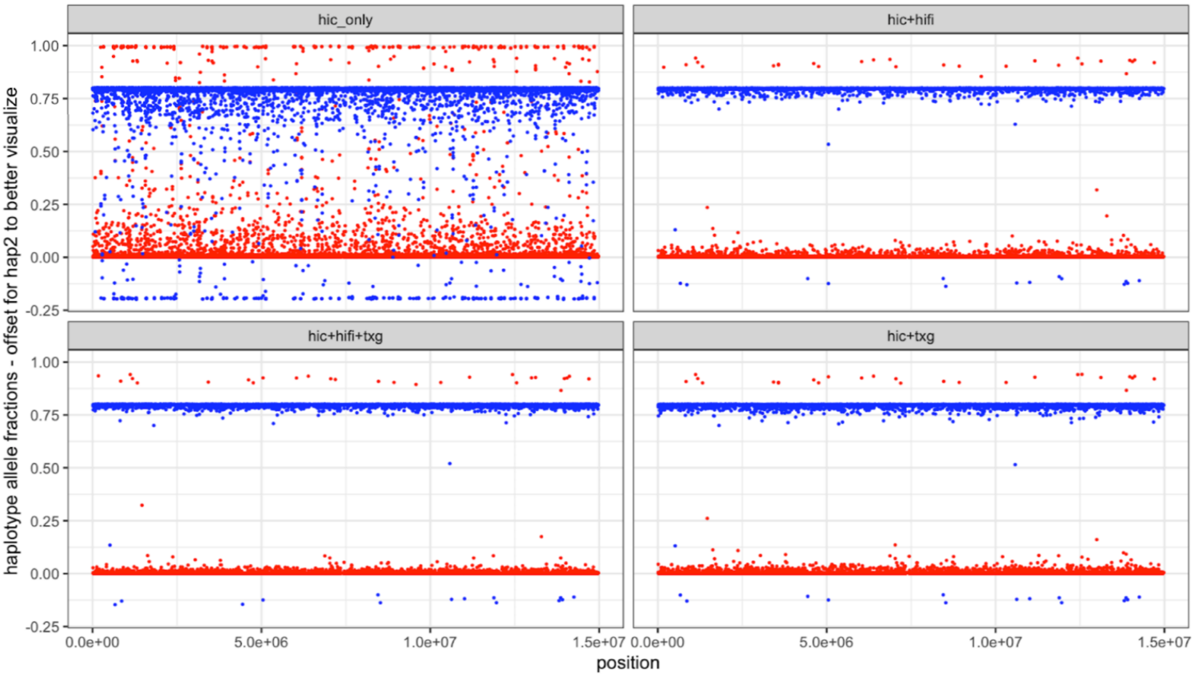
In order to handle polyploid phasing, one simply increases the number of cluster centers. And because sites are not assumed to be heterozygous, the algorithm can naturally correct miscalled genotypes.

I ran this algorithm on Hi-C Arima, 10x Genomics linked read, and PacBio HiFi data from the butterfly *Vanessa atalanta* from the Darwin Tree of Life project with ≈300x, ≈90x, and ≈40x coverage, respectively. I then compared this phasing to which allele the HiCanu assembly contained. HiCanu uses the HiFi data, and while it does not explicitly phase haplotypes, it uses such stringent filters for read overlaps, that generally one haplotype is assembled in a contig unless there is a homozygous region that no read spans. Initially I show the phasing cluster center values on a single contig to compare phasing with Hi-C alone vs with linked reads and/or long reads (see figure 4.12). And in figure 4.13 I show the phasing against the entire genome.

4.3.3 Phasst phase

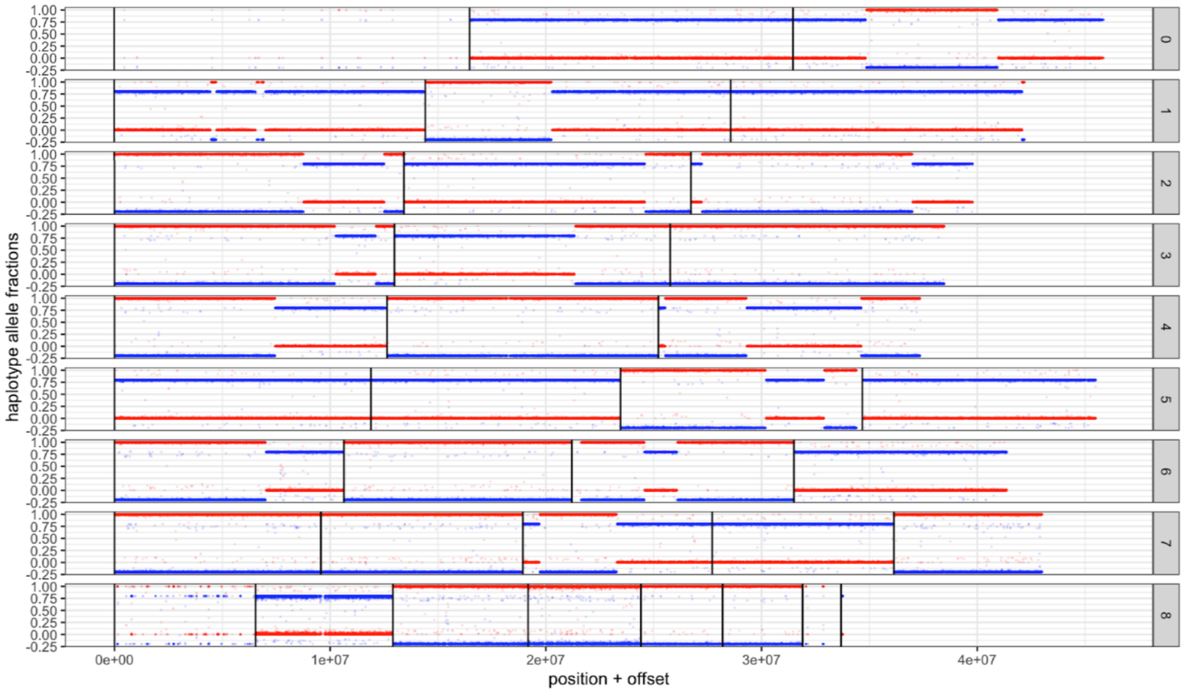
Because the sparse *Bernoulli* mixture model haplotype phasing can run into local optima, it currently requires plentiful and high quality Hi-C data. Also, for some organisms, such as diptera, chromosome pairs colocate within the nucleus[303], greatly increasing the number of cross haplotype Hi-C links. This makes it not always sufficient to have an initial solve with just the long range Hi-C. I still believe that this algorithm can be made to work either with a moving or expanding window or through deterministic

Fig. 4.12: Phasing with Hi-C alone vs combined data



Here I show the haplotype cluster center values for sparse *Bernoulli* mixture model haplotype clustering with the values for the haplotype 2 cluster offset by -0.2 such that any phase switches would be visible and not overwritten by other data points. Each point represents the haplotype cluster center value of a heterozygous kmer. The kmer observed on the HiCanu assembly is colored red while its paired kmer is colored blue. Hi-C alone (upper left) produces fairly noisy phasing probably because the chances that a Hi-C read hits two heterozygous loci is fairly low reducing the total amount of useful data. When linked reads and/or HiFi reads are added, the phasing becomes much robust. *txg = 10x Genomics linked reads

Fig. 4.13: Phasing *Vanessa atalanta* genome



Phasing for *Vanessa atalanta* genome. Vertical black bars indicate contig ends and the contigs are distributed over several rows for visibility (right facet is row number and can be ignored). Here one can see that contig 1 as well as the first and last large contigs on the last row are from sex chromosomes and thus haploid. Again, the cluster center values for haplotype cluster 2 were offset down by 0.2 for visibility. There are rare, but notable long switch errors vs the HiCanu assembly. Because my algorithm uses Hi-C data with chromosome length scale information, it is likely that these phase switch errors are errors in the assembly and not my phasing algorithm.

annealing, but we decided to take another direction for the time being. I developed a reference/assembly based phasing consistency rules based phasing algorithm. This is very similar to the phased assembly algorithm, but uses the existing assembly to determine the order in which kmers are treated. First a good seed kmer is found by looking at all pairwise phasing consistencies. The kmer must have sufficient other kmers that it is phasing consistent with. Then phasing consistency counts are added for all molecules that contain this seed kmer pair to a growing data structure and mark those molecules as used. I then proceed according to the order that kmers from the paired kmer pair set occur on the assembly. If the kmer pair's phasing consistency counts are sufficient (>90% consistent—mostly cis or mostly trans, the minor edge of the major phase—cis or trans—makes up >15% of the counts (avoiding the phasing consistency error modes outlined in figure 4.3), and the total phasing counts is at least 10), it is added to the growing phase block in the appropriate phase. Then for all molecules that contain that kmer and are not already marked used, potential new kmer counts are updated and that molecule is marked as used. This is done until the end of the contig is reached or there are a string of 10 kmer pairs with zero phasing consistency counts. I then go back to the seed kmer and proceed in the same way in the other direction. A simple diagram of this process is shown in figure 4.14.

Fig. 4.14: Building phase blocks

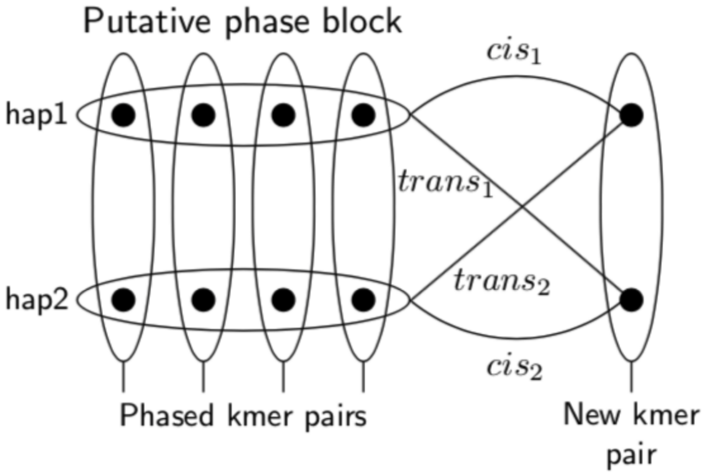


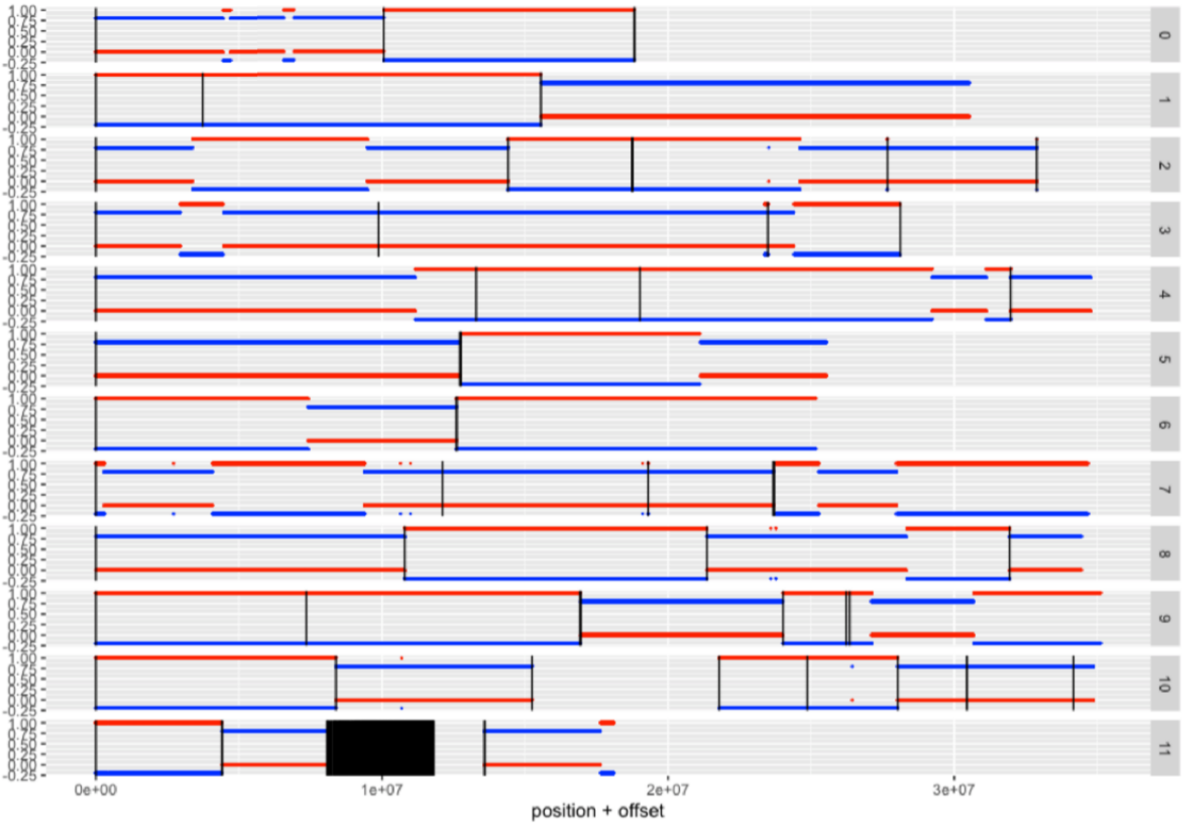
Diagram outlining the process of building phase blocks with HiFi and linked reads using phasing consistency rules.

Because only HiFi and linked reads have been used in the process thus far, it is not expected to create contig or chromosome length phase blocks. Instead, very high quality phase blocks are made that can then be phased with respect to one another using the

Hi-C reads. This is done in a way that is very similar to how Hi-C is used for phasing aware scaffolding. The Hi-C reads that have two or more phased kmer pairs in different phase blocks are used to create inter-phaseblock phasing consistency counts. Because the phase blocks are typically hundreds of kilobases long, there are plenty of Hi-C molecules linking them with two or more heterozygous kmer pairs.

I ran this on the same data used to demonstrate the sparse *Bernoulli* mixture model phasing, this time using the HiCanu assembly post purge dups but pre-scaffolding and curation and show the results in figure 4.15.

Fig. 4.15: Phasst phase results on *Vanessa atalanta*

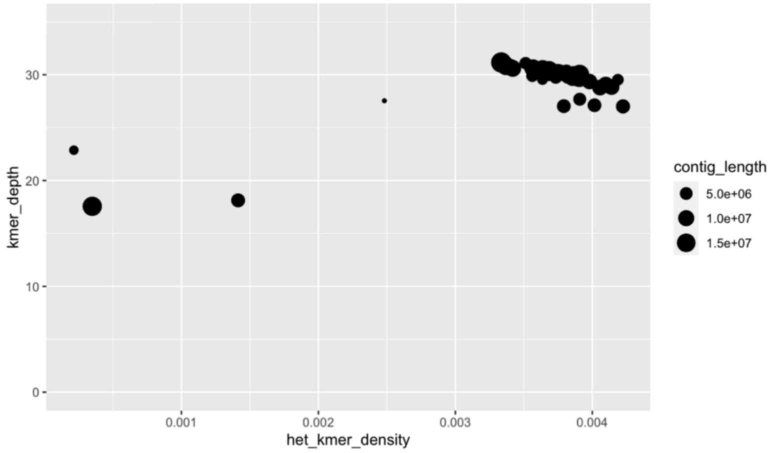


Phasst phase shows cleaner phasing than sparse *Bernoulli* mixture model phasing because it skips over any kmer pairs that are not phasing consistent, it only allows discrete phasing instead of a continuous solution. Vertical black bars are contig ends. Because we are using the pre-scaffolded and pre-curated assembly this time, there is a collection of very short contigs which were placed or removed in the curation process (this explains the black section on row 11 which is caused by many short contigs).

4.3.4 Haploid / sex contig detection

Phasing consistency counts should only be used when the chromosome is diploid (or polyploid if the rules were extended to apply to polyploid genomes). This makes contigs coming from haploid sex chromosomes problematic. It would be desirable to detect these prior to phasing and scaffolding such that they can be treated separately in a more standard fashion by using Hi-C read linkage as a signal for scaffolding. Kmer counts and kmer pair density are used as signals for haploid contigs. Using a fast kmer package from Gene Myers called FASTK[243], I find all kmers that occur exactly once in the assembly and check their counts in the HiFi data and take the mean for each contig. For kmer pair density, I simply find how many of the kmer pairs occur on each contig and divide by the contig length. Figure 4.16 shows contig kmer counts vs kmer pair density for *Vanessa atalanta* for contigs > 100kb.

Fig. 4.16: Haploid / sex contig detection for *Vanessa atalanta*



Kmer counts are a signal for haploid/sex contigs, but only a weak one because the heterozygosity for this species is so large, much of the sequencing in diploid chromosomes are haploid counts. Kmer pair density is a much better signal for haploid contigs. The three contigs on the left of the graph are sex contigs. The one intermediate contig is just over 100kb and we do not know if it is diploid or not.

While haploid sex contig detection works well for this dataset and assembly, it is quite different for different species. This could be due to pseudoautosomal regions, repeat structure, or several other factors. An alternative method would be to use orthologous genes known to be on sex chromosomes in related species for sex contig detection.

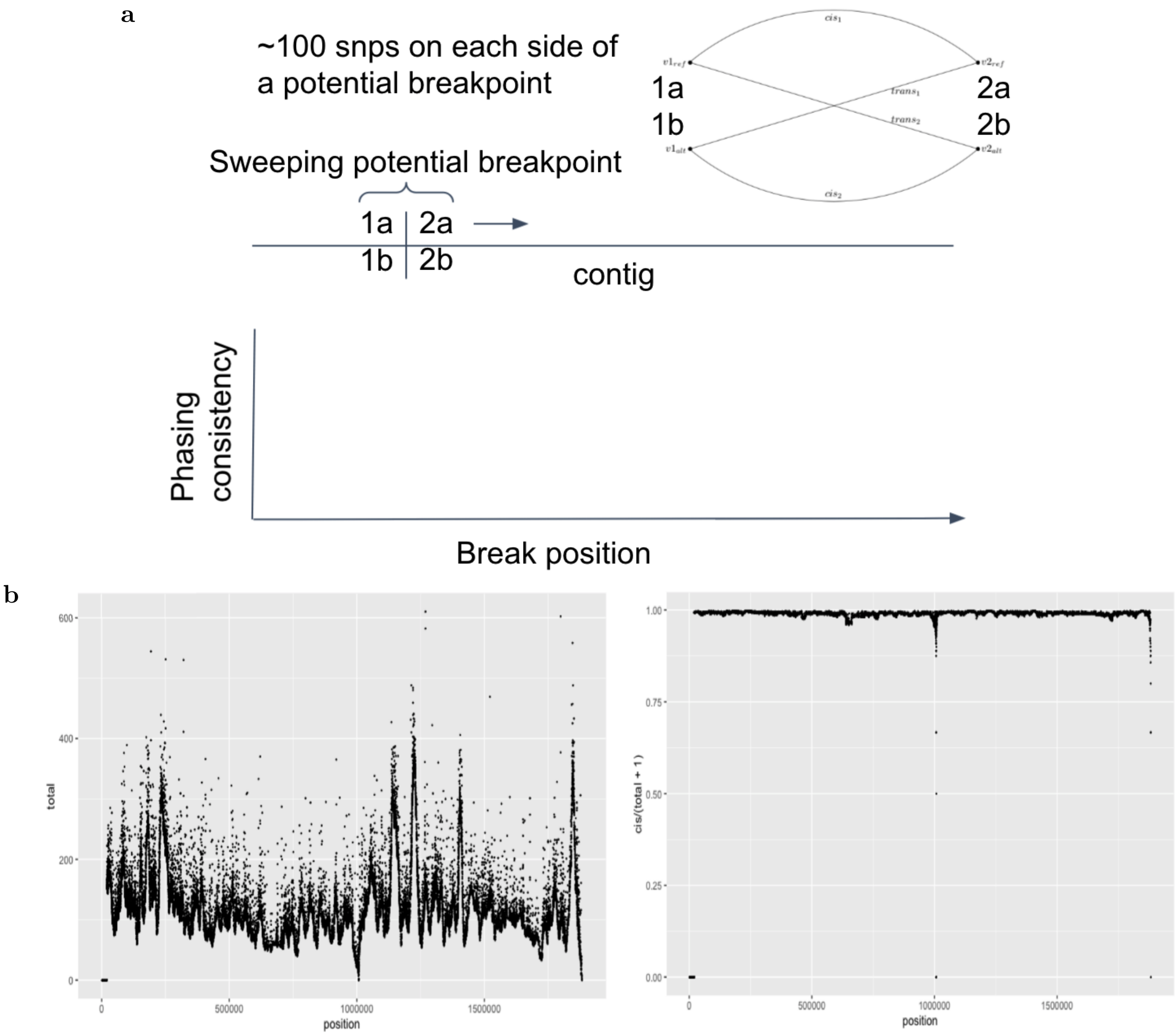
4.3.5 Contig breaking

Before scaffolding, one should break any contigs that are likely chimeric misassemblies. This is a common first step in scaffolding[104][272]. The Hi-C data is used to break contigs across regions where there are extremely low or no Hi-C links. Both paired kmers as well as a modimized sampling of likely homozygous kmers are used in order to utilize more of the data. Because Hi-C uses a restriction enzyme or a collection of restriction enzymes to create its cuts and then enriches for reads ligated by those cuts, the coverage across the genome is not uniform, but in a wide enough region there should be Hi-C links with high probability if those two regions are physically close in the 3D genome (and also in the 1D sequence). A window is swept across kmer positions on the contig and count how many Hi-C links cross the midpoint of that window. Both raw counts and phasing consistency counts of kmer pairs across the midpoint of the window are kept separately. Figure 4.17 shows a depiction of this process as well as data on a contig that contains a chimeric misassembly and requires breaking from the *Vanessa atalanta* HiCanu assembly. I show both total links across the midpoint of a 200 kmer pair window as well as the percent phasing consistency across that window. In this case, the total links drops to zero, so of course the phasing consistency also drops to zero, but one can see how phasing consistency for breaking contigs can be a very powerful signal. However, this does require that one has detected the sex contigs in order to treat them differently. With this, I found the same three breaks in the HiCanu assembly that SALSA does (within 10kb of the SALSA break).

4.3.6 Phasst scaff: phasing aware assembly scaffolding

Now that the paired kmers are phased on each contig, at least for the ones that could be phased, one can now look at the phasing consistency counts of pairs of contigs by looking at Hi-C reads that contain kmer pairs that occurs on two contigs. An example of a phasing consistent match we found has the following counts: cis1: 10150, cis2: 10278, trans1: 98, trans2: 101. A binomial test is used to compare cis vs trans counts to a random expectation of 0.5 and of course these numbers yield exceedingly low p-values. The counts between non matches are essentially randomly distributed between cis and trans. For the HiCanu assembly of the butterfly *Vanessa atalanta*, the system found the same scaffolding matches that SALSA finds.

Fig. 4.17: Contig breaking using Hi-C total linkage and phasing consistency



a) shows the setup for contig breaking. A 200 kmer pair window is used and the linkage and phasing consistency are calculated across the midpoint of a sliding window. Below, **b)** shows the results of this at each position of that sliding window. The low values on the left and right are low due to how this is calculated and should be ignored. The low value in the middle represents a contig misassembly that should be broken.

4.3.6.1 Chromosome binning

Using these matches, a graph is created where nodes are contigs and edges are phasing consistent scaffolding matches. The connected components of this graph are found and these connected components of contigs represent the chromosomes these contigs belong to.

4.3.6.2 Order and orientation

To order and orient, the heterozygous and homozygous kmers on the ends of each contig (100,000 bases on each contig end, or to the middle of the contig if shorter then 200kb) are used. Counts of how many Hi-C reads link the start and end of each contig to the start and end of each other contig in the chromosome bin are calculated. This number is then normalized by the actual length used (if less than 100,000). SALSA takes these normalized link numbers between the ends of each contig that passes its scaffolding thresholds and takes the highest weight link and makes that a new longer scaffold and then the next highest weight until completion.

I have not yet implemented order and orientation, but have sketched out how I plan to do this. I will create a stochastic process in which the edges are chosen with a probability weighted on the normalized linkage value and otherwise proceed as SALSA does. This process will be run many times and the final scaffolding will be selected that maximizes the total sum of normalized linkage weights. For each link, I will be able to report what percentage of the stochastic scaffoldings it was chosen to give a level of confidence to the ordering and orientation.

Although some of the elements of this chapters are not as polished as chapters 2 and 3, which correspond to published papers, I feel this material contains multiple significant contributions to the field and will result in publications in the future.

4.4 Discussion

Due to recent advances in long read and long range genetic information technologies, our ability to assemble high quality genomes has increased greatly. But some issues still exist around high heterozygosity and small organisms. With the advent of the Earth Biogenome project and the Darwin Tree of Life project, there is a need for new assembly and scaffolding methods that are robust to these issues and work on a wide range of organisms. While there are a number of phased assembly methods currently available, I take the approach of treating phasing consistency as a first class signal for physical linkage and make it just as important as sequence similarity in the assembly and scaffolding processes.

I presented a method for phased assembly by using heterozygous kmer pairs and their phasing consistency as a signal of physical linkage. I then use this to bin HiFi reads into haploid bins that can be assembled separately. By phasing prior to assembly, the errors and bias caused by the ambiguity between heterozygosity and paralogous sequences are removed. Furthermore, I generate a complete set of paired haplotype sequences, linked by confirmed single-copy kmer-pairs, in the process identifying large scale structural variation. Although N50 numbers are shorter than from some modern assemblers, I believe the linked haplotype output of our method is the correct goal for diploid assembly. The same approach could in principle be applied to higher ploidy genomes.

I demonstrated a new phasing aware scaffolding method and compared it to the most commonly used existing scaffolder. It is clear that phasing consistency is a powerful tool in determining if contigs should be scaffolded together or not. There is a dramatic difference between the phasing consistency counts of contigs from the same chromosome and different chromosomes using Hi-C reads.

I also described a novel algorithm for *de novo* haplotype phasing. This algorithm is capable of scaling to polyploid genomes efficiently by adding more haplotype cluster centers. It is also robust to receiving as input sites that are not actually heterozygous and is capable of correcting the genotypes of those sites. This algorithm is of general use even outside of the context of phasing aware scaffolding. I believe this will be of particular use for plant genome research as many crops are polyploid. When expanded with either a growing window approach or deterministic annealing, I believe that the local optima problems from regionality of many data types will be solved.

Chapter 5

Conclusions

Genetic variation along with natural selection, has driven the evolutionary history on earth creating us and all other life. Much work has been done to assess population variation across humans and other species and use that to link genotypes with phenotypes and infer evolutionary histories. Less work has used these as markers to disambiguate data in different problems in genomics.

ScRNAseq suffers from several error modes that arise from natural limitations of detections of small materials as well as from the strategy used to partition cells. Because cells contain miniscule amounts of mRNA, amplification methods must be used and these lead to technical artifacts between cells within an experiment and across experiments. Because cells are loaded into droplets in a Poisson process, in order to capture many singletons, the cell suspension must be in a concentration that will also randomly load two or more cells into a single droplet. And if cells lyse or if there is RNA in solution prior to partitioning, some reads will have cell barcodes of cells from which they did not originate. One experimental design promises to address each of these issues at once. Mixing cells from multiple individuals reduces the batch effects when comparing them, cross sample multiplets should be easier to identify and remove, and the skew in allele fractions away from those expected from a diploid genome may be used to measure the ambient RNA. In chapter 2, I presented souporcell, a computational tool which uses the genetic variation between individuals to cluster cells in a single cell RNAseq mixture of individuals by the variants expressed in the reads. Souporcell also calls doublets using the alleles in cells versus the alleles in each cluster. And souporcell estimates the amount of ambient RNA in the system by how far the allele fractions in the clusters vary from those one expects from a diploid organism. I validated and compared souporcell to the other relevant tools and found that it compared favorably against all of them including the previous gold standard, demuxlet, which requires more information *a priori.* I believe this is due to the rigid model based system used in demuxlet versus the simple cluster center method that I used, which is robust to any unmodeled factors. Souporcell has already been used in several million+ cell experiments and has been externally validated using cell hashing. I believe that mixture experimental designs will only get more popular over time due to the advantages this strategy has.

Long read sequencing has undoubtedly revolutionized genome assembly and has motivated large projects such as the Darwin Tree of Life and Earth Biogenome projects that seek to sequence and assemble high quality genomes for all multicellular eukaryotic organisms. This will provide the next generation of science with an invaluable resource, allow for insights into evolution not possible before, and serve as a conservation of biological information in an era of extinction unprecedented during humans' time on earth. While much progress has been made through both data improvements and algorithmic methods, some issues remain especially for small organisms and highly heterozygous genomes. In chapter 3, we present the first high quality assembly of a single mosquito. This was made possible by recent advances in library preparation for long read sequencing reducing the DNA requirements. This improved the assembly versus other mosquito assemblies which used pools of individuals or short reads because of the presence of only two haplotypes versus many haplotypes.

While many consider heterozygosity a hinderance to genome assembly, and most assemblers perform worse on highly heterozygous genomes, it can have some benefits as well. In chapter 4, I turn this idea that heterozygosity is bad on its head and instead use heterozygosity as an advantage by using the phasing consistency of reads across multiple heterozygous sites as a signal for physical linkage in both assembly and scaffolding. In doing this, I also describe two phasing algorithms (three, if phased assembly is counted). One of these algorithms is novel and has several benefits of general interest. It is both robust to, and can correct incorrect genotypes because it does not make the assumption that every input variant is heterozygous and has relaxed the discrete constraint shared by most phasing algorithms. Another benefit of this algorithm is that it scales well for polyploid because we use treat haplotypes as cluster centers and can trivially increase the number of clusters as the ploidy increases. We demonstrate our phased assembly and scaffolding on the lepidoptera *Vanessa atalanta.* While we do not meet the contiguity of some modern assemblers and do not assemble sex chromosomes, these can be assembled separately. We believe that phased assembly is the correct solution to diploid and polyploid assembly.

References

[1] Phrap. http://www.phrap.org/phredphrapconsed.html#block_phrap. Accessed: 2021-09-8.

[2] 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.

[3] 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.

[4] 10x Genomics. 10x genomics. https://www.10xgenomics.com/. Accessed: 2021-09-8.

[5] Fred Aboagye-Antwi, Nahla Alhafez, Gareth D Weedall, Jessica Brothwood, Sharanjit Kandola, Doug Paton, Abrahamane Fofana, Lisa Olohan, Mauro Pazmiño Betancourth, Nkiru E Ekechukwu, Rowida Baeshen, Sékou F Traorè, Abdoulaye Diabate, and Frédéric Tripet. Experimental swap of anopheles gambiae's assortative mating preferences demonstrates key role of x-chromosome divergence island in incipient sympatric speciation. *PLoS Genet.*, 11(4):e1005141, April 2015.

[6] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, 14(11):1083–1086, November 2017.

[7] Javier Antonio Alfaro, Peggy Bohländer, Mingjie Dai, Mike Filius, Cecil J Howard, Xander F van Kooten, Shilo Ohayon, Adam Pomorski, Sonja Schmid, Aleksei Aksimentiev, Eric V Anslyn, Georges Bedran, Chan Cao, Mauro Chinappi, Etienne Coyaud, Cees Dekker, Gunnar Dittmar, Nicholas Drachman, Rienk Eelkema, David Goodlett, Sébastien Hentz, Umesh Kalathiya, Neil L Kelleher, Ryan T Kelly, Zvi Kelman, Sung Hyun Kim, Bernhard Kuster, David Rodriguez-Larrea, Stuart Lindsay, Giovanni Maglia, Edward M Marcotte, John P Marino, Christophe Masselon, Michael Mayer, Patroklos Samaras, Kumar Sarthak, Lusia Sepiashvili, Derek Stein, Meni Wanunu, Mathias Wilhelm, Peng Yin, Amit Meller, and Chirlmin Joo. The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods*, 18(6):604–617, June 2021.

[8] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors, 2013.

[9] Jose Alquicira-Hernandez, Anuja Sathe, Hanlee P Ji, Quan Nguyen, and Joseph E Powell. scpred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.*, 20(1):264, December 2019.

[10] Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, 21(1):30, February 2020.

[11] Sasan Amini, Dmitry Pushkarev, Lena Christiansen, Emrah Kostem, Tom Royce, Casey Turk, Natasha Pignatelli, Andrew Adey, Jacob O Kitzman, Kandaswamy Vijayan, Mostafa Ronaghi, Jay Shendure, Kevin L Gunderson, and Frank J Steemers. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.*, 46(12):1343–1349, December 2014.

[12] Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Res.*, 7:1740, November 2018.

[13] Tallulah S Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Mol. Aspects Med.*, 59:114–122, February 2018.

[14] Michael Angelo, Sean C Bendall, Rachel Finck, Matthew B Hale, Chuck Hitzman, Alexander D Borowsky, Richard M Levenson, John B Lowe, Scot D Liu, Shuchun Zhao, Yasodha Natkunam, and Garry P Nolan. Multiplexed ion beam imaging of human breast tumors. *Nat. Med.*, 20(4):436–442, April 2014.

[15] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. June 2006.

[16] David Arthur and Sergei Vassilvitskii. How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry - SCG '06*, New York, New York, USA, 2006. ACM Press.

[17] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M.

McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J. M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT, Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and NIH National Eye Institute. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393. URL https://doi.org/10.1038/nature15393.

[18] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79(2):137–158, 1944.

[19] James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods, 2005.

[20] Matthew N Bainbridge, René L Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, Malachi Griffith, Matthew Hickenbotham, Vincent Magrini, Elaine R Mardis, Marianne D Sadar, Asim S Siddiqui, Marco A Marra, and Steven J M Jones. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7:246, September 2006.

[21] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[22] Lyam Baudry, Nadège Guiglielmoni, Hervé Marie-Nelly, Alexandre Cormier, Martial Marbouty, Komlan Avia, Yann Loe Mie, Olivier Godfroy, Lieven Sterck, J Mark Cock, Christophe Zimmer, Susana M Coelho, and Romain Koszul. instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. *Genome Biol.*, 21(1):148, June 2020.

[23] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, December 2018.

[24] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo A Baybayan, Vincent A Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John A Bridgham, Rob C Brown, Andrew A Brown, Dale H Buermann, Abass A Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip A Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish,

Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T A Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008.

[25] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.*, 33(6):623–630, June 2015.

[26] Nicholas J Bernstein, Nicole L Fong, Irene Lam, Margaret A Roy, David G Hendrickson, and David R Kelley. Solo: Doublet identification in Single-Cell RNA-Seq via Semi-Supervised deep learning. *Cell Syst*, 11(1):95–101.e5, July 2020.

[27] Sayantan Bose, Zhenmao Wan, Ambrose Carr, Abbas H Rizvi, Gregory Vieira, Dana Pe'er, and Peter A Sims. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol.*, 16:120, June 2015.

[28] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One*, 12(6): e0178751, June 2017.

[29] Gerard Brady, Mary Barbara, and Norman N Iscove. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol. Cell. Biol.*, 2(1):17–25, 1990.

[30] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(8):888, August 2016.

[31] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10(11):1093–1095, November 2013.

[32] S Brenner, F Jacob, and M Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581, May 1961.

[33] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81(5):1084–1097, November 2007.

[34] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12(10):703–714, September 2011.

[35] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, 33(2):155–160, February 2015.

[36] W N Burnette. “western blotting”: electrophoretic transfer of proteins from sodium dodecyl sulfate–polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein a. *Anal. Biochem.*, 112(2): 195–203, April 1981.

[37] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18(5):810–820, May 2008.

[38] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726): 203–209, October 2018.

[39] José Antonio Cabezas, Marian Morcillo, María Dolores Vélez, Luis Díaz, Juan Segura, María Teresa Cervera, and Isabel Arrillaga. Haploids in conifer species: Characterization and chromosomal integrity of a maritime pine cell line. *For. Trees Livelihoods*, 7(11):274, November 2016.

[40] B Canard and R S Sarfati. DNA polymerase fluorescent substrates with reversible 3’-tags. *Gene*, 148(1):1–6, October 1994.

[41] Sadi Carnot. Reflections on the motive power of fire, and on machines fitted to develop that power. *Paris: Bachelier*, 108, 1824.

[42] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13:238, September 2012.

[43] Mark J P Chaisson, Richard K Wilson, and Evan E Eichler. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.*, 16(11):627–640, November 2015.

[44] Richard Challis, Edward Richards, Jeena Rajan, Guy Cochrane, and Mark Blaxter. BlobToolKit–Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, 10(4):1361–1374, 2020.

[45] Erika Check Hayden. Rna editing study under intense scrutiny. *Nature News*, 2012.

[46] Xi Chen, Sarah A Teichmann, and Kerstin B Meyer. From tissues to cell types and back: Single-Cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science*, July 2018.

[47] Ying Chen, Weicai Ye, Yongdong Zhang, and Yuesheng Xu. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.*, 43(16):7762–7768, September 2015.

[48] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, 18(2):170–175, February 2021.

[49] Chen-Shan Chin, Jon Sorenson, Jason B Harris, William P Robins, Richelle C Charles, Roger R Jean-Charles, James Bullard, Dale R Webster, Andrew Kasarskis, Paul Peluso, Ellen E Paxinos, Yoshiharu Yamaichi, Stephen B Calderwood, John J Mekalanos, Eric E Schadt, and Matthew K Waldor. The origin of the haitian cholera outbreak strain. *N. Engl. J. Med.*, 364(1):33–42, January 2011.

[50] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, 10(6): 563–569, June 2013.

[51] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, 10(6): 563–569, June 2013.

[52] 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.

[53] Maureen Coetzee, Richard H Hunt, Richard Wilkerson, Alessandra Della Torre, Mamadou B Coulibaly, and Nora J Besansky. Anopheles coluzzii and anopheles amharicus, new members of the anopheles gambiae complex. *Zootaxa*, 3619:246–274, 2013.

[54] Mouse Genome Sequencing Consortium and Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome, 2002.

[55] Albert H Coons, Hugh J Creech, and R Norman Jones. Immunological properties of an antibody containing a fluorescent group. *Proc. Soc. Exp. Biol. Med.*, 47(2): 200–202, June 1941.

[56] W H Coulter. US patent 2656508. 1953.

[57] F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970.

[58] F H Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

[59] F H Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

[60] Matei David, L J Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson. Nanocall: an open source basecaller for oxford nanopore sequencing data. *Bioinformatics*, 33 (1):49–55, January 2017.

[61] D W Deamer and M Akeson. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.*, 18(4):147–151, April 2000.

[62] Ethan C Degner and Laura C Harrington. A mosquito sperm's journey from male ejaculate to egg: Mechanisms, molecules, and methods for exploration. *Mol. Reprod. Dev.*, 83(10):897–911, October 2016.

[63] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.

[64] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.

[65] Olivier Delaneau, Jean-François Zagury, Matthew R Robinson, Jonathan L Marchini, and Emmanouil T Dermitzakis. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.*, 10(1):5436, November 2019.

[66] Megan Y Dennis and Evan E Eichler. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.*, 41:44–52, December 2016.

[67] Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Szymon Grabowski. Disk-based k-mer counting on a PC, 2013.

[68] Sebastian Deorowicz, Marek Kokot, Szymon Grabowski, and Agnieszka Debudaj-Grabysz. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10): 1569–1576, May 2015.

[69] Erica A K DePasquale, Daniel Schnell, Phillip Dexheimer, Kyle Ferchen, Stuart Hay, Kashish Chetal, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, and Nathan Salomonis. cellharmony: cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic Acids Res.*, 47(21):e138, December 2019.

[70] Erica A K DePasquale, Daniel J Schnell, Pieter-Jan Van Camp, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, Harinder Singh, and Nathan Salomonis. DoubletDecon: Deconvoluting doublets from Single-Cell RNA-Sequencing data. *Cell Rep.*, 29(6):1718–1727.e8, November 2019.

[71] 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.

[72] Jun Ding, Chieh Lin, and Ziv Bar-Joseph. Cell lineage inference from SNP and scRNA-Seq data. *Nucleic Acids Res.*, 47(10):e56, June 2019.

[73] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[74] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, February 2015.

[75] 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.

[76] R Drmanac, I Labat, I Brukner, and R Crkvenjakov. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics*, 4(2):114–128, February 1989.

[77] 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.

[78] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, April 2017.

[79] Olga Dudchenko, Muhammad S Shamim, Sanjit S Batra, Neva C Durand, Nathaniel T Musial, Ragib Mostofa, Melanie Pham, Brian Glenn St Hilaire, Weijie Yao, Elena Stamenova, et al. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under 1000.*BioRxiv*,*page*254797, 2018.

[80] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas S P Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments, 2016.

[81] Richard Durbin. A toolset for efficient analysis of 10x genomics linked read data sets. https://github.com/richarddurbin/hash10x, 2018.

[82] Richard Durbin. A toolset for fast and space efficient dna read set matching and assembly using a simple kmer sampling approach. https://github.com/richarddurbin/modimizer, 2018.

[83] Dent Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R Zerbino, Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne, Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T Simpson, Nuno A Fonseca, İnanç Birol, T Roderick Docking, Isaac Y Ho, Daniel S Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis, Delphine Naquin, Nicolas Maillet, Michael C Schatz, David R Kelley, Adam M Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu, Wen-Chi Chou, Anuj Srivastava, Timothy I Shaw, J Graham Ruby, Peter Skewes-Cox, Miguel Betegon, Michelle T Dimon, Victor Solovyev, Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez, Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li, Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall, Paul J Kersey, Richard Durbin, Shaun D Jackman, Jarrod A Chapman, Xiaoqiu Huang, Joseph L DeRisi, Mario Caccamo, Yingrui Li, David B Jaffe, Richard E Green, David Haussler, Ian Korf, and Benedict Paten. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.*, 21(12): 2224–2241, December 2011.

[84] 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.

[85] 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.

[86] Adam C English, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M Muzny, Jeffrey G Reid, Kim C Worley, and Richard A Gibbs. Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS One*, 7(11):e47768, November 2012.

[87] Shigeyuki Esumi, Sheng-Xi Wu, Yuchio Yanagawa, Kunihiko Obata, Yukihiko Sugimoto, and Nobuaki Tamamaki. Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors. *Neurosci. Res.*, 60(4): 439–451, April 2008.

[88] Ester Falconer, Mark Hills, Ulrike Naumann, Steven S S Poon, Elizabeth A Chavez, Ashley D Sanders, Yongjun Zhao, Martin Hirst, and Peter M Lansdorp. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods*, 9(11):1107–1112, November 2012.

[89] H C Fan, G K Fu, and S P A Fodor. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 2015.

[90] Omid R Faridani, Ilgar Abdullayev, Michael Hagemann-Jensen, John P Schell, Fredrik Lanner, and Rickard Sandberg. Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, 34(12):1264–1266, December 2016.

[91] P Ferragina and G Manzini. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398, November 2000.

[92] Paolo Ferragina, Giovanni Manzini, Veli Mäkinen, and Gonzalo Navarro. An Alphabet-Friendly FM-Index. In *String Processing and Information Retrieval*, pages 150–160. Springer Berlin Heidelberg, 2004.

[93] M J Fulwyler. Electronic separation of biological cells by volume. *Science*, 150(3698): 910–911, November 1965.

[94] 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.

[95] Shilpa Garg, Marcel Martin, and Tobias Marschall. Read-based phasing of related individuals. *Bioinformatics*, 32(12):i234–i242, June 2016.

[96] Shilpa Garg, Mikko Rautiainen, Adam M Novak, Erik Garrison, Richard Durbin, and Tobias Marschall. A graph-based approach to diploid genome assembly. *Bioinformatics*, 34(13):i105–i114, July 2018.

[97] Shilpa Garg, Arkarachai Fungtammasan, Andrew Carroll, Mike Chou, Anthony Schmitt, Xiang Zhou, Stephen Mac, Paul Peluso, Emily Hatas, Jay Ghurye, Jared Maguire, Medhat Mahmoud, Haoyu Cheng, David Heller, Justin M Zook, Tobias Moemke, Tobias Marschall, Fritz J Sedlazeck, John Aach, Chen-Shan Chin, George M Church, and Heng Li. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.*, 39(3):309–312, March 2021.

[98] Shilpa Garg, Arkarachai Fungtammasan, Andrew Carroll, Mike Chou, Anthony Schmitt, Xiang Zhou, Stephen Mac, Paul Peluso, Emily Hatas, Jay Ghurye, Jared Maguire, Medhat Mahmoud, Haoyu Cheng, David Heller, Justin M Zook, Tobias Moemke, Tobias Marschall, Fritz J Sedlazeck, John Aach, Chen-Shan Chin, George M Church, and Heng Li. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.*, 39(3):309–312, March 2021.

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

[100] Jellert T Gaublomme, Bo Li, Cristin McCabe, Abigail Knecht, Yiming Yang, Eugene Drokhlyansky, Nicholas Van Wittenberghe, Julia Waldman, Danielle Dionne, Lan Nguyen, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nature communications*, 10(1):1–8, 2019.

[101] Maryam Ghareghani, David Porubsk?, Ashley D Sanders, Sascha Meiers, Evan E Eichler, Jan O Korbel, and Tobias Marschall. Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics*, 34(13):i115–i123, July 2018.

[102] Shila Ghazanfar, Yingxin Lin, Xianbin Su, David Ming Lin, Ellis Patrick, Ze-Guang Han, John C Marioni, and Jean Yee Hwa Yang. Investigating higher-order interactions in single-cell data with scHOT. *Nat. Methods*, 17(8):799–806, August 2020.

[103] Jay Ghurye, Arang Rhie, Brian P Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M Phillippy, and Sergey Koren. Integrating Hi-C links with assembly graphs for chromosome-scale assembly.

[104] Jay Ghurye, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, 18(1):527, July 2017.

[105] Jay Ghurye, Sergey Koren, Scott T Small, Seth Redmond, Paul Howell, Adam M Phillippy, and Nora J Besansky. A chromosome-scale assembly of the major african malaria vector anopheles funestus. *Gigascience*, 8(6), June 2019.

[106] 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.

[107] Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11(4):417–422, April 2014.

[108] Peter Goldstein, William Heaton, Franco Preparata, and Eli Upfal. Distance maps using multiple alignment consensus construction, September 18 2014. US Patent App. 14/212,458.

[109] 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.

[110] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, May 2016.

[111] F Gros, H Hiatt, W Gilbert, C G Kurland, R W Risebrough, and J D Watson. Unstable ribonucleic acid revealed by pulse labelling of escherichia coli. *Nature*, 190:581–585, May 1961.

[112] Dengfeng Guan, Shane A McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, and Richard Durbin. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, 36(9):2896–2898, May 2020.

[113] Chuner Guo, Wenjun Kong, Kenji Kamimoto, Guillermo C Rivera-Gonzalez, Xue Yang, Yuhei Kirita, and Samantha A Morris. CellTag indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.*, 20(1):90, May 2019.

[114] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, March 2017.

[115] Ishaan Gupta, Paul G Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B Smit, Steven A Sloan, Wenjie Luo, Olivier Fedrigo, M Elizabeth Ross, and Hagen U Tilgner. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*, October 2018.

[116] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, 36(5):421–427, June 2018.

[117] R W Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, April 1950.

[118] S Hannenhalli, W Feldman, H F Lewis, S S Skiena, and P A Pevzner. Positional sequencing by hybridization. *Comput. Appl. Biosci.*, 12(1):19–24, February 1996.

[119] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[120] J Hawks, K Hunley, S H Lee, and M Wolpoff. Population bottlenecks and pleistocene human evolution. *Mol. Biol. Evol.*, 17(1):2–22, January 2000.

[121] Ning Z HE. Scaff10x v4.2: Pipeline for scaffolding and breaking a genome assembly using 10x genomics linked-reads. https://github.com/wtsi-hpag/Scaff10X, 2018.

[122] James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, January 2016.

[123] Haynes Heaton, Arthur M Talman, Andrew Knights, Maria Imaz, Daniel J Gaffney, Richard Durbin, Martin Hemberg, and Mara K N Lawniczak. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods*, 17(6):615–620, June 2020.

[124] Michelle E H Helinski, Laura Valerio, Luca Facchinelli, Thomas W Scott, Janine Ramsey, and Laura C Harrington. Evidence of polyandry for aedes aegypti in semifield enclosures. *Am. J. Trop. Med. Hyg.*, 86(4):635–641, April 2012.

[125] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.*, 37(6):685–691, June 2019.

[126] Paul G Higgs and Ralph E Pudritz. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*, 9(5):483–490, June 2009.

[127] Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840, 2002.

[128] R W Holley, J Apgar, G A Everett, J T Madison, M Marquisee, S H Merrill, J R Penswick, and A Zamir. STRUCTURE OF a RIBONUCLEIC ACID. *Science*, 147(3664): 1462–1465, March 1965.

[129] 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.

[130] Rui Hou, Elena Denisenko, and Alistair R R Forrest. scmatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics*, 35(22): 4688–4695, November 2019.

[131] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.*, 21(1):218, August 2020.

[132] Kerstin Howe, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey, and Jonathan Wood. Significantly improving the quality of genome assemblies through curation.

[133] Kerstin Howe, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey, and Jonathan Wood. Significantly improving the quality of genome assemblies through curation. *Gigascience*, 10(1):giaa153, 2021.

[134] Virginia M Howick, Andrew J C Russell, Tallulah Andrews, Haynes Heaton, Adam J Reid, Kedar Natarajan, Hellen Butungi, Tom Metcalf, Lisa H Verzier, Julian C Rayner, Matthew Berriman, Jeremy K Herren, Oliver Billker, Martin Hemberg, Arthur M Talman, and Mara K N Lawniczak. The malaria cell atlas: Single parasite transcriptomes across the complete plasmodium life cycle. *Science*, 365(6455), August 2019.

[135] Shengfeng Huang, Mingjing Kang, and Anlong Xu. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33(16):2577–2579, August 2017.

[136] X Huang. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14(1):18–25, September 1992.

[137] Xianjie Huang and Yuanhua Huang. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics*, May 2021.

[138] Yuanhua Huang, Davis J McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.*, 20(1):273, December 2019.

[139] T Hunkapiller, R J Kaiser, B F Koop, and L Hood. Large-scale and automated DNA sequence determination. *Science*, 254(5028):59–67, October 1991.

[140] Julian Huxley and Others. Evolution. the modern synthesis. *Evolution. The Modern Synthesis.*, 1942.

[141] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, 50(8):1–14, August 2018.

[142] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, 17:29, February 2016.

[143] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono,

Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[144] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat. Genet.*, 44(2): 226–232, January 2012.

[145] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2):163–166, February 2014.

[146] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS One*, 9(6):e98679, June 2014.

[147] Chirag Jain, Alexander Dilthey, Sergey Koren, Srinivas Aluru, and Adam M Phillippy. A fast approximate algorithm for mapping long reads to large reference databases, 2018.

[148] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang

Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, 36(4):338–345, April 2018.

[149] Miten Jain, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson, and Karen H Miga. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.*, 36(4):321–323, April 2018.

[150] 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.

[151] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemp. Math.*, 26, 1984.

[152] Emil Jørsboe, Kristian Hanghøj, and Anders Albrechtsen. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics*, 33(19):3148–3150, October 2017.

[153] JP Joule. I. on the œconomical production of mechanical effect from chemical forces. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 5 (29):1–5, 1853.

[154] 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.

[155] Kasper Karlsson and Sten Linnarsson. Single-cell mRNA isoform diversity in the mouse brain. *BMC Genomics*, 18(1):126, February 2017.

[156] J J Kasianowicz, E Brandin, D Branton, and D W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, 93 (24):13770–13773, November 1996.

[157] John D Kececioglu and Eugene W Myers. Combinatorial algorithms for dna sequence assembly. *Algorithmica*, 13(1):7–51, 1995.

[158] W J Kent and D Haussler. Assembly of the working draft of the human genome with GigAssembler. *Genome Res.*, 11(9):1541–1548, September 2001.

[159] W James Kent. BLAT—The BLAST-Like alignment tool. *Genome Res.*, 12(4):656–664, April 2002.

[160] Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M Luber, Scott B Ouellette, Alaleh Azhir, Nikhil Kumar, Jeewon Hwang, Soohyun Lee, Burak H Alver, Hanspeter Pfister, Leonid A Mirny, Peter J Park,

and Nils Gehlenborg. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.*, 19(1):125, August 2018.

[161] Helena Kilpinen, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, Dalila Bensaddek, Francesco Paolo Casale, Oliver J Culley, Petr Danecek, Adam Faulconbridge, Peter W Harrison, Annie Kathuria, Davis McCarthy, Shane A McCarthy, Ruta Meleckyte, Yasin Memari, Nathalie Moens, Filipa Soares, Alice Mann, Ian Streeter, Chukwuma A Agu, Alex Alderton, Rachel Nelson, Sarah Harper, Minal Patel, Alistair White, Sharad R Patel, Laura Clarke, Reena Halai, Christopher M Kirton, Anja Kolb-Kokocinski, Philip Beales, Ewan Birney, Davide Danovi, Angus I Lamond, Willem H Ouwehand, Ludovic Vallier, Fiona M Watt, Richard Durbin, Oliver Stegle, and Daniel J Gaffney. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*, 546(7658):370–375, June 2017.

[162] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, April 2013.

[163] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, 12(4):357–360, April 2015.

[164] 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.

[165] Mirna Kirin, Ruth McQuillan, Christopher S Franklin, Harry Campbell, Paul M McKeigue, and James F Wilson. Genomic runs of homozygosity record population history and consanguinity. *PLoS One*, 5(11):e13996, November 2010.

[166] S Kirkpatrick, C D Gelatt, Jr, and M P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[167] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data, 2017.

[168] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, 15(5):359–362, May 2018.

[169] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, November 2011.

[170] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

[171] Dmitry Kobak and George C Linderman. UMAP does not preserve global structure any better than t-SNE when using the same initialization.

[172] Marek Kokot, Maciej Dlugosz, and Sebastian Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, September 2017.

[173] 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.

[174] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy P L Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*, October 2018.

[175] Jonas Korlach, Gregory Gedman, Sarah B Kingan, Chen-Shan Chin, Jason T Howard, Jean-Nicolas Audet, Lindsey Cantin, and Erich D Jarvis. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience*, 6(10):1–16, October 2017.

[176] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15:356, November 2014.

[177] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12):1289–1296, December 2019.

[178] Zev N Kronenberg, Richard J Hall, Stefan Hiendleder, Timothy PL Smith, Shawn T Sullivan, John L Williams, and Sarah B Kingan. Falcon-phase: integrating pacbio and hi-c data for phased diploid genomes. *BioRxiv*, page 327064, 2018.

[179] 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.

[180] Kazuki Kurimoto, Yukihiro Yabuta, Yasuhide Ohinata, Yukiko Ono, Kenichiro D Uno, Rikuhiro G Yamada, Hiroki R Ueda, and Mitinori Saitou. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.*, 34(5):e42, March 2006.

[181] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2):R12, January 2004.

[182] S Kyriazopoulou-Panagiotopoulou, P Marks, and others. Systems and methods for determining structural variation and phasing using variant call data, 2016.

[183] Dominik R Laetsch and Mark L Blaxter. BlobTools: Interrogation of genome assemblies. *F1000Res.*, 6(1287):1287, July 2017.

[184] E S Lander and M S Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, April 1988.

[185] 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.

[186] 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.

[187] Ben Langmead. Aligning short sequencing reads with bowtie. *Curr. Protoc. Bioinformatics*, Chapter 11:Unit 11.7, December 2010.

[188] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

[189] 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.

[190] Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, March 2002.

[191] 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.

[192] Richard M Levenson, Alexander D Borowsky, and Michael Angelo. Immunohistochemistry and mass spectrometry for highly multiplexed cellular molecular imaging. *Lab. Invest.*, 95(4):397–405, April 2015.

[193] 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.

[194] 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.

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

[196] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, July 2016.

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

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

[199] 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.

[200] Mingyao Li, Isabel X Wang, Yun Li, Alan Bruzel, Allison L Richards, Jonathan M Toung, and Vivian G Cheung. Widespread rna and dna sequence differences in the human transcriptome. *science*, 333(6038):53–58, 2011.

[201] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[202] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke,

John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[203] R J Lipshutz, S P Fodor, T R Gingeras, and D J Lockhart. High density synthetic oligonucleotide arrays. *Nat. Genet.*, 21(1 Suppl):20–24, January 1999.

[204] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, 12(8):733–735, August 2015.

[205] J A Luckey, H Drossman, A J Kostichka, D A Mead, J D'Cunha, T B Norris, and L M Smith. High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.*, 18 (15):4417–4421, August 1990.

[206] Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, April 2016.

[207] Aaron T L Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.*, 20(1):63, March 2019.

[208] Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5, 2016.

[209] Nicholas Lytal, Di Ran, and Lingling An. Normalization methods on Single-Cell RNA-seq data: An empirical survey. *Front. Genet.*, 11:41, February 2020.

[210] Nicholas Lytal, Di Ran, and Lingling An. Normalization methods on Single-Cell RNA-seq data: An empirical survey. *Front. Genet.*, 11:41, February 2020.

[211] 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.

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

[213] Udi Manber and Gene Myers. Suffix arrays: A new method for On-Line string searches. *SIAM J. Comput.*, 22(5):935–948, October 1993.

[214] 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.

[215] 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.

[216] Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Jean-François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, Nancy Guillén, Antoine Margeot, Christophe Zimmer, and Romain Koszul. High-quality genome (re)assembly using chromosomal contact data, 2014.

[217] Patrick Marks, . URL https://kb.10xgenomics.com/hc/en-us/articles/115003822406-How-does-Cell-Ranger-correct-barcode-sequencing-errors-.

[218] Patrick Marks, . URL https://kb.10xgenomics.com/hc/en-us/articles/115003822406-How-does-Cell-Ranger-correct-barcode-sequencing-errors-.

[219] 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.

[220] 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.

[221] James Clerk Maxwell and Peter Pesic. *Theory of heat*. Courier Corporation, 2001.

[222] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, Scott Vrieze, Laura J Scott, He Zhang, Anubha Mahajan, Jan Veldink, Ulrike Peters, Carlos Pato, Cornelia M van Duijn, Christopher E Gillies, Ilaria Gandin, Massimo Mezzavilla, Arthur Gilly, Massimiliano Cocca, Michela Traglia, Andrea Angius, Jeffrey C Barrett, Dorrett Boomsma, Kari Branham, Gerome Breen, Chad M Brummett, Fabio Busonero, Harry Campbell, Andrew Chan, Sai Chen, Emily Chew, Francis S Collins, Laura J Corbin, George Davey Smith, George Dedoussis, Marcus Dorr, Aliki-Eleni Farmaki, Luigi Ferrucci, Lukas Forer, Ross M Fraser, Stacey Gabriel, Shawn Levy, Leif Groop, Tabitha Harrison, Andrew Hattersley, Oddgeir L Holmen, Kristian Hveem, Matthias Kretzler, James C Lee, Matt McGue, Thomas Meitinger, David Melzer, Josine L Min, Karen L Mohlke, John B Vincent, Matthias Nauck, Deborah Nickerson, Aarno Palotie, Michele Pato, Nicola Pirastu, Melvin McInnis, J Brent Richards, Cinzia Sala, Veikko Salomaa, David Schlessinger, Sebastian Schoenherr, P Eline Slagboom, Kerrin Small, Timothy Spector, Dwight Stambolian, Marcus Tuke, Jaakko Tuomilehto, Leonard H Van den Berg, Wouter Van Rheenen, Uwe Volker, Cisca Wijmenga, Daniela Toniolo, Eleftheria Zeggini, Paolo Gasparini, Matthew G Sampson, James F Wilson, Timothy Frayling, Paul I W de Bakker, Morris A Swertz, Steven McCarroll, Charles Kooperberg, Annelot Dekker, David Altshuler, Cristen Willer, William Iacono, Samuli Ripatti, Nicole Soranzo, Klaudia Walter, Anand Swaroop, Francesco Cucca, Carl A Anderson, Richard M Myers, Michael Boehnke, Mark I McCarthy, Richard Durbin, and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, 48(10): 1279–1283, October 2016.

[223] Rajiv C McCoy, Ryan W Taylor, Timothy A Blauwkamp, Joanna L Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A Petrov, and Anna-Sophie Fiston-Lavier. Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PloS one*, 9(9):e106689, 2014.

[224] Mark A McElwain, Rebecca Yu Zhang, Radoje Drmanac, and Brock A Peters. Long fragment read (LFR) technology: Cost-Effective, High-Quality Genome-Wide molecular haplotyping. *Methods Mol. Biol.*, 1551:191–205, 2017.

[225] C S McGinnis, L M Murrow, and Z J Gartner. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *bioRxiv*, 2018.

[226] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, July 2019.

[227] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection, 2018.

[228] Joana I Meier, Patricio A Salazar, Marek Kučka, Robert William Davies, Andreea Dréau, Ismael Aldás, Olivia Box Power, Nicola J Nadeau, Jon R Bridle, Campbell Rolian, Nicholas H Barton, W Owen McMillan, Chris D Jiggins, and Yingguang Frank Chan.

Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl. Acad. Sci. U. S. A.*, 118(25), June 2021.

[229] Gregor Mendel. Experiments in Plant-Hybridization. https://essayzilla.org/wp-content/uploads/2020/12/20190405221854mendel_1866___1__.pdf. Accessed: 2021-8-20.

[230] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, September 2020.

[231] Nan Miles Xi and Jingyi Jessica Li. Protocol for benchmarking computational doublet-detection methods in single-cell rna sequencing data analysis. *arXiv e-prints*, pages arXiv–2101, 2021.

[232] Ashley Moffett and Francesco Colucci. Co-evolution of NK receptors and HLA ligands in humans is driven by reproduction. *Immunol. Rev.*, 267(1):283–297, 2015.

[233] A P Monaco and Z Larin. YACs, BACs, PACs and MACs: artificial chromosomes as research tools. *Trends Biotechnol.*, 12(7):280–286, July 1994.

[234] Aleksandr Morgulis, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, 22(2):134–141, January 2006.

[235] Adam Moroz. *The common extremalities in biology and physics: Maximum energy dissipation principle in chemistry, biology, physics and evolution.* Elsevier Science Publishing, Philadelphia, PA, 2 edition, November 2011.

[236] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7): 621–628, July 2008.

[237] 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.

[238] Kary B Mullis. Process for amplifying nucleic acid sequences, July 1987.

[239] Kary B Mullis and Fred A Faloona. Specific synthesis of DNA in vitro via a Polymerase-Catalyzed chain reaction, 1989.

[240] 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.

[241] 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.

[242] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2: ii79–85, September 2005.

[243] Gene Myers. Fastk. https://github.com/thegenemyers/FASTK, 2020.

[244] 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.

[245] 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.

[246] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, March 1970.

[247] Jim Nemesh and Alec Wysoker. Drop-seq. https://github.com/broadinstitute/Drop-seq, 2015.

[248] Sergey Nurk, Brian P Walenz, Arang Rhie, Mitchell R Vollger, Glennis A Logsdon, Robert Grothe, Karen H Miga, Evan E Eichler, Adam M Phillippy, and Sergey Koren. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.*, 30(9):1291–1305, September 2020.

[249] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina V Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G de Lima, Philip C Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T Fiddes, Giulio Formenti, Robert S Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G S Grady, Tina A Graves-Lindsay, Ira M Hall, Nancy F Hansen, Gabrielle A Hartley, Marina Haukness, Kerstin Howe, Michael W Hunkapiller, Chirag Jain, Miten Jain, Erich D Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V Maduro, Tobias Marschall, Ann M McCartney, Jennifer McDaniel, Danny E Miller, James C Mullikin, Eugene W Myers, Nathan D Olson, Benedict Paten,

Paul Peluso, Pavel A Pevzner, David Porubsky, Tamara Potapova, Evgeny I Rogaev, Jeffrey A Rosenfeld, Steven L Salzberg, Valerie A Schneider, Fritz J Sedlazeck, Kishwar Shafin, Colin J Shew, Alaina Shumate, Yumi Sims, Arian F A Smit, Daniela C Soto, Ivan Sović, Jessica M Storer, Aaron Streets, Beth A Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P Walenz, Aaron Wenger, Jonathan M D Wood, Chunlin Xiao, Stephanie M Yan, Alice C Young, Samantha Zarate, Urvashi Surti, Rajiv C McCoy, Megan Y Dennis, Ivan A Alexandrov, Jennifer L Gerton, Rachel J O'Neill, Winston Timp, Justin M Zook, Michael C Schatz, Evan E Eichler, Karen H Miga, and Adam M Phillippy. The complete sequence of a human genome. May 2021.

[250] Ludovic Orlando, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, Enrico Cappellini, Bent Petersen, Ida Moltke, Philip L F Johnson, Matteo Fumagalli, Julia T Vilstrup, Maanasa Raghavan, Thorfinn Korneliussen, Anna-Sapfo Malaspinas, Josef Vogt, Damian Szklarczyk, Christian D Kelstrup, Jakob Vinther, Andrei Dolocan, Jesper Stenderup, Amhed M V Velazquez, James Cahill, Morten Rasmussen, Xiaoli Wang, Jiumeng Min, Grant D Zazula, Andaine Seguin-Orlando, Cecilie Mortensen, Kim Magnussen, John F Thompson, Jacobo Weinstock, Kristian Gregersen, Knut H Røed, Véra Eisenmann, Carl J Rubin, Donald C Miller, Douglas F Antczak, Mads F Bertelsen, Søren Brunak, Khaled A S Al-Rasheid, Oliver Ryder, Leif Andersson, John Mundy, Anders Krogh, M Thomas P Gilbert, Kurt Kjær, Thomas Sicheritz-Ponten, Lars Juhl Jensen, Jesper V Olsen, Michael Hofreiter, Rasmus Nielsen, Beth Shapiro, Jun Wang, and Eske Willerslev. Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. *Nature*, 499(7456):74–78, July 2013.

[251] Daniel Osorio and James J Cai. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics*, 37(7):963–967, May 2021.

[252] Alexander F Palazzo and Eliza S Lee. Non-coding RNA: what is functional and what is junk? *Front. Genet.*, 6:2, January 2015.

[253] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4): 417–419, April 2017.

[254] 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.

[255] Patrick Marks et al. Paul Ryvkin. Cell ranger. https://github.com/10XGenomics/cellranger, 2016.

[256] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, 12(8):780–786, August 2015.

[257] P A Pevzner, H Tang, and M S Waterman. An eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.*, 98(17):9748–9753, August 2001.

[258] Ernesto Picardi, Anna Maria D'Erchia, Claudio Lo Giudice, and Graziano Pesole. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.*, 45(D1):D750–D757, January 2017.

[259] 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.

[260] 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.

[261] Hannah A Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, 16(10):983–986, October 2019.

[262] Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, February 2020.

[263] David Porubský, Ashley D Sanders, Niek van Wietmarschen, Ester Falconer, Mark Hills, Diana C J Spierings, Marianna R Bevova, Victor Guryev, and Peter M Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing, 2016.

[264] David Porubsky, Peter Ebert, Peter A Audano, Mitchell R Vollger, William T Harvey, Katherine M Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, et al. A fully phased accurate assembly of an individual human genome. *bioRxiv*, 2019.

[265] David Porubsky, Peter Ebert, Peter A Audano, Mitchell R Vollger, William T Harvey, Pierre Marijon, Jana Ebler, Katherine M Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, Human Genome Structural Variation Consortium, Peter M Lansdorp, Benedict Paten, Scott E Devine, Ashley D Sanders, Charles Lee, Mark J P Chaisson, Jan O Korbel, Evan E Eichler, and Tobias Marschall. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.*, 39(3):302–308, March 2021.

[266] Vladimir Potapov, Xiaoqing Fu, Nan Dai, Ivan R Corrêa Jr, Nathan A Tanner, and Jennifer L Ong. Base modifications affecting rna polymerase and reverse transcriptase fidelity. *Nucleic acids research*, 46(11):5753–5763, 2018.

[267] F P Preparata and E Upfal. Sequencing-by-hybridization at the information-theory bound: an optimal algorithm. *J. Comput. Biol.*, 7(3-4):621–630, 2000.

[268] Robert F Purnell, Kunal K Mehta, and Jacob J Schmidt. Nucleotide identification and orientation discrimination of DNA homopolymers immobilized in a protein nanopore. *Nano Lett.*, 8(9):3029–3034, September 2008.

[269] Xiaojie Qiu, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. Single-cell mRNA quantification and differential analysis with census. *Nat. Methods*, 14 (3):309–315, March 2017.

[270] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[271] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

[272] Edward S Rice and Richard E Green. New approaches for genome assembly and scaffolding. *Annu Rev Anim Biosci*, 7:17–40, February 2019.

[273] 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.

[274] Michael Roberts, Wayne Hayes, Brian R Hunt, Stephen M Mount, and James A Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20 (18):3363–3369, December 2004.

[275] Peter Robinson and Tomasz Zemo Jtel. Integrative genomics viewer (IGV): Visualizing alignments and variants, 2017.

[276] K Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, November 1998.

[277] Julien Rougeot, Ania Zakrzewska, Zakia Kanwal, Hans J Jansen, Herman P Spaink, and Annemarie H Meijer. RNA sequencing of FACS-sorted immune cell populations from zebrafish infection models to identify cell specific responses to intracellular pathogens. *Methods Mol. Biol.*, 1197:261–274, 2014.

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

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

[280] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, 37(5):547–554, May 2019.

[281] Steven L Salzberg, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, Guillaume Marçais, Mihai Pop, and James A Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 22(3):557–567, March 2012.

[282] F Sanger and H Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. 1. the identification of lower peptides from partial hydrolysates. *Biochem. J*, 49(4):463–481, September 1951.

[283] F Sanger, G G Brownlee, and B G Barrell. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.*, 13(2):373–398, September 1965.

[284] F Sanger, J E Donelson, A R Coulson, H Kössel, and D Fischer. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 70(4):1209–1213, April 1973.

[285] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467, December 1977.

[286] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, May 2015.

[287] Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Res.*, 20(9):1165–1173, September 2010.

[288] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, Hyeran Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J Levy, Linda McMahan, Peter Van Buren, Matthew W Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W Brad Barbazuk, Regina S Baucom, Thomas P Brutnell, Nicholas C Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C Estill, Yan Fu, Jeffrey A Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C McCann, Phillip SanMiguel, Alan M Myers, Dan Nettleton, John Nguyen, Bryan W Penning, Lalit Ponnala, Kevin L Schneider, David C Schwartz, Anupma Sharma, Carol Soderlund, Nathan M Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L Bennetzen, R Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G Presting, Susan R Wessler, Srinivas Aluru, Robert A Martienssen, Sandra W Clifton, W Richard McCombie, Rod A Wing, and Richard K Wilson. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956): 1112–1115, November 2009.

[289] 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.

[290] Erwin Schrodinger. What is life? *J. Philos.*, 43(7):194, March 1946.

[291] Andrew W Schroeder, Swastika Sur, Priyanka Rashmi, Izabella Damm, Arya Zarinsefat, Matthias Kretzler, Jeff Hodgin, George Hartoularos, Tara Sigdel, Jimmie Chun Ye, et al. Novel human kidney cell subsets identified by mux-seq. *bioRxiv*, 2020.

[292] D Schwartz, X Li, L Hernandez, S Ramnarain, E Huff, and Y Wang. Ordered restriction maps of saccharomyces cerevisiae chromosomes constructed by optical mapping, 1993.

[293] 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.

[294] 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.

[295] 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.

[296] 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.

[297] Jared T Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.*, 22(3):549–556, March 2012.

[298] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19(6):1117–1123, June 2009.

[299] Mandeep Singh, Ghamdan Al-Eryani, Shaun Carswell, James M Ferguson, James Blackburn, Kirston Barton, Daniel Roden, Fabio Luciani, Tri Giang Phan, Simon Junankar, Katherine Jackson, Christopher C Goodnow, Martin A Smith, and Alexander Swarbrick. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.*, 10(1):3120, July 2019.

[300] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, November 2013.

[301] T F Smith and M S Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, March 1981.

[302] Edward C Stack, Chichung Wang, Kristin A Roman, and Clifford C Hoyt. Multiplexed immunohistochemistry, imaging, and quantitation: a review, with an assessment of tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods*, 70 (1):46–58, November 2014.

[303] N M Stevens. A study of the germ cells of certain diptera, with reference to the heterochromosomes and the phenomena of synapsis, 1908.

[304] 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.

[305] Ian Streeter, Peter W Harrison, Adam Faulconbridge, Paul Flicek, Helen Parkinson, and Laura Clarke. The human-induced pluripotent stem cell initiative—data resources for cellular genetics. *Nucleic Acids Res.*, 45(D1):D691–D697, January 2017.

[306] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

[307] Granger G Sutton, Owen White, Mark D Adams, and Anthony R Kerlavage. TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19, January 1995.

[308] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, 14(4):381–387, April 2017.

[309] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, 13(4):599–604, April 2018.

[310] 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.

[311] 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.

[312] 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.

[313] 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.

[314] Takashi Tokuda, Kunihiro Tanaka, Masamichi Matsuo, Keiichiro Kagawa, Masahiro Nunoshita, and Jun Ohta. Optical and electrochemical dual-image cmos sensor for on-chip biomolecular sensing applications. *Sensors and Actuators A: Physical*, 135(2): 315–322, 2007.

[315] H Towbin, T Staehelin, and J Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U. S. A.*, 76(9):4350–4354, September 1979.

[316] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[317] Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, 38(15):e159, August 2010.

[318] Naonori Ueda and Ryohei Nakano. Deterministic annealing variant of the EM algorithm. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 545–552, Cambridge, MA, USA, January 1994. MIT Press.

[319] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. BASiCS: Bayesian analysis of Single-Cell sequencing data. *PLoS Comput. Biol.*, 11(6):e1004333, June 2015.

[320] Laurens van der Maaten. Visualizing data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwA, 2008. Accessed: 2021-8-30.

[321] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[322] R N Van Gelder, M E von Zastrow, A Yool, W C Dement, J D Barchas, and J H Eberwine. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. U. S. A.*, 87(5):1663–1667, March 1990.

[323] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, 27(5):737–746, May 2017.

[324] Andrea Vattani. k-means requires exponentially many iterations even in the plane. *Discrete Comput. Geom.*, 45(4):596–616, June 2011.

[325] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato,

V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, February 2001.

[326] Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, Kerstin B Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, Lucy Gardner, Staffan Holmqvist, Johan Henriksson, Angela Zou, Andrew M Sharkey, Ben Millar, Barbara Innes, Laura Wood, Anna Wilbrey-Clark, Rebecca P Payne, Martin A Ivarsson, Steve Lisgo, Andrew Filby, David H Rowitch, Judith N Bulmer, Gavin J Wright, Michael J T Stubbington, Muzlifah Haniffa, Ashley Moffett, and Sarah A Teichmann. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353, November 2018.

[327] Beate Vieth, Swati Parekh, Christoph Ziegenhain, Wolfgang Enard, and Ines Hellmann. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature communications*, 10(1):1–11, 2019.

[328] 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.

[329] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, November 2014.

[330] 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.

[331] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[332] Pacbio website. Pacbio website. https://www.pacb.com/. Accessed: 2021-09-8.

[333] Peter Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, pages 1–11, October 1973.

[334] Neil I Weisenfeld, Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, Diana Tabbaa, Louise Williams, Carsten Russ, Chad Nusbaum, Eric S Lander, Iain MacCallum, and David B Jaffe. Comprehensive variation discovery in single human genomes. *Nat. Genet.*, 46(12):1350–1355, December 2014.

[335] 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.

[336] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-Cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, June 2019.

[337] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162, October 2019.

[338] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162, October 2019.

[339] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162, October 2019.

[340] Ryan R Wick, Mark B Schultz, Justin Zobel, and Kathryn E Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, October 2015.

[341] S L Wolock, R Lopez, and A M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv*, 2018.

[342] Douglas C Wu, Jun Yao, Kevin S Ho, Alan M Lambowitz, and Claus O Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1):510, July 2018.

[343] Nan Miles Xi and Jingyi Jessica Li. Benchmarking computational Doublet-Detection methods for Single-Cell RNA sequencing data. *Cell Syst*, 12(2):176–194.e6, February 2021.

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

[345] Li Yang, Michael O Duff, Brenton R Graveley, Gordon G Carmichael, and Ling-Ling Chen. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.*, 12(2): R16, February 2011.

[346] Keren Yizhak, François Aguet, Jaegil Kim, Julian M Hess, Kirsten Kübler, Jonna Grimsby, Ruslana Frazer, Hailei Zhang, Nicholas J Haradhvala, Daniel Rosebrock, Dimitri Livitz, Xiao Li, Eila Arich-Landkof, Noam Shoresh, Chip Stewart, Ayellet V Segrè, Philip A Branton, Paz Polak, Kristin G Ardlie, and Gad Getz. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444), June 2019.

[347] Hubert P Yockey. *Information Theory, Evolution, and the Origin of Life*. Cambridge University Press, April 2005.

[348] Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data, 2020.

[349] Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, 7(7), jul 2018. doi: 10.1093/gigascience/giy083. URL http://dx.doi.org/10.1093/gigascience/giy083.

[350] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.*, 18(5):821–829, May 2008.

[351] Bin Zhang, Meichun Hsu, and Umeshwar Dayal. K-harmonic means-a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 55, 1999.

[352] Fan Zhang. *Leveraging Genetic Variants for Rapid and Robust Upstream Analysis of Massive Sequence Data*. PhD thesis, 2019.

[353] Fan Zhang, Lena Christiansen, Jerushah Thomas, Dmitry Pokholok, Ros Jackson, Natalie Morrell, Yannan Zhao, Melissa Wiley, Emily Welch, Erich Jaeger, Ana Granat, Steven J Norberg, Aaron Halpern, Maria C Rogert, Mostafa Ronaghi, Jay Shendure, Niall Gormley, Kevin L Gunderson, and Frank J Steemers. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol.*, 35(9):852–857, September 2017.

[354] 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.

[355] 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.

[356] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4): 631–643.e4, February 2017.

[357] Boris Zinshteyn and Kazuko Nishikura. Adenosine-to-inosine RNA editing. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 1(2):202–209, September 2009.

[358] Justin M Zook, Jennifer McDaniel, Nathan D Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A Irvine, Len Trigg, Rebecca Truty, Cory Y McLean, Francisco M De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.*, 37(5): 561–566, May 2019.